Improving credit risk measurement and management: A new application of statistical techniques

Nico Kritzinger

26817950

Thesis submitted for the degree Philosophiae Doctor

in Risk Analysis at the Potchefstroom Campus of the North-West University

Promoter: Prof Gary van Vuuren

Johannesburg, South Africa

October 2017

To

My parents, Ronnie and Susanna Catherina Kritzinger

Preface

All of the theoretical work in this thesis was carried out whilst in the employment of Nedbank Ltd, South Africa. All practical, and some theoretical, work was carried out in collaboration with colleagues from Nedbank Ltd and North-West University (South Africa) under the supervision of Prof Gary van Vuuren.

The papers presented in this thesis represent the original work of the author and have not been submitted in any form to any other university. In cases where work of others have been used, this has been duly acknowledged in the thesis. All data used in this thesis were sourced from a South African retail bank and bureau specific data were obtained from TransUnion South Africa.

This thesis is presented in article format in accordance with the policies of the North-West University's Faculty of Natural Sciences. The literature study on statistical techniques used in building credit scorecards and the developed credit scoring matrix model (CSMM) with results have been submitted to the *South African Journal of Economics and Management Sciences* (Chapter 2). The literature review on statistical measures used to indicate the level to separate good and bad clients for credit scorecards together with the statistical measure swap-set Gini-coefficient has been submitted to *Applied Economics* (Chapter 3).

The swap-set Gini-coefficient allows a comparison of credit scorecard predictiveness from development against post implementation accepted clients more effectively. Research on the Bayesian theorem, which provided valuable insight into the origins of probability theory and the application of the explicit form of the Bayesian formulae on retail bank data, has been submitted to *Applied Economics* (Chapter 4).

The robustness of the illustrated calibration model was evident in the various scenarios presented. The articles submitted comply with the writing style requirements (i.e. abstract, spelling, grammar and referencing requirements) of the specific journal in which the respective article has been submitted.

Acknowledgements

I acknowledge an enormous debt of gratitude to everyone that contributed in some way to the completion of this thesis.

The people mentioned below deserve special mention in appreciation for assistance rendered:

- my parents for their unconditional support and without whom the completion of this thesis would not have been possible,
- my wife Louise for the encouragement, patience, support, love and understanding,
- fellow colleagues providing many years of experience, and
- Prof Gary van Vuuren for being a fantastic and inspirational supervisor.

Abstract

Keywords: Credit Risk, credit scoring, credit risk management, calibration, Bayes

In an ever-growing economy, increased competition and pressures for increased revenue has led financial institutions to search for more effective ways to attract creditworthy clients. Since the 1950s, credit scoring has been widely adopted to guide credit decisions, however literature on credit scoring has been limited. Credit scoring plays a critical role in the banking environment which affects future impairments, capital and profits. In light of the limited research and the importance of credit scoring, the need to augment existing techniques and develop new techniques to improve credit risk measurement and management are paramount. This thesis explores three significant problems in the world of credit scoring that affect credit risk measurement and management.

The first article addresses the issue that no optimal technique exists in building credit scorecard models and also provides a solution for the disagreement on the appropriate cut-off score. An optimal Credit Scoring Matrix Model (CSMM) to determine which clients will go bad in the future is proposed. The CSMM gives uplift to the Gini-coefficient compared with a one-dimensional credit scorecard and provides a solution to determine an appropriate cut-off score on a more granular level.

The second paper explores the effect of scorecard implementation on the performance measures for the accept population. In credit scoring, much focus has been placed on modelling techniques for building credit scorecards. Less focus has been put on credit scorecard implementations. Performance measures on the accept population appear to change after the implementation of the credit scorecard against the development sample. In this paper a statistical technique called swap-set Gini-coefficient provides a more comparable statistic between development and post implementation of the credit scorecard. The swap-set Gini-coefficient performance measure results indicate significant improvement for monitoring the credit application scorecard for the accepts population.

In the third paper, the procyclicality problem on credit scorecards is investigated. The performance of a bureau scorecard during a downturn and an upturn period is analysed with the results indicating the necessity of calibration to account for procyclicality. Various calibration

scenario results are presented indicating the significant contribution of the proposed calibration model.

Table of Contents

Chapter 1
Introduction

1.1 Introduction and Background

Credit scoring is defined as the use of statistical models to transform data into numerical measures which may be used to guide credit decisions (Anderson, 2007, p. 6). The origins of credit scoring surfaced when Sir Ronald Aylmer Fisher (1936) published an article on a technique called linear discriminant analysis (Fisher, 1936). Using the same technique Durand (1941) indicated that it can also be used to discriminate between good and bad business (Durand, 1941). Wells (1992) recognised the use of statistical techniques to develop decision models leading to the first ever credit scoring system developed (Lewis, 1992). Wonderlic (1946), using his knowledge in statistics, also developed a credit score, however it was never accepted (Johnson, 2004). Sears (1950s) also used statistical derived models for decision purposes (Anderson, 2007, p. 40), however, the best known pioneers of credit scoring are Fair and Isaac (1956), with their consultancy Fair Isaac (FI) in San Francisco (Anderson, 2007, p. 40). In 1958, FI (presently known as FICO: Fair Isaac Corporation) developed its first application risk scorecards. Ever since credit scoring began to make its appearance on the world stage it has been widely adopted (Anderson, 2007, pp. 39-40). Even though credit scoring goes as far back as the 1950s, literature is very limited (Abdou & Pointon, 2011, p. 60).

The introduction of credit scoring initially had the objective of minimising risk with primary benefits of better and consistent decisions at a lower cost (Anderson, 2007, p. 512). However, since the mid-1990s the business focus shifted to the objective of improving profits. Over the years credit scoring has provided significant advantages to both decision making and to the customer. On operational processes, credit scoring provided consistency, speed, comprehensiveness and objectivity of decisions, however it comes with a cost of complexity. On the financial side, credit scoring provides less collateral management, reduced manpower cost and reduced bad debts, however this comes with significant capital investment. On the strategy side, credit scoring provides monitoring at a detailed level, improved controls over strategies and the ability to adapt in a changing environment, however credit scoring models are backward looking. On the human resource side, credit scoring provides more productive staff allocation with the cost of change management, however new skills required in credit scoring

are scarce. For the customer, credit scoring improves access to credit, provides greater mobility between lenders, is accompanied by lower costs, but this comes at the expense of losing the personal relationship with the bank (Anderson, 2007, p. 15). Credit scoring has become a significant player in credit decisions and granting credit to the right customer is of critical importance which affects future impairments for expected losses, level of capital for unexpected losses and eventually profits. Growing pressures for increased revenue generation and increased competition led financial institutions searching for more effective ways to attract creditworthy clients while at the same time controlling losses, hence the use of credit scoring (Siddiqi, 2006, p. 1).

In credit scoring, two general types of scorecards exist known as application scoring and behavioural scoring. For application credit scoring models the credit risk assessment is performed at the application stage of the loan for acquisition purposes. Behavioural scoring is a credit risk assessment after the application stage that indicate the manner in which an accepted borrower's characteristics of payment behaviour changes and is based on time dependent characteristics (Lim & Sohn, 2007, p. 427). The credit bureau score can also be related to application scoring which is based on data provided by various credit organisations to the credit bureau. In South Africa, four major credit bureaus exist, known as TransUnion, Experian, Compuscan and XDS (TransUnion, 2015). At TransUnion, the credit bureau score is known as the Empirica score and uses demographics, client judgments, client default experiences, client enquiries and client payment profiles to build the Empirica score (TransUnion & FICO, 2009, p. 2).

Retail banking within South Africa is a highly regulated environment. The Bank of International Settlements (BIS) states that "an effective system of banking supervision has clear responsibilities and objectives for each authority involved in the supervision of banks and banking groups" (Basel Committee on Banking Supervision, 2012, p. 10). In South Africa, banks are regulated by the South African Reserve Bank (SARB) which was established in 1921 by the special act of parliament and the currency and banking act of 1920 (South African Reserve Bank, 2007, p. 2). The Bank Supervision department of SARB supervises all activities of all registered banks within South Africa to achieve a sound, effective banking system in the interest of depositors of banks as well as the economy as a whole (South African Reserve Bank, 2007, p. 4). In 2005 the National Credit Regulator was established as the regulator under the

National Credit Act (NCA) that was approved in March 2006 and which went live in June 2007 (Government Gazette, 2006). The NCR is responsible for the South African credit industry to promote a fair and non-discriminatory marketplace (Government Gazette, 2006, p. 2). In June 2004, a new framework for capital measurement and standards was introduced by the BIS. In this framework it is stated under paragraph 444 that internal ratings and default and loss estimates must play an essential role in the credit approval and risk management functions of a bank under the internal ratings based (IRB) approach (Basel Committee on Banking Supervision, 2006, pp. 98-99). In September 2006 the BIS issued a newsletter on the IRB use test whereby the Basel committee recognises that some divergence is possible between IRB components and estimates used for internal purposes. One possible divergence is pricing which is more likely to use estimates based on the life of an asset (Basel Committee on Banking Supervision, 2006, pp. 1-2). It is important to note that this thesis focusses on credit scoring from an acquisition point of view for secured retail banking, meaning that the estimates of IRB components for capital purposes should be regarded as independent from the credit scoring point of view for acquisitions which uses lifetime outcome of the asset.

In this thesis, three important areas of credit scoring are explored; statistical techniques to build a credit scorecard model, credit scorecard monitoring and credit scorecard model calibration. The exploration in these three areas of credit scoring is done by means of solving significant problems identified in each area respectively. In the early developments of credit scoring models the statistical techniques used was discriminant analysis and linear probability modelling. However, due to developments in statistical software and increased computer power during the 1980s allowed credit scorecard developers to experiment with other statistical techniques to develop credit scoring models (Anderson, 2007, pp. 41-42). Various statistical techniques exist to build credit scorecards, ranging from the non-parametric statistical techniques such as expert systems to parametric statistical techniques such as logistic regression (Abdou & Pointon, 2011, pp. 66-68). Although there are various benefits for building credit scorecard models, problems that exist include reasons for including certain characteristics in the final scorecard, the general belief that there is no optimal number of characteristics in a scorecard, disagreement on the appropriate cut-off score when assessing credit at implementation stage and the determination of the sample size when building a credit scorecard. In addition, given the range of statistical techniques to build credit scoring models, no

optimal technique exists to construct a credit scorecard (Abdou & Pointon, 2011, pp. 66-68). A two-dimensional credit scorecard model is proposed to optimise the guidance of credit decisions which is compared to a one-dimensional credit scorecard model. The proposed model also provides the ability to determine a more appropriate cut-off score.

An important concept in credit scoring is called monitoring. The aim of monitoring is to observe whether the credit scoring models are working and what is happening within the processes they are used in. Monitoring can be split up into two types of report categories called front-end and back-end reports. Front-end reports focus on the stability monitoring of the credit scorecard which can be produced as soon as the credit scorecard has been implemented. Back-end reports focus on the performance of the credit scorecard based on the accepted, or booked, accounts, however time is needed to observe outcomes from the booked accounts (Anderson, 2007, p. 467). Within application scorecards, there is the added complexity of reject inference (Siddiqi, 2006, p. 98). Reject inference infer rejected applications of which the performance outcome is not known. The reason to include the reject inference process is because one would be biased if one only accepted applicants from the past were the application scorecard to be built on only accepted applicants and then implemented.

When building a credit scorecard one generally has a development sample and an independent hold-out sample within development to test the performance of the scorecard (Siddiqi, 2006, p. 127). Given the development sample and hold out sample a problem that exists when implementing an application scorecard is the comparison of performance measures between the development data and the post implementation data of the accepts population. Although reject inference can be applied to compare performance measures, to include the reject inference process for rejected applications regularly is not readily easy or possible making the comparison of performance measures for the accepts more important for monitoring purposes. It is still the general belief that application scorecard performance monitoring is done on only the accept population as a result of the effort risk managers apply regarding risk strategies and making cut-off decisions (Huang & Scott, 2007, p. 2).

A methodology and a statistical performance measure is proposed to monitor the accepts population more effectively between the development data and the post implementation data. More effectively in this context means taking scorecard implementations and the macro-environment simultaneously into account within the statistical performance measure.

The effect of economic downturns and upturns establishes another important aspect of credit scoring which is the adjustment or calibration of a credit scorecard at a certain point in the economic cycle. This effect is related to the term called procyclicality which surfaced during the introduction of the Basel capital accord that aims for capital stability though the cycle. Calibration is a requirement whenever the scores from the credit scorecard model cannot be directly associated with the required probability estimates. This non association can happen as a result of the statistical modelling technique used to build the credit scorecard model, the passing of time or the difference between the model's target variable and metric to be estimated. The latter is due to good/bad definition differences or time frames which include short- versus long-term (Anderson, 2007, p. 419).

Credit scorecards can be affected by changes in the economy, market changes, operational changes, behaviour of people and the age of the scorecard (Siddiqi, 2006, pp. 167-168; Anderson, 2007, pp. 83-90). Anderson (2007) asked a very important question whether a scorecard can be built in one point of the economic cycle and used in another. As scorecards are relatively robust, the answer is usually yes (Anderson, 2007, p. 84). Although the monotonicity of a credit scorecard could hold during the economic cycle, the level of credit risk across the scores could differ making the calibration of credit scorecards a necessity for future acquisition purposes and cut-off changes. A methodology and calibration model is proposed to address this procyclicality problem in which a credit scorecard model is calibrated to expected future bad rates. Calibration scenarios are presented which is also compared to actual bad rates experienced.

1.2 Thesis outline

Credit scoring plays an important role in guiding credit decisions which has a direct impact on future impairments, capital requirements and profits. In the current (2016) banking environment and with constant pressure from regulators it is of utmost importance that prudent credit scorecards are in place to guide credit decisions. It is important that the credit risk measurement of clients is as accurate as possible and that constant vigilance of risk management and monitoring of credit scorecards is achieved. Chapter 2 addresses the issue raised in obtaining an optimal model and determining an appropriate cut-off score when assessing credit risk of borrowers at the application stage of a loan. A literature review on non-parametric and parametric statistical techniques is provided in building credit scorecards with a

methodology presented to optimise guidance in credit decisions and determining the appropriate cut-off score.

Generally, emphasis is put on modelling techniques when building credit scorecards with less focus on the implementation of credit scorecards. It is a general belief that risk managers monitor scorecard performance of the accepts population as a result of the effort applied for risk strategies and making cut-off decisions. Chapter 3 provides a literature review on statistical performance measures that indicate the level of power or measure to separate good and bad clients. The effect on credit scorecard performance measures are investigated especially on the accept population when scorecards are implemented. It is noticeable that the level of performance measures on the accept population change after implementation and in this chapter a statistical measure is introduced to enhance the monitoring for the accepts population.

As credit scorecards are generally robust, they need not be discarded if the economy changes. Although the monotonicity across scores should hold between downturn and upturn periods the level of credit risk is not the same due to procyclicality. Chapter 4 presents a literature review on scenarios affecting scorecards and methods for identifying downturn and upturn periods. The literature review also includes historical calibration approaches followed with emphasis put on the probability origins of the Bayesian theorem and deriving an explicit form which is to be used for calibration. Chapter 4 illustrates the effect of a downturn and upturn periods on a bureau scorecard and presents a calibration model to calibrate the bureau scorecard to expected future bad rate levels.

1.3 Research design and procedure

The thesis design follows the outline below:

Research questions: Broad questions were posed in the field of credit scorecards. Although credit scorecards surfaced as far back as the 1950s, literature in the field of credit scoring has been limited. With the goal of improving credit risk measurement and management and with limited literature in credit scoring, three topics were selected.

Literature review: A critical literature review was conducted for each topic based on work done in the past in the credit scoring field. Given the problems posited in this thesis, literature

was limited to the issues raised. Where a problem was to be resolved, a new methodology and technique was presented with the effectiveness illustrated.

Theory building and testing: Current credit scoring methodologies were augmented to improve credit risk measurement and management. The introduction and development of new techniques in this thesis required thorough investigations and analysis with the effectiveness of the new techniques illustrated through validation and back-testing. All results reported in this thesis were ultimately achieved through empirical analysis, based on real retail banking data.

Data: Original data from a retail bank in South Africa were used for this thesis. Bureau related data were obtained from TransUnion South Africa. Data were abundant under each topic discussed and relevant for the analysis required.

Methodology development: The primary objective of this thesis is to improve credit risk measurement and management with the aim to implement in the retail banking environment. Statistical Analysis Software (SAS) programming was used to prepare and analyse data. The primary source of analytical work was done in Microsoft Excel with FICO's Model Builder (MB7) used to build a credit scoring model under Chapter 2. Results were validated and back tested where required. Calibrated results under Chapter 4 agreed well with actual data.

Reflection/theory extension: Results obtained from the models built and statistical measures obtained in this thesis were critically assessed, analysed and displayed. The background behind and the reason for developing methodologies to address stated problems are presented, accompanied by relevant, proven results where necessary.

State/disseminate findings: With the data analysed, relevant and meaningful results obtained were displayed effectively, all findings were written up in article style reports for peer review and publication.

Future research: Within each topic to ensure continuation of work in the field of credit scoring, other possible future related topics was proposed in this thesis for risk theorists and practitioners.

1.4 Conclusion

Credit scoring has been part of credit risk management for decades, however literature in this field has been limited. Credit scoring not only directly plays an essential role to guide credit decisions, but indirectly influences future returns. Pressure from regulators and the mind-set to critically assess the credit risk has become important. Pressures for increased revenue generation and increased competition leads financial institutions to be more effective to attract creditworthy clients.

A thesis outline, research design and procedure was described to carry out solving the problems stated in this thesis. Models, methodologies and results presented in this thesis all aim to enhance credit risk measurement and management with the objective of implementation in the retail banking environment.

The three chapters that follow address credit scoring related problems. Each comprises a literature review and methodology employed. The thesis concludes with a summary of findings, conclusion and recommendations.

Chapter 2
An optimised credit scorecard to enhance cut-off score determination

Nico Kritzinger[1] and Gary van Vuuren[2]

ABSTRACT

Literature in the world of credit scoring is limited. In this paper parametric and non-parametric statistical techniques which – used in credit scoring – are reviewed with the aim to build an optimal credit scoring matrix model to predict which clients will go bad in the future. The matrix provides uplift in the Gini-coefficient when compared to a one-dimensional model. There has been disagreement on the most appropriate cut-off score as indicated in the literature; this article also illustrates the use of the credit scoring matrix model to determine an appropriate cut-off score on a more granular level.

Keywords: Credit risk, credit scoring, credit risk management

JEL classification: C20, G32

2.1 Introduction

One of the most important elements driving a bank's existence and continuance is the ability to grant credit to the appropriate client that is less likely to go bad. Two general techniques exist to assess possible future bads; firstly this can be done subjectively by a credit manager or loan officer, and secondly it can be accomplished objectively by means of credit scoring (Allen, De Long, Saunders, 2004, pp. 734 – 736). Credit scoring (which surfaced in the mid-1900s) is a statistical tool allowing banks to distinguish between good and bad clients. In banking especially credit scoring has grown in the last 25 years (2015) mainly due to a wider range of banking products and large number of credit applications (Abdou & Pointon, 2011, p. 65). The general idea behind credit scoring is to use client characteristics from the past to predict whether the client to be financed would either be good or bad in the future. Credit scoring has proved its worth in credit evaluation, however the literature has been limited (Abdou &

[1] PhD-student at the Faculty of Natural Sciences, North-West University.
[2] Visiting professor, Faculty of Natural Sciences, North-West University.

Pointon, 2011, p. 60). Various modelling techniques exist to model credit scoring with most of them being statistical e.g. linear regression, discriminant analysis, probit analysis, logistic regression, decision trees, expert systems, neural networks and genetic programming (Abdou & Pointon, 2011, pp. 66-68). Given these range of statistical techniques, however, no optimal technique for scorecard construction exists (Abdou & Pointon, 2011, p. 68).

Various benefits exist for using scorecards for credit assessment e.g. less information is required to make a decision. Credit scoring removes the *bias* by not only looking at accepted applicants, but by considering the entire application population, credit scoring considers characteristics of both good and bad applicants, credit scorecards are based on large data samples, credit scorecards include legally acceptable characteristics and credit scorecards demonstrate correlation between variables and repayment/bad behaviour. Credit scorecards also include a large number of characteristics, enjoy efficient processing time, minimise process time and cost and produce fewer errors. Credit scorecards are based on real historical data and the interrelation between characteristics is considered (Abdou & Pointon, 2011, pp. 61-62). Given these benefits, however, various criticisms also exist for using scorecards for credit assessment e.g. no economic factors are considered, there are misclassification problems, any characteristic is up for consideration, indirect discrimination is possible, no standardisation exists across the market, training analysts is expensive, the final model is statistically "incomplete" as not all characteristics are part of the final model, the data are historical, characteristics are assumed to be constant over time and credit scoring imposes a dichotomous outcome (Abdou & Pointon, 2011, pp. 62-63).

Credit scoring – which allows banks to determine a level of risk of an applicant or borrower – is based on statistical numerical scores to determine whether they would either be a good or bad client in the future (Siddiqi, 2006, p. 5). In the credit scoring milieu, two types of credit scoring models have been employed, namely: application scoring and behavioural scoring (Lim & Sohn, 2007, p. 427). Application scoring (AS) is used where the scorecard is built for a specific credit organisation, by utilising the credit organisation's historical data. It is a credit assessment performed at the application stage. Behavioural scoring (BS) indicates the way a borrower's characteristics of payment behaviour changes after the loan is made, and is based on time dependent characteristics of borrowers (Lim & Sohn, 2007, p. 427).

Another form of scoring that exists is the credit bureau score (CBS) which can be described as an economic barometer of the way the borrower performs with other organisations. The CBS is a representation of how borrowers manage his existing credit obligations and the CBS is an external score with *external* meaning that the scorecard was built by the bureau using data from various credit organisations. As with all scorecards, the credit bureau scoring model is constructed using history, but it does consider the probability of change in behaviour when the score is updated. The model is a dynamic one in this sense, but becomes historic in actual use. One of the key strengths of the CBS is the relative complete nature of economic information on which it is based, i.e. it consists of external risk-related data from diverse businesses, contributing information on their own debtors to the credit bureau. In this manner a borrower is assessed on the actual credit performance with all his or her accounts with different institutions such as banks, finance houses, service providers, credit card companies, retailers, public authorities, tax registers, etc. (Anderson, 2007, p. 285 and TransUnion, 2015). Anderson (2007) also indicates the importance of credit bureaus: the facilitation of information gathering from public sources especially court orders and the sharing of borrower performance information (Anderson, 2007, p. 10). The private nature of this kind of information and their ability gives a monetary incentive to credit bureaus to collect, to record and to exchange reliable, valuable and up to date information on all credit performance (TransUnion, 2015).

In South Africa there are four major credit bureaus: TransUnion, Experian, Compuscan and XDS which capture, update and store the credit histories of the credit active consumers in South Africa. The credit bureaus obtain and build their scorecards on data regularly supplied by the credit lenders to the credit bureaus (TransUnion, 2015). Transunion emphasise that to reduce exposure to risk a predictive scoring system is needed (TransUnion & FICO, 2009, p. 1). The bureau score of Transunion is referred to as the Empirica score which aims to assess risk at the origination stage, evaluate clients risk in conjunction with expected performance, assess clients with no historical credit history, manage existing clients from a credit limit and collections perspective, rank clients according to credit risk and identify cross-sell opportunities and to grow the client base. The Empirica score consists of five categories to assess the credit risk of a client: demographics, judgments of the client, defaults experiences of the client, payment profiles and enquiries of the client (TransUnion & FICO, 2009, p. 2).

As the objective of credit scoring is to distinguish between good (usually up to date or 1 payment in arrear) and bad (usually three or more payments in arrears, in legal or written-off) clients the problems that arise from credit scoring would be related to classification (Abdou & Pointon, 2011, p. 66). Problems that exist in credit scoring include the lack of a theoretical reason as to why characteristics are chosen in the final scorecard. Siddiqi (2006) gave some guidelines regarding number of variables (characteristics) to be used within scorecards. Scorecards should comprise between eight to 15 characteristics to ensure a stable scorecard as the predictive power will remain strong even if the profile of one or two variables change. Scorecards containing too few variables are more susceptible to minor changes from the applicants profile making the scorecard unable to withstand the test of time (Siddiqi, 2006, p. 88). Anderson (2007) mention that although a huge number of variables can potentially be used for credit scorecard developments that usually only between six and 15 variables best explain consumer behaviour (Anderson, 2007, p. 393). However, it is generally believed that there is no optimal number of variables to be used in scorecards (Abdou & Pointon, 2011, p. 67).

In general, two scenarios which will be explained in the next section are considered when implementing a credit scorecard, i.e. keeping at the same reject rate or approval rate or keeping the same bad rate. However there is disagreement on the appropriate cut-off score when assessing credit at implementation and the determination of the sample size when building a scorecard (Abdou & Pointon, 2011, pp. 67-68). In addition there is no optimal technique when building scorecards (Abdou & Pointon, 2011, p. 68).

2.2 Credit scorecard implementation strategies

Once an organisation has its credit scorecard built it is mostly used to set minimum levels of score which represents a threshold of risk called the cut-off score. This cut-off score could represent a level of accepting clients, profit level or any other level based on the objectives of the organisation (Siddiqi, 2006, p. 146). A simple illustration of implementing a credit scorecard with a cut-off score strategy can be presented by Figure 2.1.

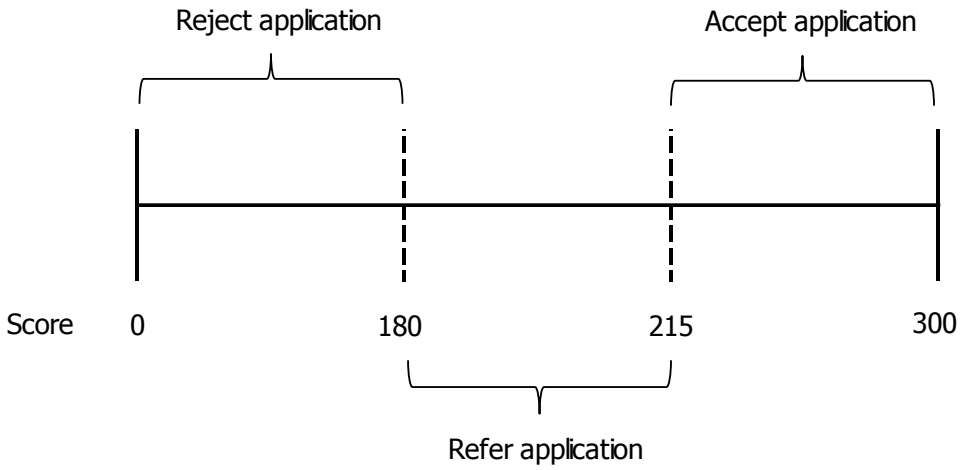


Figure 2.1: Simple cut-off score strategy.

From Figure 2.1 the implementation strategy indicates that every client scoring below 180 will be rejected, in other words not being considered, a score above 215 will be accepted and clients scoring between 180 and 215 will for example be referred to credit managers for further analysis to determine whether the client will be rejected or accepted.

In general when building a new credit scorecard it should outperform the previous credit scorecard which should lead to an outcome of a higher approval rate when the same bad rate is maintained or leading to a lower bad rate when the same current approval rate is maintained (Siddiqi, 2006, p. 148). This describes the two general scenarios when implementing a credit scorecard; firstly the lender may wish to implement its credit scorecard such that the lender keep at the same reject rate or approval rate of clients to reduce bad rates in the future. This first scenario could be chosen when the company wishes to be conservative where the aim is better risk management in a tough economic environment. Secondly the lender may wish to keep at the same bad rate to gain market share to expand the business, this scenario could be chosen by the company in a competitive environment. According to Anderson (2007), one of the most obvious approaches when setting a cut-off score was where any account was accepted which provides a profit. Since the introduction of credit scoring, lenders have gained experience to apply scientific approaches to enhance the business, use credit scoring for forecasting and portfolio valuation, take potential profitability into consideration, use credit scoring for risk-based pricing, account management and incorporate other aspects of borrower behaviour such as response, revenue and retention. In the event of choosing the cut-off score strategy it should be noted that several assumptions were made during the

credit scorecard development and a credit scorecard is a tool that must fit in with a company's general strategies (Anderson, 2007, pp. 67, 240). Credit scorecards play an important role when distinguishing good clients from bad clients, however there has been disagreement on the most appropriate cut-off score when strategically implementing a credit scorecard. It is commonly known that there is no optimal cut-off score decision which is differently derived based on the environment and country (Abdou & Pointon, 2011, p. 67).

2.3 Problem statement and objective

The aim of this article is to build a Credit Scoring Matrix Model (CSMM) to address the issues raised especially obtaining an optimal model (most favourable) and determining the more appropriate cut-off score when assessing credit of borrowers at the application stage of a loan. The more appropriate cut-off score, in this sense, refers to the cut-off made from the CSMM. The primary objectives of this study are to improve credit risk measurement and management in the world of credit scoring and accurately determine the appropriate cut-off score. The focus includes credit risk, credit scoring and credit risk management. Historical credit scoring modelling techniques are evaluated and the effectiveness of proposed primary objectives on credit risk management are assessed with the aim of implementing these in the retail banking environment.

2.4 Literature review

Statistical techniques used in credit scoring can be divided into two categories namely non-parametric techniques and parametric techniques. Non-parametric statistical techniques do not require many assumptions about the underlying data, if any, whereas parametric statistical techniques require several (Anderson, 2007, p. 172). This section gives an overview of the different types of non-parametric and parametric statistical techniques used in the credit scoring world.

2.4.1 Non-parametric statistical techniques

2.4.1.1 Expert systems

Expert systems involve the use of expert judgment or human expert knowledge to solve problems and explain the outcomes as to why certain credit applicants are rejected. Abdou and Pointon (2011) presented the three components of an expert system: relying on facts and rules, an interface communicating the expert's conclusion and updating the expert's decisions

and recommendations (Abdou & Pointon, 2011, p. 72). Disadvantages of expert systems include subjectivity, inconsistency and individual expert preferences. Advantages of expert systems include the use of qualitative characteristics within its judgmental evaluation and the vast experience of the expert from the past (Abdou & Pointon, 2011, p. 61).

2.4.1.2 Decision trees

Thomas et al. (2002) mentioned one of the earliest decision trees which was a type of expert system with a rule set (Anderson, 2007, pp. 172-173). A decision tree or recursive partitioning analysis (RPA) is defined as a classification technique where a dependent variable is analysed as a function of continuous explanatory variables (Abdou & Pointon, 2011, p. 71). A decision tree comprises three nodes with the root node the start (at the top of the tree) followed by subsequent levels called the child nodes and the bottom of the decision tree referred to as the terminal nodes. The aim of a decision tree is to indicate possible turn of events or consequences with each event indicating the outcome. Figure 2.2 presents a simple example of a decision tree.

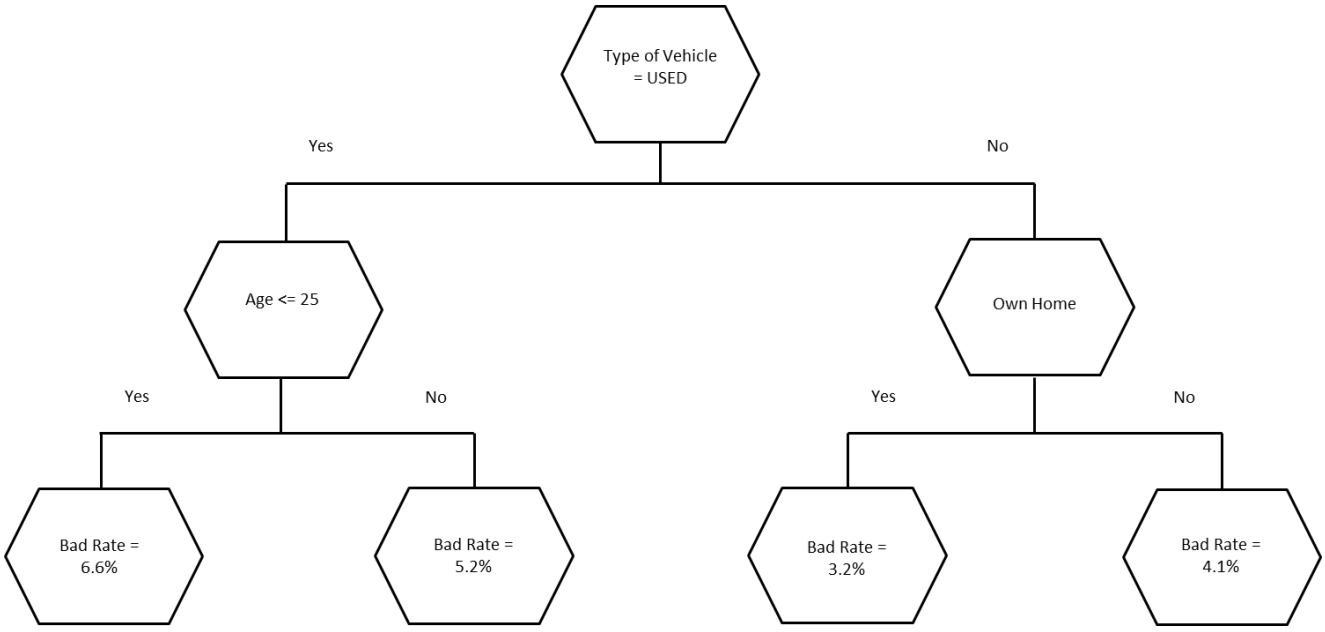


Figure 2.2: Decision tree.

Advantages of decision trees include ability to identify patterns, finding and exploiting interactions, results are transparent and easy to implement, computationally simple and quick and easy identification of extremely high and low risk categories. Disadvantages include decision

trees becoming too busy which could result in over fitting and unreliable results. Where interactions are not an issue regression techniques provide better results and decision trees prove to be relatively inflexible (Anderson, 2007, pp. 172-174).

2.4.1.3 Neural networks

Neural networks (NNs) are structures which allow training processes to take place in which the linear and non-linear variables help to distinguish variables to obtain better decision making results. Andou and Pointon (2011) alluded to the use of NNs which could be successful in credit card fraud, bankruptcy prediction, bank failure prediction, option pricing and mortgage application (Abdou & Pointon, 2011, p. 73). Disadvantages of NNs include computational intensity, multiple iterations are required before a final model is obtained, it is expensive to implement and maintain, they are opaque and there is always a significance chance of over fitting. In credit scoring NNs are seldom used, but have advantages where there are fewer data. NNs are also used for fraud scoring (Anderson, 2007, p. 175). Compared to more generally-used techniques, such as discriminant analysis and logistic regressions, NNs have the highest average correct classification rate and statistical measures indicate that they represent the data better than logistic regression (Abdou & Pointon, 2011, p. 73).

2.4.1.4 Genetic algorithms

Genetic algorithms – first proposed in the 1960s – are the use of genetic operations to transform data according to fitness value (Anderson, 2007, p.176; Abdou & Pointon, 2011, p.74). Genetic algorithms have been used in financial services, computer sciences and engineering. Disadvantages of genetic algorithms include computational problems and slower run times. An advantage of genetic algorithms is that alternate solutions may be obtained when they are not readily apparent. The primary uses for genetic algorithms are; providing an exhaustive search if many possible solutions are possible, the aim is optimisation and not necessarily the best model, identifying good solutions which are not easy to find and when there are multiple targets. Genetic algorithms work best in a rapidly changing environment where new solutions are to be found (Anderson, 2007, p. 176). Table 2.1 summarise the assumptions for the non-parametric techniques.

Table 2.1: Non-Parametric Techniques Assumptions.

Non-Parametric Techniques (Few assumptions if any)				
Assumption no.	**Expert Systems**	**Decision Trees**	**Neural Networks**	**Genetic Algorithms**
1.	Non-statistical approach that makes no assumptions	Non-statistical approach that makes no assumptions	Non-statistical approach that makes no assumptions	Existence of a fitness function which is to be maximised
2.	*Primitive form of a decision tree*			A mapping from bit-strings to potential solutions

2.4.2 Parametric statistical techniques

2.4.2.1 Discriminant analysis

Discriminant analysis is another statistical technique where the aim is to determine group membership where there are two or more known groups (Anderson, 2007, p. 169). Andou and Pointon (2011) indicated the earliest proposal of discriminant analysis where multiple discriminant analysis was applied while examining car loan applications (Abdou & Pointon, 2011, p. 69). Discriminant analysis uses a classification tool to minimise the distance between cases in a group and maximising the differences between cases from different groups. The problem with discriminant analysis is that it suffers from all the assumptions associated with the statistical technique used. Linear discriminant analysis is the most common which suffers from high misclassification errors when predicting rare groups (Anderson, 2007, pp. 169-170). Problems such as using linear functions instead of quadratic functions, group definitions, inappropriateness of prior probabilities and classification error prediction also surface when using discriminant analysis and multivariate normal distributions and equal variances are assumed (Abdou & Pointon, 2011, p. 70).

2.4.2.2 Probit analysis

Probit analysis aims to transform a linear combination of independent variables into its cumulative probability value from a normal distribution. Under probit analysis normal distributions of the threshold values are assumed and the coefficient estimates can be tested individually for significance using a likelihood ratio test which is not possible within discriminant

analysis. A problem with probit analysis is that multicollinearity can cause incorrect signs for the coefficients which are not an issue within discriminant analysis applications (Abdou & Pointon, 2011, p. 70).

2.4.2.3 Linear regression

Galton (1889) introduced linear regression, a statistical technique to explain linear relationships described by:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_j x_{jt} \quad (1)$$

where

$y_t =$ dependent variable OR endogenous variable value at time t,

$x_{jt} =$ independent variable OR exogenous variable j at time t, and

$\beta_j =$ coefficient OR change in predicted value y per unit of change in x_j at time t.

In (1) the dependent variable y is predicted by using the values of independent variables x_j. The prediction of y is determined by calculating the coefficients β by minimising the sum of the squared error terms. The problem with linear regression is that it makes numerous assumptions such as linearity, homoscedasticity, a normally distributed error term, independent error terms, additivity, uncorrelated predictors and the use of relevant variables (Anderson, 2007, pp. 166-167).

2.4.2.4 Logistic regression

Logistic regression is one of the widely used statistical techniques used in credit scoring. The difference between linear regression and logistic regression is that with the latter the outcome variable is binary, i.e. 1 or 0, good or bad etc. (Abdou & Pointon, 2011, p. 71). Anderson (2007) indicated that investigations on human populations were the roots of where logistic regression originated (Anderson, 2007, p. 171). Hand and Henley (1997) indicted that a comparison between discriminant analysis and logistic regression in the world of credit scoring concluded that logistic regression gave superior classification results (Hand & Henley, 1997, p. 533).

A logistic regression function can be presented as:

$$\ln\left(\frac{p(G)}{1-p(G)}\right) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_j x_{jt} \quad (2)$$

where

$p(G) =$ probability of being good,

$\beta =$ coefficients, and

$x =$ independent variables.

In (2) the left-hand side represents the natural logarithm of the odds of being a good client. For logistic regression the regression coefficients (i.e. the β coefficients) in (2) is obtained by using the maximum likelihood (ML) method (Harrell, 2015, p. 220). ML is a general statistical technique to estimate parameters and making statistical conclusions in various situations (Harrell, 2015, p. 181). Logistic regression follows assumptions which include categorical target variable, log odds linear relationship function, independent error terms, uncorrelated error terms and the use of relevant variables. A disadvantage of logistic regression is its intense computations, however logistic regression has been the primary choice for building credit scoring models because it predicts a binary outcome, the final probability cannot fall outside of the boundaries 0 to 1 and logistic regression provides robust estimates of the actual probability (Anderson, 2007, pp. 170-171). Table 2.2 summarise the assumptions for the parametric techniques.

Table 2.2: Parametric Techniques assumptions.

Parametric Techniques (Lot of assumptions)				
Assumption no.	**Discriminant Analysis (DA)**	**Probit Analysis**	**Linear Regression**	**Logistic Regression**
1.	*Suffer from any and all assumptions associated with the statistical technique used*	Linearity with log odds function	Linearity	Linearity with log odds function
2.	Linear DA was the original, however logistic regression DA is now preferred due to fewer assumption violations		Homoscedasticity	
3.		Normally distributed error terms	Normally distributed error term	Logistic distributed error terms
4.		Independent error terms	Independent error terms	Independent error terms
5.		Additivity	Additivity	Additivity
6.		Uncorrelated predictors	Uncorrelated predictors	Uncorrelated predictors
7.		Use of relevant variables	Use of relevant variables	Use of relevant variables

2.5 Data and methodology

2.5.1 Data

Data that were used in the research study are based on a bank in South-Africa and would be on retail banking specific. External credit bureau data were also used where applicable and were based on credit bureau data South-Africa specific. Table 2.3 presents the frequency and source data used in more detail.

Table 2.3: Data used for Research Study.

Topic	Data requirement	Frequency	Source
Scorecard development and CSMM	Historical client data with performance	Monthly data	Retail bank in South Africa
Bureau data	Historical client data with performance	Monthly data	South-African credit bureau

Data availability for Table 2.3 includes data from January 2002 up to and including December 2014.

2.5.2 Methodology

The Credit scoring matrix model (CSMM) may be split up into two components, firstly an internal application scorecard which is built from internal information and secondly a credit bureau score obtained from credit bureaus. The methodology which is required to construct the internal application scorecard is presented by Figure 2.3 which is further explained in the sections that follow.

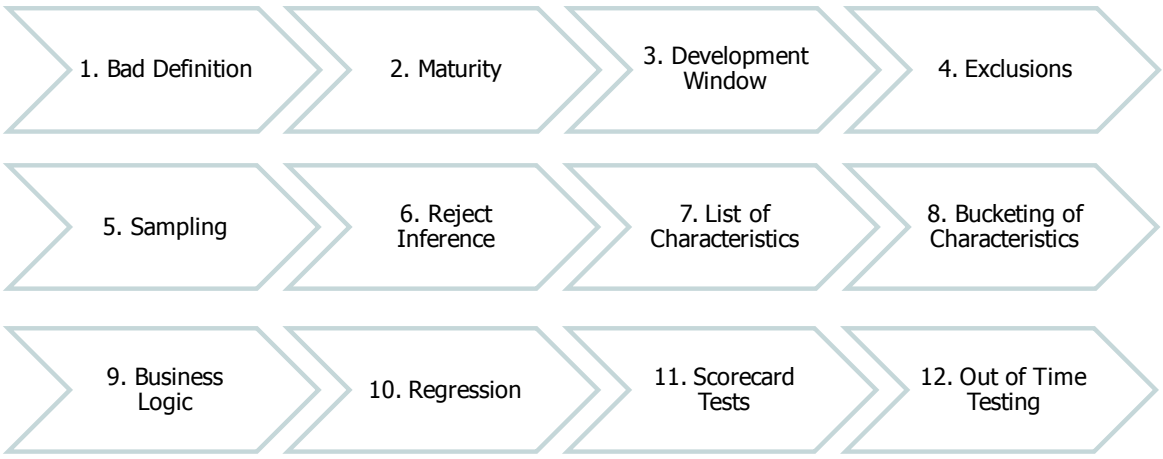


Figure 2.3: Scorecard Methodology.

2.5.2.1 Bad definition

A client's historical performance can be divided into three categories namely "good", "bad" and "Indeterminate" (Anderson, 2007, p. 138). A client which is seen as a "bad" can be based on several considerations such as organisational objectives, product, being risk-averse, being risk-seeking, an easily interpretable definition, accounting policies on write-offs, consistent definitions or regulatory requirements (Siddiqi, 2006, pp. 38-40). Classifying a client as a "good" client requires the same considerations as mentioned as when classifying a client as a

"bad" client. A client is classified as "indeterminate" if the client does not conclusively fall in either the "bad" or "good" category. As a rule of thumb "indeterminates" should not exceed 10% to 15% of the development sample used to build the scorecard (Siddiqi, 2006, pp. 43-44).

2.5.2.2 Maturity

Various methods exist to obtain comfort that the correct "bad" definition is determined. Firstly we have the analytical methods such as roll rate analysis (Comparing delinquency buckets from the past with delinquency buckets observed currently and then calculating which clients remained in the same delinquency bucket, cured to a better bucket and which rolled to a worse delinquency bucket) (Siddiqi, 2006, p. 41) and current versus worst delinquency comparison (comparing the worst ever delinquency status with the most current delinquency status) (Siddiqi, 2006, p. 42). Secondly, the consensus method where various stakeholders from the credit risk department, operational departments or other areas comes together to obtain consensus on the definition of a "bad" client (Siddiqi, 2006, pp. 40-43). One can also use the maturity approach where an outcome is determined which is sufficient to capture the "bad" clients, e.g. of all clients booked at a specific vintage date which of them were "bad" clients in the next "X" number of months (Siddiqi, 2006, p. 77). Although one would wish to use as many outcomes as possible to capture most of the "bad" clients, the current portfolio of the bank could be vastly different from the portfolio in earlier years.

2.5.2.3 Development window

Building a scorecard requires the identification of two windows, firstly the development or observation window which is the period where the data are observed and secondly the outcome or performance window which measure the level of clients being "bad" after a certain time period have occurred for clients from the observation window (Anderson, 2007, pp. 77, 344). Siddiqi (2006: 35) mentions that the development window should be chosen where the level of bad rate is deemed to be stable, however Anderson (2007: 344) indicated that one could come at a point where applications is not a representation of today's applications the longer the outcome period as also mentioned in Section 2.5.2.2.

2.5.2.4 Exclusions

Data used to build a scorecard should contain only information on which one intends to use them. Clients such as frauds, staff, deceased, out of country, preapproved or underage should

be excluded when building a scorecard because these clients are given finance based on a non-score-dependent criteria. Any client that is not a normal client or is not going to be scored should be excluded from the scorecard development (Siddiqi, 2006, p. 31; Anderson, 2007, p. 338).

2.5.2.5 Sampling

To ensure that a group of representative clients are used during the scorecard development sampling is used (Anderson, 2007, p. 61). It is the general believe that a minimum of 1 500 bads, 1 500 goods and 1 000 rejects are used when developing a credit scorecard. These minimum numbers were derived in the 1960s as the work required to obtain data was demanding and sufficient computer power was lacking (Anderson, 2007, p. 350). Two steps are required during the sampling phase when developing a scorecard, first data are sampled which represents the client population (Siddiqi, 2006, p. 63). Generally, the total client population is sampled into a smaller dataset especially when working with large amount of data which can take longer to process. Secondly the representative total sample is split up into a development sample on which the scorecard is built and a validation sample on which the scorecard is tested (Siddiqi, 2006, p. 63). It is important to note that random samples can pose difficulties as important subgroups of the total sample must also be adequately represented. To achieve this a technique such as over-sampling can be used. Over-sampling in essence adjust the subgroups distribution against the total sample (Anderson, 2007, p. 351).

2.5.2.6 Reject inference

Application scorecards are built based on the through the door population known as the "All Good Bad (AGB)" scorecard, i.e. all applications not just accepted applications (Siddiqi, 2006, p. 101). If application scorecards are built using only the accepted clients (known as the "Known Good Bad (KGB) scorecard") there will be bias to accepted clients from the past when the new application scorecard is applied as the rejected clients from the past was ignored during development (Anderson, 2007, p. 65). Reject inference accounts for influence from past decision making during the AGB scorecard development process. As an example, reject inference accounts for "cherry-picked" applicants, e.g. suppose 100 out of 1 000 applications have a very high delinquency, credit managers decline 90 of the 100 and accept 10 with subsequent performance indicating that the 10 accepted applicants perform well and is classified as a good account. Building only a KGB scorecard would see these high delinquency accounts

as good applicants. Reject inference accounts for these "cherry-picked" accounts. From a decision making view reject inference provides expected performance based on all applications, i.e. the through-the-door (TTD) population. Suppose a bank rejects all applications below a score of 180, however the bank feels it has been too conservative and now rejects all applications below a score of 165. If the bank has never accepted these applicants in the past how will the bank account for this additional risk being taken when moving the cut-off? Reject inference accounts for this through the estimated bad applicants for the rejected applications. It is important to realise that with the reject inference process there will always be a level of uncertainty, however can be minimised using better reject inference techniques. Reject inference lead to better decision making, however is not 100% accurate (Siddiqi, 2006, pp. 99-101). Various rejection inference techniques exist such as random supplementation, augmentation, extrapolation, cohort performance and bivariate two-step (Anderson, 2007, p. 79).

2.5.2.7 List of characteristics

A scorecard consists of a group of characteristics to separate good clients from bad clients (Siddiqi, 2006, p. 5). Characteristics to be included in the development sample are an important step when building a scorecard and business input are recommended when characteristics are included. Characteristics should be selected based on: Expected predictive power, reliability and robustness, ease in collection, interpretability, human intervention, legal issues surrounding the usage of certain types of information, creation of ratios based on business reasoning, future availability and changes in competitive environment (Siddiqi, 2006, pp. 60-62).

2.5.2.8 Bucketing of characteristics

Bucketing of characteristics before the regression step has advantages such as: it is easier to deal with outliers, relationships are easier to understand, nonlinear dependencies can be modelled with linear models, unprecedented control in the development process of the scorecard is allowed, increased knowledge of the portfolio and the user may develop insights into the behaviour of risk predictors (Siddiqi, 2006, p. 78).

The predictive power of each characteristic attribute (a group of attributes form a characteristic) after bucketing can be determined by weights of evidence (WoE) presented by (3) for each attribute i:

$$\left[\ln\left(\frac{\text{Distribution}_{Good_i}}{\text{Distribution}_{Bad_i}}\right)\right] \times 100 \tag{3}$$

The predictive power of each characteristic is measured by the information value (IV) presented by (4):

$$IV_{Characteristc} = \sum_{i=1}^{n}\left(\%Distribution_{Good_i} - \%Distribution_{Bad_i}\right) \times \ln\left(\frac{\%Distribution_{Good_i}}{\%Distribution_{Bad_i}}\right) \tag{4}$$

Where $\%Distribution_{Good_i}$ is the distribution of the good clients for attribute i, $\%Distribution_{Bad_i}$ is the distribution of the bad clients for attribute i. The IV of each characteristic is interpreted as follows:

- $IV < 0.02$ Unpredictive
- $0.02 \leq IV < 0.1$ Weak
- $0.1 \leq IV < 0.3$ Medium
- $IV \geq 0.3$ Strong

(Siddiqi, 2006, p. 78-79 and Anderson, 2007, p. 192-193).

2.5.2.9 Business Logic

WoE and IV are used as statistical measures when bucketing characteristics as discussed in Section 2.5.2.8, however business logic also needs to be considered. Suppose a company under its policy rules refer clients to the credit managers that have a debt service ratio greater than 50%. Then the debt service ratio if a characteristic in the scorecard should be bucketed with a break at 50% to minimise the distortion of the policy rule on the scorecard as the clients affected by the policy rule are now isolated (Siddiqi, 2006, p. 87).

2.5.2.10 Regression

The pioneers of credit scoring are Fair and Isaac who started their consultancy, Fair Isaac, in 1956 (Anderson, 2007, p. 40). In recent times Fair Isaac has become known as FICO. In Section 2.4 various statistical techniques were described which may be used in the world of credit scoring. FICO in addition have their own methods which is part of their modelling software called FICO Model Builder (FICO MB7) and one of these methods is called “Scorecard – Divergence”. The divergence statistic measures the distance between the scores of the bad client’s

distribution and the scores of the good client's distribution and is presented by (5) (Anderson, 2007, pp. 189-190):

$$\text{Divergence} = \frac{(Mean_{Goods} - Mean_{Bads})^2}{\left[\frac{(Variance_{Goods} + Variance_{Bads})}{2}\right]} \quad (5)$$

The "Scorecard-Divergence" method in FICO MB7 is a generalised additive model of bucketed predictors which provides pattern constraints and a penalty term to reduce over fitting and smooth weight patterns. The fitting objective function optimises the model weights to maximise the divergence between binary outcome classes, subject to user defined constraints (FICO, 2014).

2.5.2.11 Scorecard Tests

Various scorecard measurements can be carried out to determine the predictiveness of the scorecard such as Akaike's Information Criterion (AIC), Schwarz's Bayesian Criterion (SBC) and the Kolmogorov-Smirnov (KS) statistic. The most common measurement used is called the Gini-coefficient which provides a single value representing the predictive power of the scorecard over the entire range of possible scores (Anderson, 2007, p. 205). Statistics such as correlation between characteristics scores, the population stability index (PSI) which measures the stability of the scorecard are also analysed when building a scorecard (Anderson, 2007, pp. 194-200). A PSI less than 0.1 indicates no significant change from development, a psi between 0.1 and 0.25 indicates a small change in distribution from development that needs investigation and a PSI greater than 0.25 indicates a significant shift from the development population (Siddiqi, 2006, p. 137).

2.5.2.12 Out-of-Time Testing

The out-of-time window is a validation period which falls outside of the development window of the scorecard (Anderson, 2007, p. 78). The out-of-time window is a validation period where the scorecard tests as mentioned in Section 2.5.2.11 can be carried out to test the scorecard on a period that did not form part of the scorecard building development window.

2.6 Results

The CSMM build consisted of two components namely the internal application scorecard and the credit bureau score. The first part of this section will present the results obtained when building the internal application scorecard after which the credit bureau score (Empirica

score) will be added as a second dimension to illustrate the optimal model. The section concludes with the illustration of improved cut-off score determination from the CSMM.

2.6.1 Internal Application Scorecard

The methodology followed when building the internal application scorecard was described in Section 2.5.2. Firstly the "bad", "indeterminate" and "good" definition is presented in Table 2.4 and maturity analysis in Figure 2.4.

Table 2.4: Internal Application Scorecard "Bad" definition.

	Definition
Bad	An account is seen to be bad if it is more than three payments in arrears ever and the balance outstanding is greater than ZAR100, OR if the account is in Legal OR if the account is written-off
Indeterminate	An account is seen to be indeterminate if it is ever two payments in arrears and the balance outstanding is greater than ZAR100
Good	An account is seen to be good If it is not "bad" or "indeterminate"

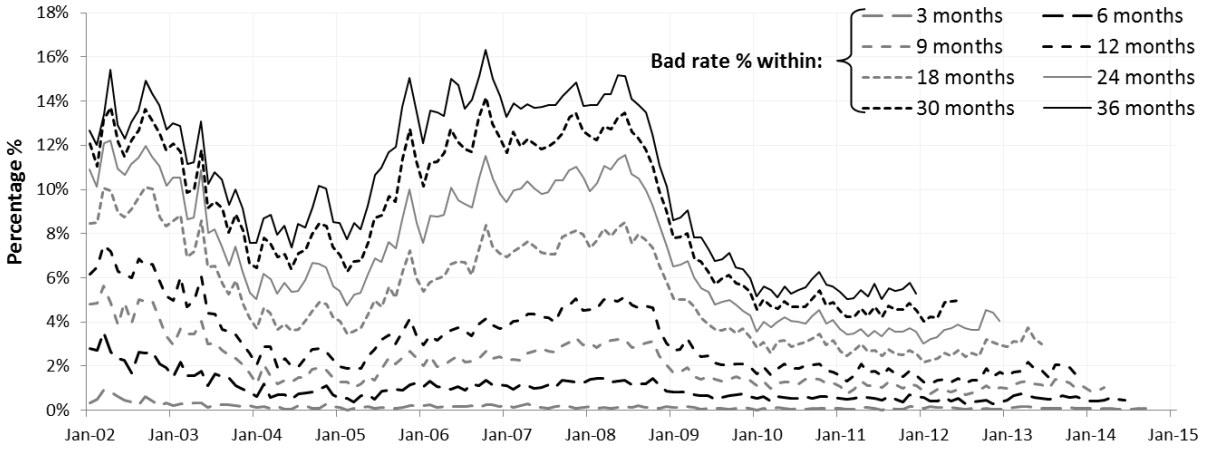


Figure 2.4: Maturity analysis.

Figure 2.4 presents the bad rate % for each vintage date booked clients, i.e. for booked clients at a specific date what percentage went bad after 3 months, 6 months, 9 months, 12 months, 18 months, 24 months, 30 months and 36 months. The 24-month outcome was used when building the internal application scorecard based on two reasons: firstly capturing as much bad clients as possible is desired and secondly too much history is undesirable as applications at development could not represent the applications of today. Based on the "Bad" outcome

decision and data availability up to and including December 2014 (as mentioned in Section 2.5.1) the development window and out-of-time window are presented in Table 2.5.

Table 2.5: Internal Application Scorecard Build Windows.

Window	Definition
Develop-ment	January 2011 up to and including December 2011 (80% development, 20% hold-out)
Out-of-Time	January 2012 up to and including December 2012

Exclusions used in the development include company clients as the scorecard is intended for individual consumers, staff clients, unemployed clients, clients with missing identification numbers (ID), applicant age less than 18, non-South-African citizens, loan amounts larger than R500 000 (as this is not normal business for the company), frauds, deceased clients, application disputes and pending decisions. Exclusions were $\sim 11\%$ for the development window. Sampling was performed using a simple random sampling technique in the statistical analysis software (SAS) and Figure 2.5 presents the bad rate percentage for accepted clients between the total population after exclusions and the sample after exclusions.

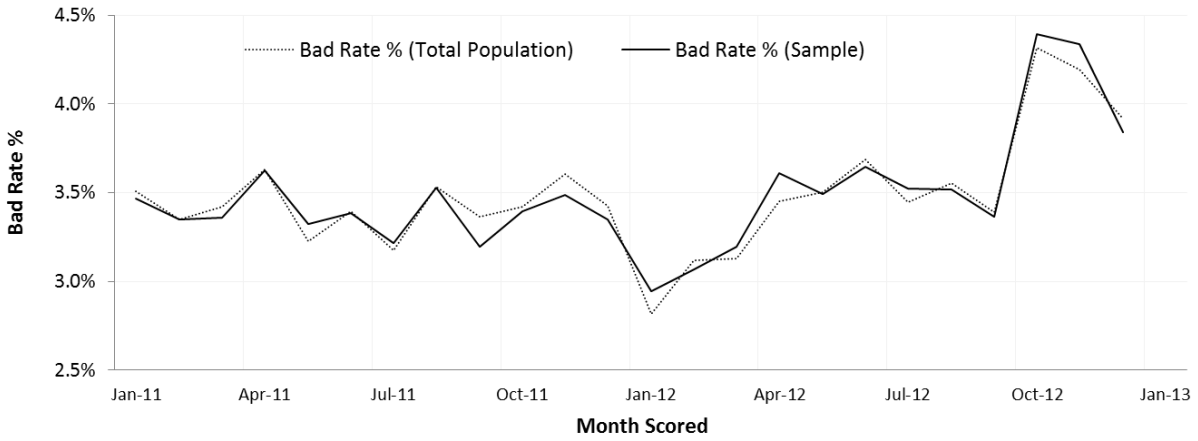


Figure 2.5: Accepted clients' bad rate (%).

Reject inference (fuzzy augmentation technique) was carried out by using the bureau score at outcome from the credit bureau TransUnion to infer the rejected applications. The Gini-coefficient for the bureau score at outcome was determined as $\sim 81\%$ indicating as a good proxy to infer the rejected applications. The bad rate after inference for the development window was calculated as 6.3%. Figure 2.6 presents the rejects inference results.

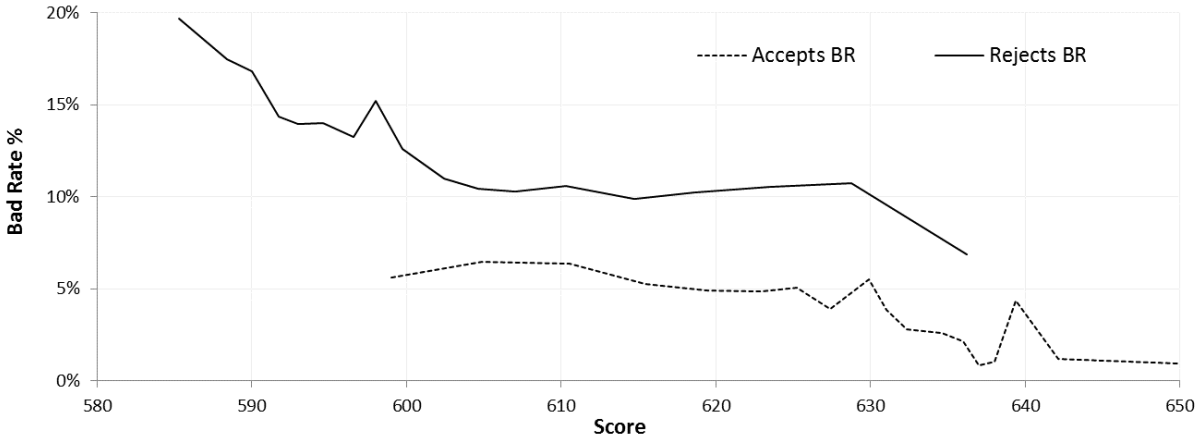


Figure 2.6: Bad Rate for Accepts & Rejects (@Observation).

Figure 2.6 illustrates the higher bad rate for rejected applications after reject inference (as expected). Characteristics considered in the development of the internal application scorecard included 32 characteristics consisting of both borrower and transaction type characteristics. The bucketing of the characteristics was done in the FICO MB7 software as follows:

- Step 1: Firstly auto binning ("Coarse Fine Supervised") in FICO Model Builder was carried out
- Step 2: Secondly after auto-binning, bucketing was carried out such that:
  - There should be a logical relationship between the bad rate and the characteristic
  - Buckets must contain at least 2% of total and not less than 5 bads per bucket
  - A tail-end bucket can contain around or less than 2% of total population of the characteristic if the characteristic has at least 3 buckets
- Step 3: Apply business rationale

After the bucketing process regression was carried out and based on the “Scorecard-Divergence” method from FICO. Table 2.6 presents the predictiveness statistics of the internal application scorecard from the FICO MB7 software.

Table 2.6: Internal application scorecard performance measures results.

	Development window	**Hold-out**
Divergence	0.316	0.253
Gini-coefficient	0.305	0.274
KS-statistic	21.838	19.843

Table 2.6 presents predictive results for both the development sample as well as the hold-out or validation sample from FICO Model Builder. Table 2.6 indicates a Gini-coefficient of 30.5%, which excludes bureau information and is based only on internal company information. No correlation between characteristics was observed as indicated in Table 2.7.

Table 2.7: Correlation (development).

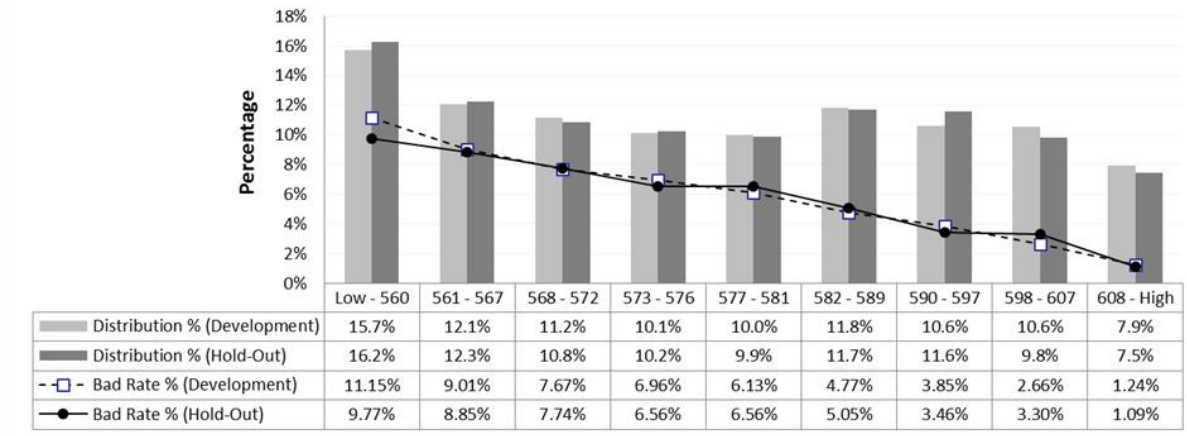
DEVELOPMENT

	Characteris-tic 1	**Characteris-tic 2**	**Characteris-tic 3**	**Characteris-tic 4**	**Characteris-tic 5**
Characteristic 1	100%	-4%	8%	3%	4%
Characteristic 2	-4%	100%	20%	13%	10%
Characteristic 3	8%	20%	100%	25%	6%
Characteristic 4	3%	13%	25%	100%	3%
Characteristic 5	4%	10%	6%	3%	100%

HOLD-OUT

	Characteris-tic 1	**Characteris-tic 2**	**Characteris-tic 3**	**Characteris-tic 4**	**Characteris-tic 5**
Characteristic 1	100%	-2%	7%	1%	6%
Characteristic 2	-2%	100%	19%	12%	11%
Characteristic 3	7%	19%	100%	26%	7%
Characteristic 4	1%	12%	26%	100%	4%
Characteristic 5	6%	11%	7%	4%	100%

The score trend is a summary of grouped scores equally distributed as far as possible and is presented in Figure 2.7.



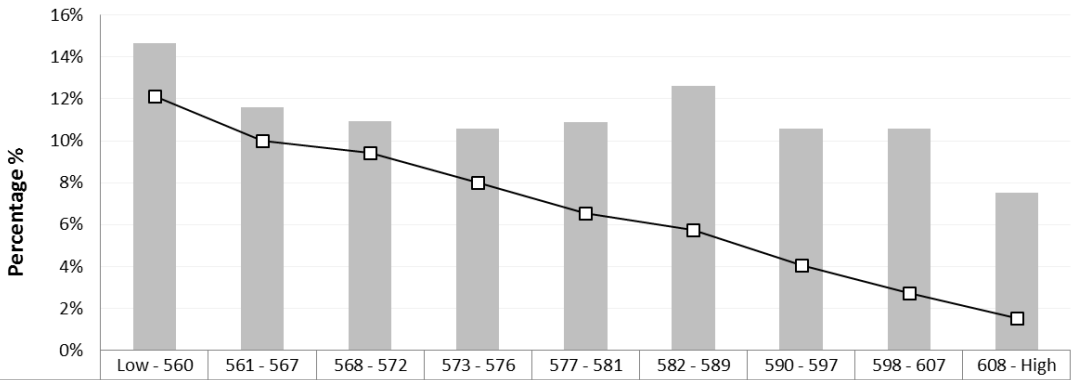
	Low - 560	561 - 567	568 - 572	573 - 576	577 - 581	582 - 589	590 - 597	598 - 607	608 - High
Distribution % (Development)	15.7%	12.1%	11.2%	10.1%	10.0%	11.8%	10.6%	10.6%	7.9%
Distribution % (Hold-Out)	16.2%	12.3%	10.8%	10.2%	9.9%	11.7%	11.6%	9.8%	7.5%
Bad Rate % (Development)	11.15%	9.01%	7.67%	6.96%	6.13%	4.77%	3.85%	2.66%	1.24%
Bad Rate % (Hold-Out)	9.77%	8.85%	7.74%	6.56%	6.56%	5.05%	3.46%	3.30%	1.09%

Figure 2.7: Score trend (internal application scorecard).

Figure 2.7 illustrates the score trend of the internal application scorecard. The score trends in Figure 2.7 illustrates the high bad rates for low scores and the higher the score the lower the bad rate for both the development and hold-out sample. The correlation analysis and score trend for the out-of-time window (OOT) is presented in Table 2.8 and Figure 2.8 respectively.

Table 2.8: Correlation (OOT).

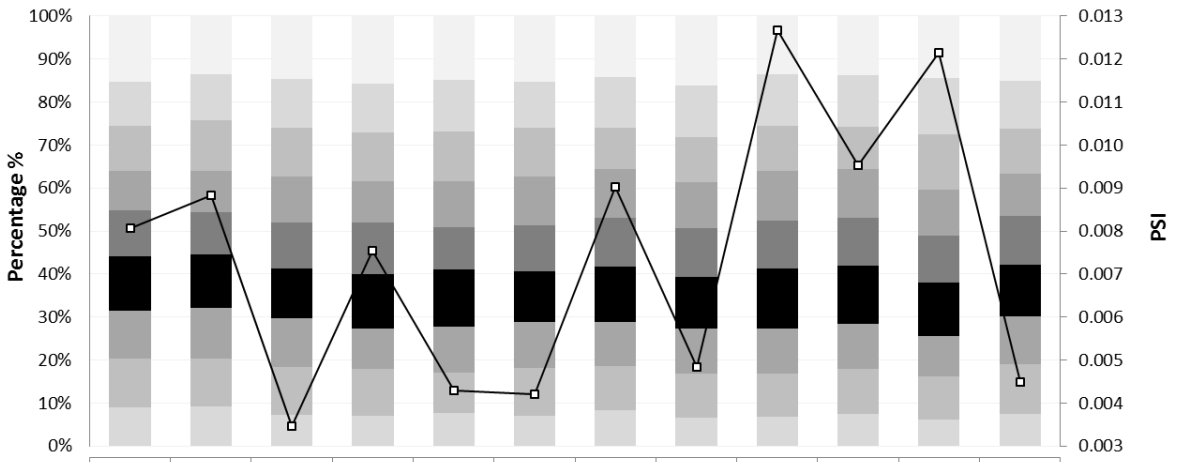
	Characteristic 1	**Characteristic 2**	**Characteristic 3**	**Characteristic 4**	**Characteristic 5**
Characteristic 1	100%	-5%	6%	3%	2%
Characteristic 2	-5%	100%	18%	13%	9%
Characteristic 3	6%	18%	100%	24%	5%
Characteristic 4	3%	13%	24%	100%	2%
Characteristic 5	2%	9%	5%	2%	100%



	Low - 560	561 - 567	568 - 572	573 - 576	577 - 581	582 - 589	590 - 597	598 - 607	608 - High
Distribution % (OOT)	14.6%	11.6%	11.0%	10.6%	10.9%	12.6%	10.6%	10.6%	7.5%
Bad Rate % (OOT)	12.15%	10.01%	9.44%	7.99%	6.54%	5.73%	4.08%	2.72%	1.51%

Figure 2.8: Score Trend (Internal Application Scorecard) for OOT Window.

The population stability index (PSI) which measure the change in distribution against the development sample is presented in Figure 2.9.



	Jan-12	Feb-12	Mar-12	Apr-12	May-12	Jun-12	Jul-12	Aug-12	Sep-12	Oct-12	Nov-12	Dec-12
Low - 560	15.3%	13.7%	14.6%	15.7%	14.8%	15.2%	14.1%	16.1%	13.5%	13.8%	14.4%	15.1%
561 - 567	10.3%	10.7%	11.4%	11.3%	11.9%	10.8%	12.0%	12.1%	12.1%	12.1%	13.1%	11.1%
568 - 572	10.5%	11.6%	11.4%	11.4%	11.6%	11.5%	9.4%	10.5%	10.4%	9.8%	12.8%	10.5%
573 - 576	9.0%	9.6%	10.6%	9.7%	10.8%	11.2%	11.4%	10.7%	11.6%	11.3%	10.9%	9.8%
577 - 581	10.9%	9.9%	10.9%	11.9%	9.9%	10.6%	11.2%	11.2%	11.0%	11.1%	10.8%	11.3%
582 - 589	12.6%	12.4%	11.4%	12.8%	13.3%	11.8%	12.9%	12.0%	14.1%	13.6%	12.5%	12.0%
590 - 597	11.2%	11.8%	11.4%	9.3%	10.6%	10.8%	10.4%	10.5%	10.4%	10.4%	9.3%	11.2%
598 - 607	11.4%	11.2%	11.1%	10.8%	9.4%	11.0%	10.3%	10.3%	10.1%	10.4%	10.1%	11.5%
608 - High	8.9%	9.1%	7.4%	7.1%	7.7%	7.1%	8.3%	6.6%	6.9%	7.5%	6.1%	7.5%
PSI	0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.00

Figure 2.9: Population Stability Index (PSI).

A $PSI < 0.1$ is considered to be stable (as evident from Figure 2.9).

2.6.2 Bureau Scorecard

The credit bureau TransUnion in South-Africa (as described in Section 2.1) provides on request bureau scores (Empirica scores) to the lending institutions. The Gini-coefficient of the Empirica score was calculated as 42.0% for the development window as shown in Figure 2.10.

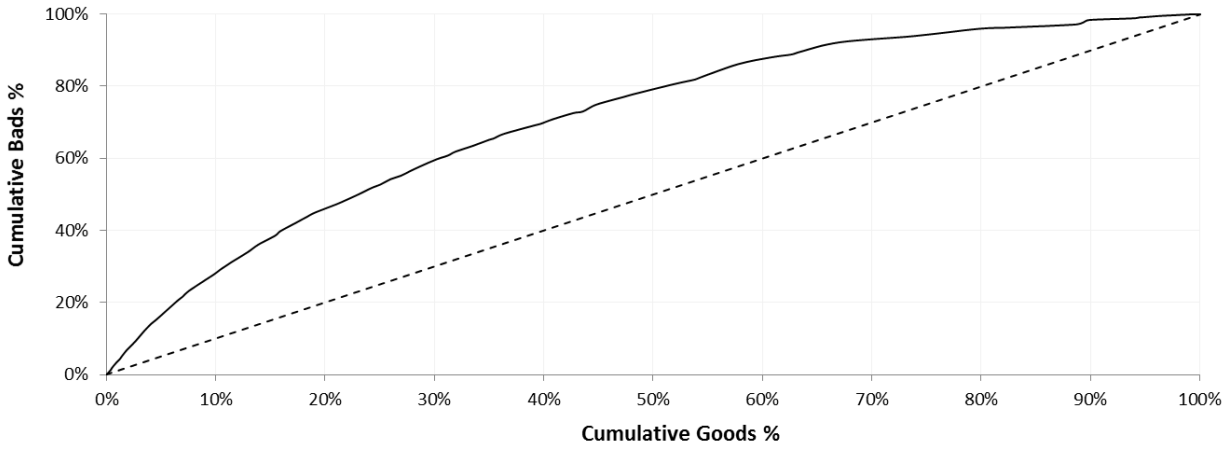
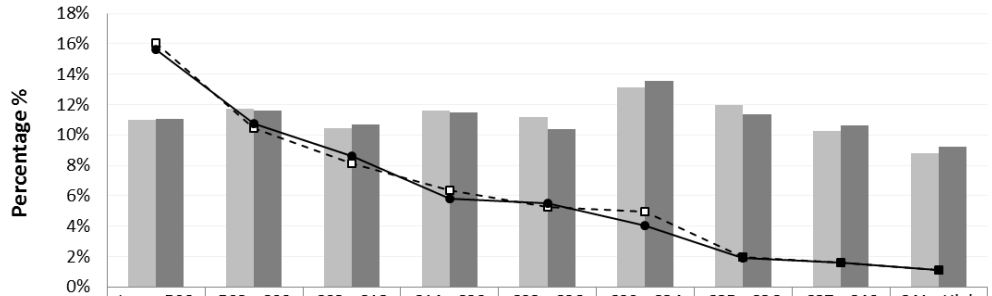


Figure 2.10: Empirica Score Gini-Coefficient (Development Window).

The score trend of the Empirica score is presented in Figure 2.11.



	Low - 592	593 - 602	603 - 613	614 - 622	623 - 628	629 - 634	635 - 636	637 - 640	641 - High
Distribution % (Development)	11.0%	11.7%	10.4%	11.6%	11.2%	13.1%	12.0%	10.3%	8.8%
Distribution % (Hold-Out)	11.1%	11.6%	10.7%	11.4%	10.4%	13.5%	11.4%	10.6%	9.2%
Bad Rate % (Development)	16.06%	10.44%	8.13%	6.38%	5.25%	4.94%	1.94%	1.60%	1.09%
Bad Rate % (Hold-Out)	15.61%	10.77%	8.59%	5.83%	5.52%	4.01%	1.90%	1.62%	1.09%

Figure 2.11: Score Trend (Empirica Score).

Figure 2.11 illustrates again how high the bad rates are for lower scores and the higher the score the lower the bad rate (as expected).

2.6.3 Credit Scoring Matrix Model (CSMM)

Siddiqi (2006) presented three approaches to implement a multi scorecard solution; the sequential approach, the matrix approach and the matrix-sequential hybrid approach. In the

sequential approach an applicant is scored sequentially on each scorecard with each scorecard having its own cut-off. The sequential approach is best used when "hurdles" are being used, i.e. an applicant must have a minimum bureau score before sequentially moving onto the next scorecard. In the matrix approach multiple scorecards are used concurrently with the decision making based on a combination of the cut-offs from each scorecard (Siddiqi, 2006, p. 144). This matrix approach is used when a balanced choice needs to be made from different types of ideally independent information. (Siddiqi, 2006, pp. 144-145). The matrix-sequential hybrid approach is used when an applicant is firstly put through the sequential approach after which the applicant goes through the matrix approach. The matrix-sequential hybrid approach is more versatile than the sequential approach and simpler than the matrix approach. This approach is best used when three independent scorecards are used, balancing several competing interests and using in conjunction with policy rules for applicants that are prequalified (Siddiqi, 2006, p. 146).

The CSMM is a matrix approach except that the decision-making is based on a cut-off determined after the internal application scorecard and bureau scorecard are combined (discussed in Section 2.6.4). The CSMM is represented as follows:

$$\begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_i \end{bmatrix} \times [B_1 \quad B_2 \quad \dots \quad B_j] = \begin{bmatrix} C_{11} & \cdots & C_{1j} \\ \vdots & \ddots & \vdots \\ C_{i1} & \cdots & C_{ij} \end{bmatrix} \tag{6}$$

where

A_i represents Internal Application Scorecard score i,

B_j represents Empirica Scorecard score j

C_{ij} represents the matrix's score ij

$$i = 1 \dots x, j = 1 \dots y$$

and x = score range of Internal Scorecard, y = score range of Empirica Scorecard

In (6) C_{11} represents the matrix score where there is a low internal application score and low Empirica score, C_{ij} represents a high internal application score and high Empirica score, scores to the bottom left represent high internal application scores and low Empirica scores and scores to the top right represent low internal application scores and high Empirica scores.

Using the scores from the internal application scorecard and combining with the Empirica score the bad rates for the CSMM are represented in Table 2.9.

Table 2.9: CSMM (Bad Rate % per Matrix Cell).

BAD RATE %		Bureau Score (Bad Rate %)										
		A	B	C	D	E	F	G	H	I	J	Bad Rate (%)
Internal Scorecard (Bad Rate)	1	20.6	16.7	15.2	13.3	11.9	10.7	9.7	9.9	9.8	4.9	**12.1**
	2	18.8	15.6	12.5	11.6	10.4	8.3	8.4	8.3	6.4	4.8	**10.0**
	3	18.7	15.5	12.8	10.2	9.3	6.7	6.9	7.3	6.8	4.0	**9.1**
	4	18.8	14.8	12.7	9.7	8.3	7.9	5.6	4.5	6.6	3.4	**8.2**
	5	17.3	12.5	12.1	10.5	10.5	6.0	6.1	5.6	5.8	2.9	**7.7**
	6	17.6	12.8	12.6	9.5	9.4	5.2	5.8	5.2	4.6	2.2	**6.8**
	7	18.5	12.1	10.7	10.3	7.7	6.0	6.4	4.7	4.9	2.1	**6.2**
	8	18.1	13.7	11.0	8.8	6.5	5.4	4.0	2.9	3.5	1.6	**4.9**
	9	16.7	10.6	10.5	6.4	6.0	4.7	4.7	3.5	2.8	1.1	**3.8**
	10	12.4	12.9	6.3	5.0	4.4	3.0	3.2	2.4	1.4	0.8	**2.1**
Bad Rate		**18.4**	**14.1**	**11.6**	**9.2**	**8.1**	**6.2**	**5.9**	**4.8**	**4.8**	**1.9**	**6.3**

Table 2.9 indicates high bad rates in the top left quadrant with low internal application scores and low Empirica scores. The lowest bad rates are at the bottom right quadrant as expected (high internal application scores and high Empirica scores). Top right and bottom left quadrant are marginal client bad rates with either a low internal application score or high Empirica score (top right quadrant) or high internal application score and low Empirica score (bottom left quadrant).

To illustrate the benefit of using the CSMM from a predictive point of view the Gini-coefficient was calculated. Firstly, the Gini-coefficient for the internal application scorecard was determined by grouping each score into the 10 score bands used in Table 2.9 and was calculated as 29%. Secondly the Gini-coefficient for the bureau scorecard (Empirica score) was determined by grouping each score into the 10 score bands used in Table 2.9 and was calculated as 41%. Thirdly the Gini-coefficient for the CSMM was calculated with a value of 46% by ranking the bad rates from the matrix from worst to best. The Gini-coefficients are summarised in Figure 2.12.

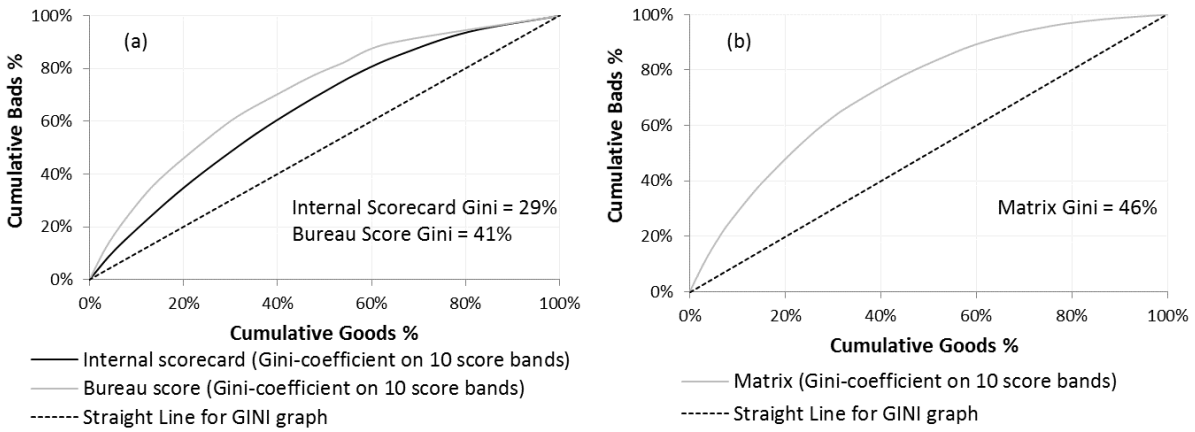


Figure 2.12: Gini-Coefficients: (a) Internal Scorecard & Bureau Score and (b) matrix.

In Figure 2.12 the uplift in Gini-coefficient is $\frac{46-41}{41} = 12\%$. The theoretically correct way to compute the uplift from the CSMM is to compare to a one-dimensional scorecard containing both internal application information and external bureau information and was calculated as 43%, indicating that the CSMM still gives uplift. However it is not recommended to implement a one-dimensional scorecard containing both internal application information and external bureau information for various reasons: External bureau information is much stronger than internal application information, hence when building a one-dimensional scorecard containing both internal application information and external bureau information the bureau information will completely overpower the internal application information. Internal application information is more susceptible to application information manipulation especially to a certain secured retail product. Given that the application information can be manipulated the second axis of the CSMM acts as a hedge. Converting the internal application information to one axis of a CSMM opens the door to marginal clients, e.g. you can have applicants with high internal application scores, but low bureau scores and you can have applicants with low internal application scores, but high bureau scores, this opens up opportunities. In addition clients with no bureau history can now also be considered as they would have internal application information. Although the Gini-coefficient of 30.5% of the internal application scorecard (excluding bureau information) might seem too low the necessity of this information within the overall CSMM is paramount, e.g. preventing the booking of high loan to value (LTV) applicants which could have devastating effects.

Whenever there is a need to change to a latest bureau score it would be easier to change one axis of the CSMM instead of doing an entire new scorecard rebuild which will be the case if

the bureau score is embedded within a one-dimensional scorecard containing both internal application information and external bureau information. The CSMM in addition does not fully put reliance on bureau information or solely on internal processes, i.e. if credit bureaus are unable to provide information or if internal processes fail one can fall back on either the internal application scorecard or bureau scorecard respectively in such extreme cases. Given the uplift in Gini-coefficient from the CSMM against the bureau scorecard and the numerous reasons given above the CSMM can be regarded as the optimal model (most favourable) from a credit scoring perspective in retail banking.

The CSMM is presented in Figure 2.13.

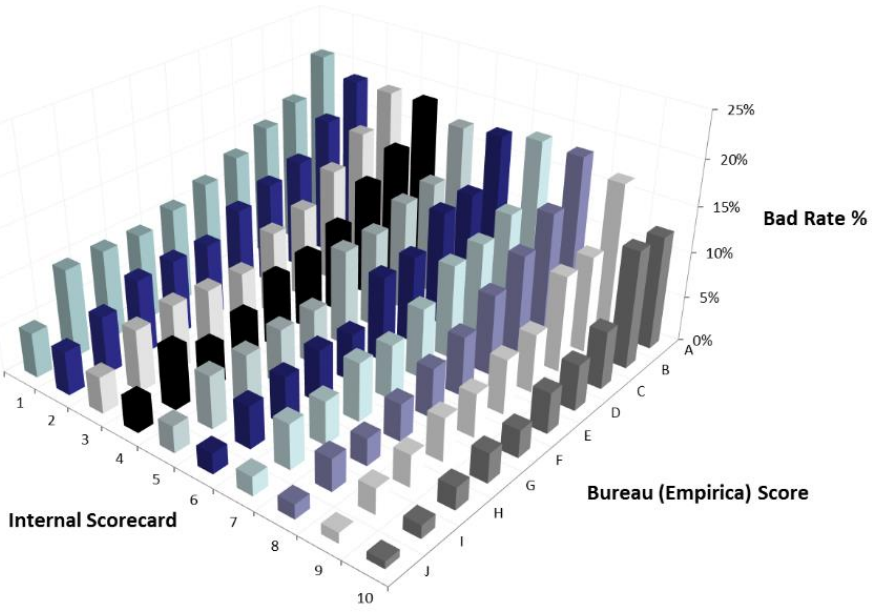


Figure 2.13: CSMM.

2.6.4 Cut-off Score Determination

To illustrate the effectiveness of the CSMM, consider a one dimensional approach and assume only the internal application scorecard is available to score clients presented in Table 2.10. The bad rate data in Table 2.10 correspond to the bad rate data in Table 2.9 for the internal application scorecard dimension.

Table 2.10: One-dimensional score (internal application score).

Internal Application Score	Bad Rate (%)
Low - 559	**12.13**
560 - 561	**9.95**
562 - 565	**9.12**
566 - 568	**8.24**
569 - 572	**7.70**
573 - 576	**6.84**
577 - 581	**6.18**
582 - 588	**4.92**
589 - 597	**3.79**
598 - High	**2.07**

Cut-off Score

Using Table 2.10, suppose that the decision was made that a bad rate ≥ 8.24 for applicants would not be considered which will set the cut-off score at 568 for the internal application score. Table 2.10 is a straight cut-off from the one-dimensional scorecard. Secondly, assume the internal application scorecard and bureau (Empirica) score in matrix format are at hand to assess an applicant. Using Table 2.9, suppose that the decision was made that a bad rate ≥ 8.24 for applicants will not be considered which will have a diagonal cut-off presented in Table 2.11.

Table 2.11: Two-dimensional CSMM.

BAD RATE (%)		Bureau Score (Bad Rate %)										
		A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**Bad Rate (%)**
Internal Scorecard (Bad Rate)	**1**	20.6	16.7	15.2	13.3	11.9	10.7	9.7	9.9	9.8	4.9	**12.13**
	2	18.8	15.6	12.5	11.6	10.4	8.3	8.4	8.3	6.4	4.8	**9.95**
	3	18.7	15.5	12.8	10.2	9.3	6.7	6.9	7.3	6.8	4.0	**9.12**
	4	18.8	14.8	12.7	9.7	8.3	7.9	5.6	4.5	6.6	3.4	**8.24**
	5	17.3	12.5	12.1	10.5	10.5	6.0	6.1	5.6	5.8	2.9	**7.70**
	6	17.6	12.8	12.6	9.5	9.4	5.2	5.8	5.2	4.6	2.2	**6.84**
	7	18.5	12.1	10.7	10.3	7.7	6.0	6.4	4.7	4.9	2.1	**6.18**
	8	18.1	13.7	11.0	8.8	6.5	5.4	4.0	2.9	3.5	1.6	**4.92**
	9	16.7	10.6	10.5	6.4	6.0	4.7	4.7	3.5	2.8	1.1	**3.79**
	10	12.4	12.9	6.3	5.0	4.4	3.0	3.2	2.4	1.4	0.8	**2.07**
Bad Rate		**18.44**	**14.07**	**11.57**	**9.15**	**8.10**	**6.15**	**5.85**	**4.83**	**4.80**	**1.90**	**6.27**

Table 2.11 presents information that is not a straight cut-off as for the one-dimensional score-card, but a more granular two-dimensional cut-off. Having such a strategy cut-off have some advantages such as: clients with a low bureau (Empirica) score, but high application score can

be considered, clients with a low application score, but high bureau score (Empirica) can be considered, negative effects of application form manipulation for internal scorecard are reduced because the bureau (Empirica) hedges this manipulation and a greater dimension of information is taken into account to determine whether an applicant will be bad in future.

2.7 Conclusion

In the last 60 years (to 2016), credit scoring played (and still plays) an important role in the credit industry which mitigates and controls future bad clients. However, literature is limited, which encourages further research in credit scoring. In the last 25 years (2015) credit scoring has grown, particularly in the banking industry, making the separation between good and bad clients more critical to have effective credit risk management and to reduce future bad debts which emphasises the significance of this research.

This article presented a literature background on statistical techniques used in the world of credit scoring illustrating both weaknesses and advantages of each technique. The methodology then led to an application scorecard which includes the decision on the initial bad definition, outcome decision or maturity, development window, exclusions to be used in the development, sampling, reject inference, characteristics considered to predict bad clients, bucketing process, business intervention, regression technique, relevant scorecard tests and out-of-time testing.

The contribution from this article is the construction of a scoring model to optimise the separation between good and bad clients in the form of a CSMM. The CSMM consists of two components namely the internal application scorecard (built from information specific to the organisation) and the credit bureau or Empirica score (based on external data). These provide insights on client's credit performance with all other credit organisations. The separation of internal organisation client credit performance information and external client credit performance information by the CSMM is important. Internal information, contained in the internal application score, indicate a direction of client credit performance, however the external information contained in the Empirica score acts as a hedge should the internal application score not capture all client-related credit performance.

During the construction phase of the internal application scorecard it was decided to use a 24-month outcome period to capture appropriate number of bad clients and to work on a population which is a reflection of the current portfolio. Reject inference results indicated the

higher bad rate for rejected clients (as expected). The internal application scorecard gave a Gini-coefficient of 30.5% for the development window with the relevant scorecard tests indicating no correlation between characteristics and stability. Combining the internal application scorecard with the Empirica score the CSMM was constructed which distinguish between good and bad clients on a more granular level which in addition enable the setting of a more appropriate cut-off score which was highlighted as a problem in past literature. It was illustrated that having a CSMM gives a relative percentage uplift in the Gini-coefficient of 12.0% to distinguish between good and bad clients more effectively which will lead to fewer clients being initially selected which would have resulted in future bad clients creating effective credit risk management and reducing future bad debts. How a CSMM establishes the cut-off score more appropriately to accept clients than a one-dimensional scorecard by providing a greater granularity option adds to the contribution of this work.

Possible future research includes the combination of an internal application scorecard with an internal bureau scorecard which can be built on specific bureau information that works best for the organisation in question. In addition, possibilities exist to investigate the optimal construction when building a matrix.

Bibliography

Abdou, H. & Pointon, J., 2011. Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance & Management,* 18(2-3), pp. 59-88.

Allen, L., De Long, G. & Saunders, A., 2004. Issues in the credit risk modelling of retail markets. *Journal of Banking and Finance,* Volume 28, pp. 727-752.

Anderson, R., 2007. *The credit scoring toolkit, theory and practise for retail credit risk. Management and decision automation.* United States: Oxford University Press Inc..

F., 2014. *FICO Model Builder.* United States of America.

Galton, F., 1889. Quoted in Anderson, R., 2007. *The credit scoring toolkit, theory and practise for retail credit risk. Management and decision automation.* United States: Oxford University Press Inc..

Hand, D. J. & Henley, W. E., 1997. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 160(3), pp. 523-541.

Harrell, F. E., 2015. *Regression Modeling Strategies.* Second ed. Switzerland: Springer International Publishing.

Lim, M. K. & Sohn, S. Y., 2007. Cluster-based dynamic scoring model. *Expert Systems with Applications,* Volume 32, pp. 427-431.

Siddiqi, N., 2006. *Credit risk scorecards, developing and implementing intelligent credit scoring.* United States of America: John Wiley & Sons, Inc..

Thomas, . L. C., Edelman, D. B. & Crook, J. N., 2002. *Credit scoring and its applications.* Philadelphia: SIAM.

TransUnion, 2015. *Credit Bureau, Credit Reporting Companies. Learn about your credit..* [Online]
Available at: https//www/mytransunion.co.za
[Accessed 2 4 2015].

TransUnion & FICO, 2009. *Empirica. Minimise your credit risk. Increase your profitability.,* Johannesburg: TransUnion.

Chapter 3
A statistical technique to enhance application scorecard monitoring

Nico Kritzinger[1] and Gary van Vuuren[2]

ABSTRACT

Application scoring plays an essential part in determining the future quality book of an organisation. The performance monitoring of an application scorecard therefore is critical to ensure the application scorecard performs as expected. As a lot of focus has been put on the modelling techniques used when building an application scorecard an area that is less in the spotlight is the implementation of an application scorecard. Performance measures on the accept population appears to change in predictive power after the implementation of an application scorecard. This article introduces and illustrates the use of a statistical measure to track the performance of the accept population after the point of implementation on a comparable basis against the development window of the application scorecard.

Keywords: Credit risk, application scoring, credit risk management

JEL classification: C20, G32

3.1 Introduction

An application scorecard is the most important tool an organisation can use during the acquisition of new business as the application scorecard determines the quality of the book for a future time frame dependent on the type of product to be financed, e.g. home loans could be financed up to 20 to 30 years whereas a vehicle loan finance product is a much shorter timeframe. Fair Isaac (presently known as FICO) introduced credit scoring approximately 60 years ago in 1956 to evaluate the level of credit risk of an application (Anderson, 2007, p. 40). Credit scoring transforms data into numerical scores through statistical models, which is then used to direct credit decisions. The well-known form of credit scoring is called application scoring which is used at the credit application stage. Application scoring determines the level

[1] PhD-student at the Faculty of Natural Sciences, North-West University.
[2] Visiting professor, Faculty of Natural Sciences, North-West University.

of risk applications present from a credit risk perspective and is generally based on historical application data from the organisation (Lim & Sohn, 2007, p. 427). Application scoring together with approval rates, profit and losses are used to make the final decision whether an application will be booked or not (Siddiqi, 2006, p. 5).

An application scorecard which consists of a group of characteristics can be modelled using various modelling techniques such as linear regression, discriminant analysis, probit analysis, decision trees, expert systems, genetic programming and neural networks (Abdou & Pointon, 2011, pp. 67-68). FICO presented the scorecard divergence method which dominated application scorecard development however logistic regression is the most common method used when building a scorecard (Huang & Scott, 2007, p. 1).

Given the vast majority of modelling techniques available and the research that has been completed regarding the modelling techniques, an area that has been less in the spotlight is the effect of the implementation of an application scorecard in which the statistical level of power of the application scorecard for the accept population could be misinterpreted when an application scorecard is implemented. Although statistical measures during development of an application scorecard could indicate a certain level of predictive performance, the statistical measures indicate a different level of predictive performance after the implementation of the application scorecard (Huang & Scott, 2007, p. 1).

Various statistical measures exist to determine the power or the measure of separation between good and bad clients of an application scorecard such as the Kolmogorov-Smirnov (KS) statistic (Siddiqi, 2006, p. 123). The misclassification matrix and divergence statistic are also measures used to determine the power or measure of separation for application scorecards (Anderson, 2007, pp. 189-190). The most common statistical measure used when evaluating the power or measure of separation of an application scorecard is called the Gini-coefficient where the area under the Lorenz curve is determined. Given the above mentioned statistical measures the interpretation of these measures plays an essential role both at the development stage of the application scorecard as well as after implementation of the application scorecard. These statistical measures play an important role in application scorecards not only indicating power and measure of separation, but it also indicates whether the application scorecard adds value and creates better risk ranking ability (Anderson, 2007, p. 188). The statistical measures mentioned above are mainly used for testing a credit scorecard's power or

ranking ability which is the focus area of Chapter 3. However there exist statistical measures to test accuracy when credit scorecards are used to provide probability estimates in areas such as pricing, forecasting or capital allocations. These accuracy tests include the binomial test which is used to compare observed and estimated success rates for a single group, the Homer-Lemeshow test which is based on the binomial test, but applied on all groups or the log-likelihood which provide measures of power and accuracy for ungrouped cases (Anderson, 2007, p. 212; Van Gestel & Baesens, 2009, p. 270). Another method to validate credit scorecards which assigns probability estimates across all rating grades is the Brier score which requires the calculation of the average quadratic deviation of the forecasted PD and actual default rates (Engelman & Rauhmeier, 2011, p. 300).

As application scoring plays an important role in the future book quality of an organisation it is important that all concepts and elements both during development and implementation of the application scorecard are understood and correctly interpreted. In this paper the statistical measures of power or measure of separation between good and bad accounts of an application scorecard are explored after an application scorecard is implemented. How the level of these measures compare during the development phase of the application scorecard is also examined. Emphasis should be placed on the role of an application scorecard for a credit organisation which, in essence, affects future profits and level of impairments.

3.2 The effect of implementing an application scorecard

When building an application scorecard it is common practise to have a development sample to illustrate the power of the application scorecard and an independent holdout sample within development to validate the performance of the application scorecard (Siddiqi, 2006: 127). It is recommended that an application scorecard is built on 80% of the development sample and tested on the independent 20% holdout sample (Siddiqi, 2006: 127). A complexity within application scoring is the added dimension of rejects built into the scorecard to have a through-the-door (all applications) population scorecard. The technique to include rejects is called reject inference (Siddiqi, 2006: 98). During an application scorecard development one generally has only performance data for the accept population and not for the reject population which requires the use of reject inference. The reason why rejects are included in the application scoring development is because the population with known performance (accepts

population) in the development sample would be biased to only those type of clients should the application scorecard be implemented (Siddiqi, 2006: 98).

Given the development sample and holdout sample within an application scorecard development a problem that exists within the scorecard development is the possible deterioration of the predictive power of the application scorecard when performance measures are compared between development and post implementation for the accept population. This creates a problem when the accept population is monitored after the point of implementation. Although reject inference can be used to infer rejected applications it is still the general belief that application scorecard performance monitoring is done on only the accept population as a result of the effort risk managers apply regarding risk strategies and making cut-off decisions (Huang & Scott, 2007, p. 2). At the point of implementation, the "cut-off score" would effectively change as pre-implementation another application scorecard was in place, in other words a swap-set has occurred where clients that have been accepted in the past are now rejected and clients that have been rejected in the past are now accepted. To illustrate this consider the following three application scorecards:

- "Previous scorecard": The application scorecard was used in the past to assess the credit risk of a client
- "Current scorecard": This application scorecard replaced the previous scorecard and is currently in use
- "New scorecard": This application scorecard is a newly developed scorecard which would replace the current scorecard.

Figure 3.1 present's possible implementation points of the three different application scorecards together with the development period for the "Current scorecard" and "New scorecard".

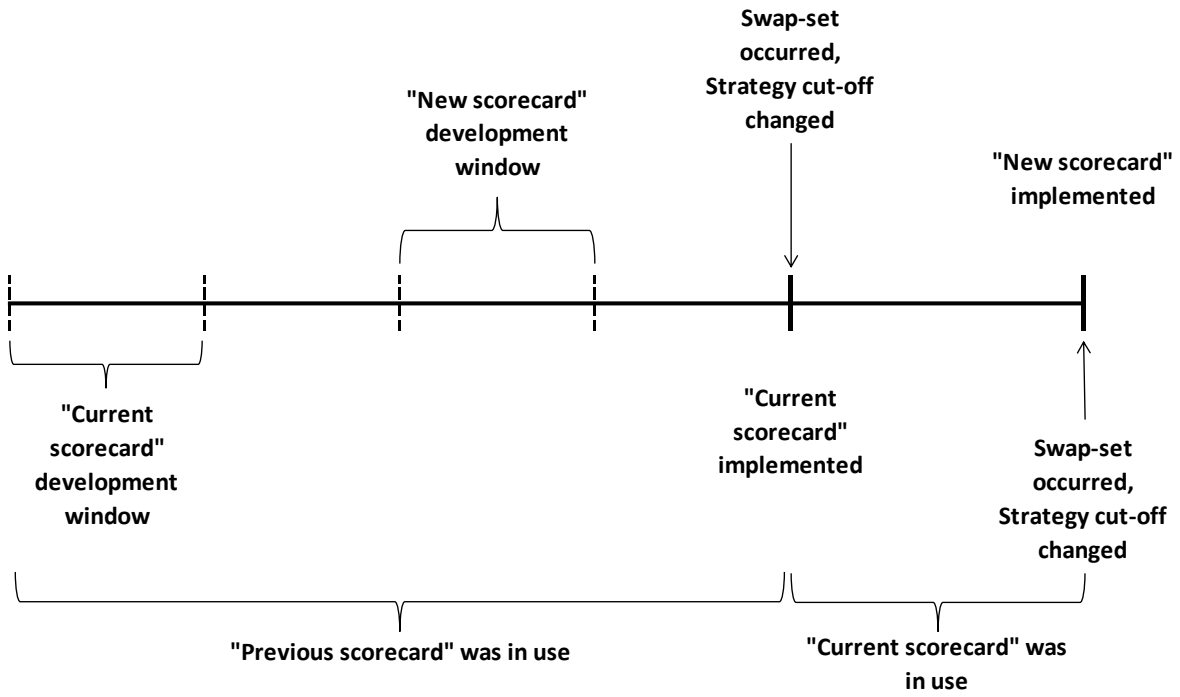


Figure 3.1: Time horizon for scorecard developments and implementations.

To compare scorecard scores pre and post implementation one can retrospective re-score applications pre implementation by the "New scorecard" such that one have a like for like comparison. The "Current scorecard" was implemented after the "Previous scorecard" was in use; and the "New scorecard" was implemented after the "Current scorecard" was in use. This creates a constant complexity since with each application scorecard implementation a swap-set occurs which would affect statistical performance measures of application score-cards between development and post implementation of the application scorecard for the accept population, hence affecting the application scorecard performance monitoring levels. The reasons for this complexity are twofold; firstly, when building application scorecard models one normally needs a sufficient amount of outcome to determine whether a client would be good or bad in future and secondly, although one would wish to employ a sufficient amount of data as possible to determine good and bad clients one preferably would not want to go back too far in history because the organisation's book profile could have changed from the past.

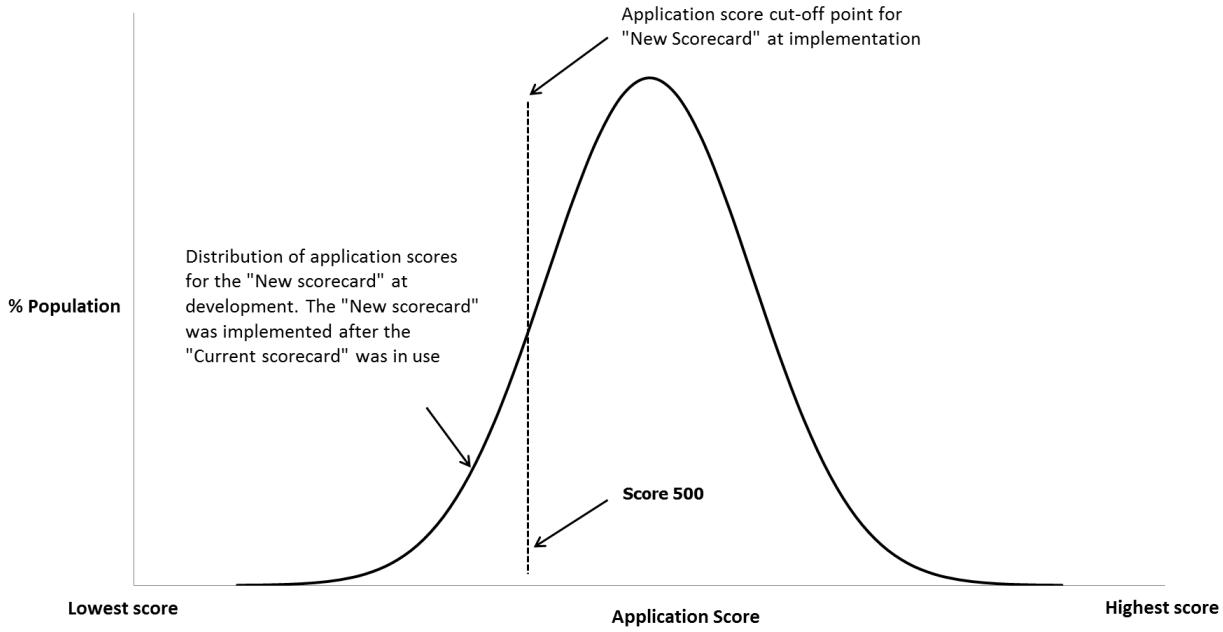


Figure 3.2: Application scorecard score distribution.

Figure 3.2 presents a theoretical illustration for a population with its application scorecard scores for the "New scorecard" at development, i.e. when the "Previous scorecard" was in use and the "New scorecard" was implemented after the "Current scorecard" was in use. The scores move from the lowest scores on the left to the highest scores on the right. Assuming from a business strategy perspective that the "New scorecard" would be implemented such that no client will be considered with a score less than 500 would result in a different distribution of scores after implementation.

As the "New scorecard" suggest that client with low scores are bad clients at development when the "Previous scorecard" was in use the population of scores post implementation would not include some of these bad clients as the cut-off score is 500, effectively again the swap-set occurring as explained earlier. This would have an effect on measures such as divergence and Gini-coefficient as previously mentioned when comparing performance measures pre and post implementation for the accept population. This would, in addition, influence effective risk management. Because one is unable to have effective risk management one could also be misguided that quality clients are being booked and swapped in and out which could lead to a worsening of book profile, increase in losses and higher capital.

Huang and Scott (2007) indicated that the loss in predictive power for the application scorecard on the accept population from development to post implementation is not due to reject

inference, but because a portion of the uplift in Gini-coefficient for a new application scorecard has already been absorbed by the current application scorecard (Huang & Scott, 2007, pp. 2 - 7). The aim of this study is to investigate the connection between scorecard development and implementation following a similar approach as Huang and Scott (2007) and to assemble a statistical technique to have comparable performance measures pre and post implementation for the accept population.

3.3 Problem statement and objective

The aim of this article is to introduce and illustrate a statistical measure called the swap-set Gini-coefficient to enhance the performance monitoring of an application scorecard. The primary objectives of this study are to improve application scorecard monitoring for the accept population and to enhance credit risk measurement and management. The focus includes credit risk, application scoring and credit risk management. Historical credit scoring statistical performance measures are investigated and the effect of the proposed primary objectives on credit risk management is assessed with the aim of implementing these in the retail banking environment for application scorecard purposes.

3.4 Literature review

3.4.1 Swap-sets

Acquisition scorecards are traditionally built on the accepted applications known as the "Known Good Bad (KGB)" scorecard after which reject inference is applied to obtain the "All Good Bad (AGB)" scorecard based on all applications. Swap-sets are generally observable after the reject inference process. Within the reject inference process the unknown performance of rejected applicants are inferred to obtain inferred good and bad applications. Having both the known performance of the accepted applications and the inferred performance of the rejected applications one can identify the "swap set". Inferred goods are applications rejected in the past, however indicating good behaviour after the inference process which would be accepted in future. By swapping the inferred good applications with accepted applications from the past exhibiting bad performance one obtain enhanced performance through better decision making (Siddiqi, 2006, p. 100).

Thomas (2000) indicated another concept of swap-sets occurring due to economic changes. A scorecard may be built in good economic conditions, but applied on a time period where

the economy changes or vice versa. Thomas (2000) presented an example comparing scorecards with one developed in good economic conditions and the other one in worse economic conditions. Adjusting the cut-off for accepting applicants for both scorecards on the same level it was found that 25% of the population that was rejected in the one year would be accepted in the other and vice versa, resulting in the swap-sets, however Thomas points out that this economic conditions are not the only cause of risk changing behaviour (Thomas, 2000, pp. 163-165). Anderson (2007) indicated that application scorecards can also be affected by changes in product features or customer demographics, operational changes, payment behaviour changes over time by people or the age of the application scorecard (Siddiqi, 2006, pp. 167-168; Anderson, 2007, pp.83-90). Thomas (2000) mentioned ways to incorporate economic conditions within the scorecard development. Firstly one can build a scorecard for good economic conditions and bad economic conditions, however this puts responsibility on the scorecard developer or credit manager to decide what the future holds. In addition data can become old when the time comes to apply a scorecard for a certain economic condition with changes in population. Other methods include adding onto the normal credit score, using graphical methods, Bayesian learning networks or expanding on the ideas of cliques and Markov blankets (Thomas, 2000, pp. 163-165).

Anderson (2007) indicated important testing requirements once an application scorecard is implemented which could indicate change or swap-sets. Testing during implementation ensure the system is working as according to original design relying on comparing test versus expected results. Operational drift where a characteristic is not populated or when the calculation or process has changed is the most difficult to measure and identify. Operational drift can occur due to changes on the questions within the application form, changes to pre-capture screening criteria, difference in calculations between development and operational data; and differences between development and operational processes (Anderson, 2007, pp. 454-455). Anderson (2007) presents methods tracking change or drift in population during the monitoring process. Front-end reports refer to the monitoring of the stability (score drift), selection process (cover inputs and outputs of the selection process, e.g. volumes processed) and override reasons (cover changes in decision to ensure maximum benefit obtained from the application scores) of the application scorecard and can be generated as soon as the application scorecard is implemented (Anderson, 2007, pp. 467-468). The monitoring of stability

or drift can be done through the population stability reports (PSI) which track changes in characteristic distributions and through the score shift report which identifies the source of changes in the application score distribution (Anderson, 2007, pp. 480-481).

Huang and Scott (2007) investigated the problem faced by application scorecard developers when comparing the performance of a new application scorecard ("New Application Scorecard") developed in the development window and comparing the performance in an out-of-time (OOT) window that indicates significant drop in predictability from the "New Application Scorecard" for the accepts population. Huang and Scott (2007) focussed on two hypothesis to resolve the problem.

Firstly the effect of reject inference was investigated where the "New Application Scorecard" sensitivity towards the reject inference sample size at development was investigated using different scenarios. It was found that as the percentage of rejects within the development sample increases the Gini-coefficient for the development accepts decreases gradually however steadily. Also systematic reduction in the Gini-coefficient does not occur for all the scenarios tested when measured in the OOT accepts only population. Within the first hypothesis it was concluded that reject inference is not the root cause for the loss of power of the "New Application Scorecard".

Secondly the hypothesis was investigated of a possible population shift between development and the OOT window. It was noted that the OOT window was in a time frame when an application scorecard (call it "Current Application Scorecard") went live with the "New Application Scorecard" being in a period when a "Previous Application Scorecard" was used. This caused a swap-set of accounts where applicants that was accepted on the "Previous Application Scorecard" was rejected by the "Current Application Scorecard" and vice versa. Huang and Scott (2007) concluded on this hypothesis that the "Current Application Scorecard" absorbs a significant portion of the uplift of the "New Application Scorecard" from the development window. Huang and Scott (2007) completed their investigation by comparing Gini-coefficients on all applications with the Gini-coefficient decreasing for both the "Current Application Scorecard" and the "New Application Scorecard" maybe due to population shift. The "New Application Scorecard" significantly outperforms the "Current Application Scorecard" and although there is loss in power from the "New Application Scorecard" in the development window compared to the OOT window on the accepted population, most of the improvement

on the total population (i.e. all applications) comes from reject inference samples confirmed by swap-set analysis. Huang and Scott (2007) suggested that a new application scorecard development window should be chosen from a period after the current application scorecard went live, consideration should be applied on whether the performance window should be reduced when developing an application scorecard and investigating bureau samples of rejected applications which was approved on a similar product (Huang & Scott, 2007, pp. 1-11).

Various statistical performance measures exist which indicate the level of power, ranking ability or measure of separation between good and bad accounts of an application scorecard. In the remainder of this section a background on these statistical measures is given which is used in credit scoring. These statistical measures originated from various disciplines which include economics, mathematics, electronics and psychology (Anderson, 2007, pp. 187 - 188). In this section the statistical measures indicating measure of separation, power or ranking ability to be investigated include the divergence statistic, misclassification matrix, Kolmogorov-Smirnov (KS) statistic and the Gini-coefficient.

3.4.2 Divergence statistic

The divergence statistic is a measure of dispersion used in credit scoring (Siddiqi, 2006, pp. 128 - 129). According to Anderson (2007) divergence is a parametric statistic where the assumption is made that each of the good and bad groups are normally distributed (Anderson, 2007, p. 189). (2) illustrates how divergence is calculated:

$$\text{Divergence} = \frac{(\mu_G - \mu_B)^2}{(\sigma_G^2 + \sigma_B^2)/2}$$

where

μ represents the mean, and

σ^2 is the variance.

The divergence statistic is determined by the squared difference between the mean of the good and bad groups divided by the average variance between the good and bad groups. The divergence statistic can be graphically illustrated and is represented in Figure 3.3.

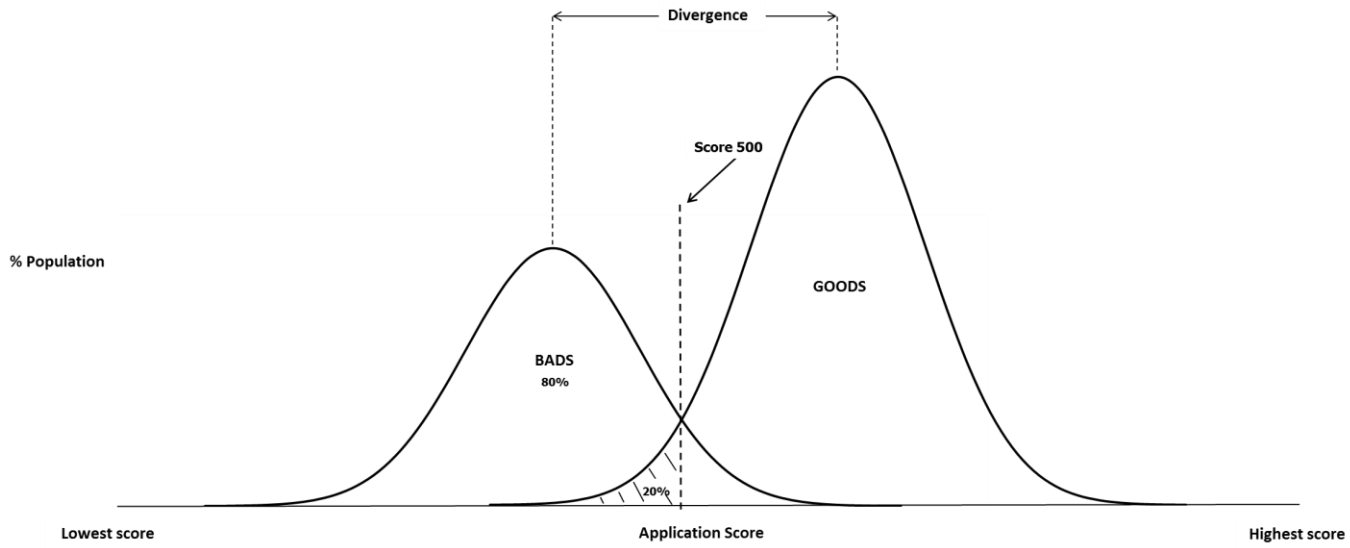


Figure 3.3: Divergence.

Considering Figure 3.3 assume we want to implement an application scorecard and we set the cut-off at a score of 500, i.e. an application below an application score of 500 we reject and above 500 we accept. In Figure 3.3 with a score cut-off of 500 20% of good applications would be turned down and 80% of the bad applications would be turned down. If an application scorecard produce a higher divergence and the cut-off score remains the same at 500 one can expect to turn down more than 80% bad applications at the same level of 20% good applications turned down.

The divergence statistic can be applied to continuous characteristics such as application scores and is closely related to the information value. Divergence is not generally used because of the assumption of normality between the two groups, its limited focus on continuous characteristics, distorting effect of outliers, other statistical measures are more generally used and divergence does not necessarily identify the better scorecard at a particular threshold (Anderson, 2007, p. 190; FICO, 2011, p.2). Advantages of the divergence statistic includes the measurement across the entire application score distribution, takes into account the separation between the two groups and the variance of the application score distributions (FICO, 2011, p. 2).

3.4.3 Misclassification matrix

The misclassification matrix is another method to determine how well an application scorecard separate good and bad accounts (Siddiqi, 2006, p. 120). The steps to be carried out to setup a misclassification matrix are illustrated in Figure 3.4 (Anderson, 2007, p. 190).

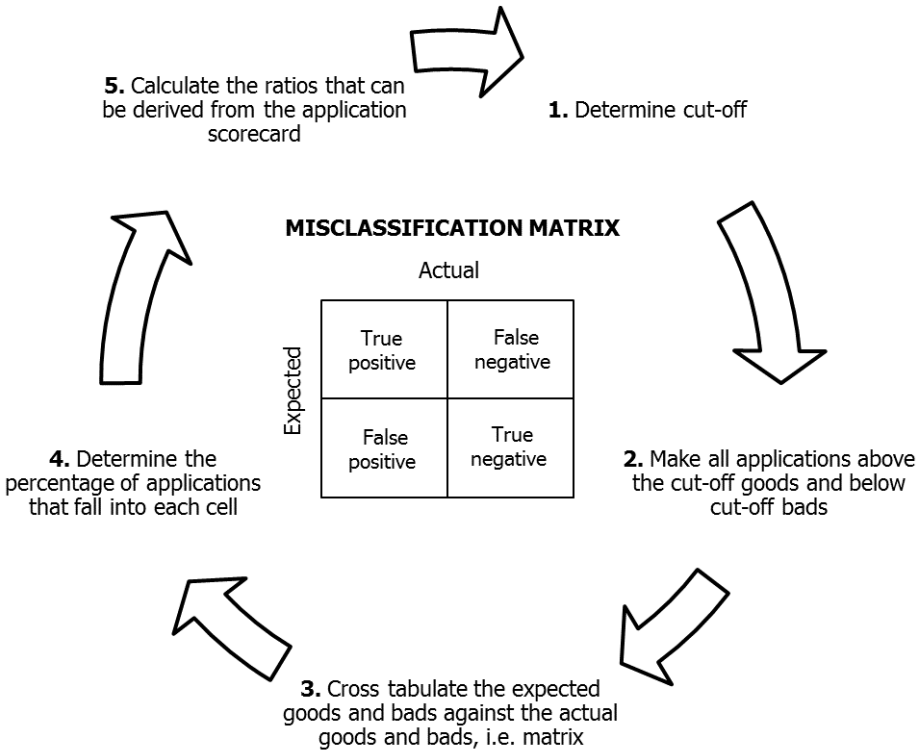


Figure 3.4: Misclassification matrix steps.

Four categories result from the misclassification matrix, firstly we have the true positives which is the correctly predicted bads, secondly we have the true negatives which is the correctly predicted goods, thirdly we have the false positives where bads were predicted however were goods and fourthly we have false negatives where goods were predicted however were bads. The misclassification rate is determined as follows:

$$\text{Misclassification rate} = \sum(\text{False positive, False negative})$$

An issue within the misclassification matrix is the choice of application score cut-off and although this method is generally used for model accuracy its generally insufficient unless the misclassification cost can be determined. A similar method to compare scorecard accuracy is determining swap sets which is useful when comparing recent development scorecards with scorecards currently in place (Anderson, 2007, pp. 190 - 191). It is also important that when the misclassification matrix is used, it should be connected to business goals where the objective of the scorecard is critical (Siddiqi, 2006, p. 120).

3.4.4 Kolmogorov-Smirnov (KS) statistic

Kolmogorov and Smirnov (1933) proposed this generally used credit scoring KS-statistic which is used to indicate how well a scorecard separate goods from bads (Anderson, 2007, p. 195). Anderson (2007) indicated that the KS-statistic is widely used in the United States of America (USA), but is not as popular outside the USA where the Gini-coefficient is mostly used. Anderson (2007) further states that it is dangerous to use only a single measure in isolation. The KS-statistic generally ranges between 20% and 70% with a value below 20% indicating that the model should be investigated and a value above 70% indicates that the scorecard seems to be over fitted (Anderson, 2007, p. 196). The KS-statistic is determined by first calculating the cumulative distribution percentage for both the good and bad population. Secondly, the greatest difference between the two distributions is known as the KS-statistic.

$$\text{KS Statistic} = \max[abs(Cd_B - Cd_G)]$$

where

Cd_B represents the cumulative distribution percentage for the bad population, and

Cd_G is the cumulative distribution percentage for the good population.

Figure 3.5 illustrates how the KS-statistic is determined.

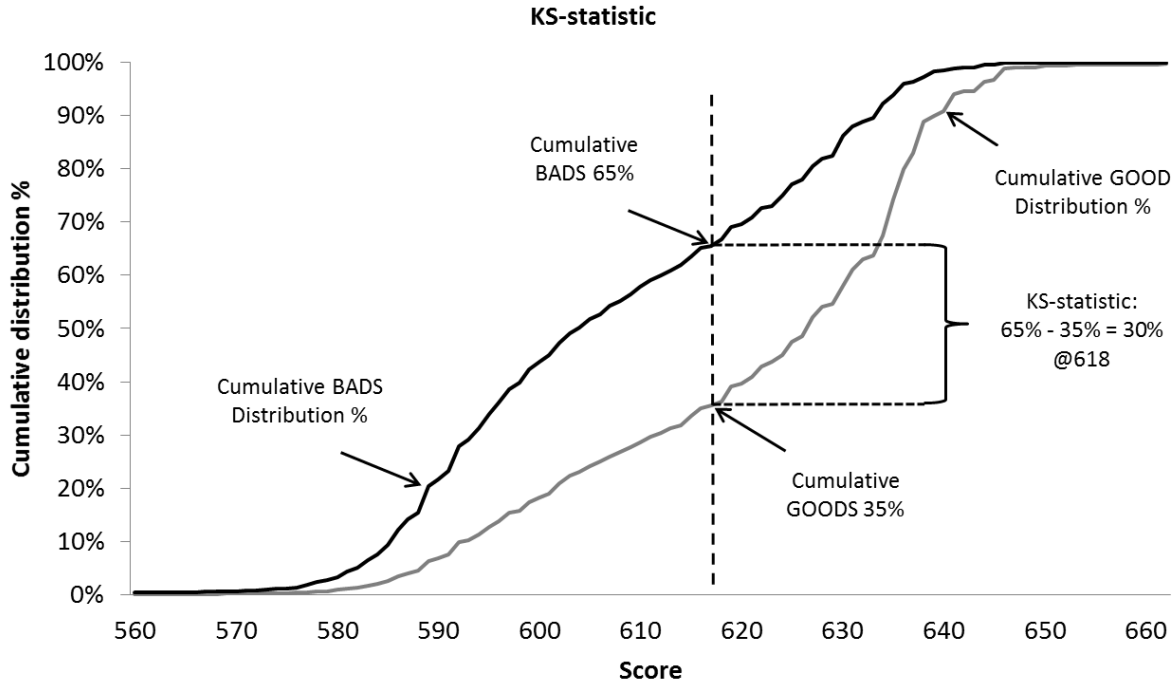


Figure 3.5: KS-statistic.

In Figure 3.5, the cumulative good and bad population is plotted against the score. The greatest difference between the two distributions is at a score of 618. At a score of 618 the cumulative bads distribution percentage is 65% with the cumulative goods distribution percentage of 35%. The KS-statistic is then given as 30% (65% - 35%) at a score of 618.

A weakness of the KS-statistic is that it is determined at a single point which may not represent the expected cut-off. In addition, the KS-statistic is not measured over the entire score range which is the case for other measures (Siddiqi, 2006, p. 123). Care must be taken when using the KS-statistic to assess an application scorecard, especially when comparing KS-statistics between the accept population and the full population which include rejected accounts (Anderson, 2007, p. 197).

3.4.5 Gini-coefficient

Vilfredo Pareto (1896) noted that 80% of the land in Italy was owned by 20% of the population which was also the case in other instances (Anderson, 2007, p. 203). The 80/20 ratio is known as the “Pareto principal” or the “80/20 principle”. Lorenz (1905) took the work of Pareto further and developed a data-visualisation tool called the “Lorenz curve” illustrating income inequality (Anderson, 2007, p. 203). One contention of Pareto was that income inequality would reduce in richer societies, however, Gini (1910) proved Pareto wrong using the well-known Gini-coefficient (Anderson, 2007, p. 204). Figure 3.6 illustrates the Gini-coefficient and how it is determined.

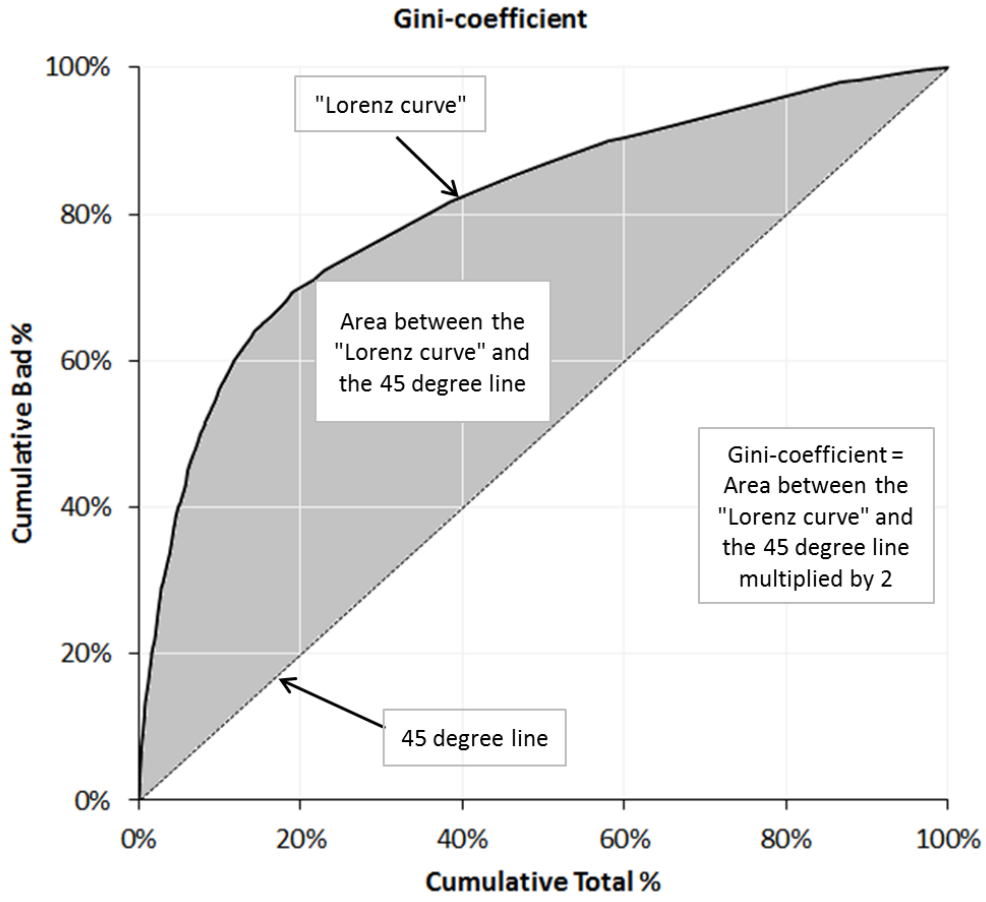


Figure 3.6: Gini-coefficient.

The Gini-coefficient is used in credit scoring to measure the accuracy of prediction of a score-card to separate good from bad accounts. The Gini-coefficient can be determined by firstly calculating the cumulative distribution percentage of the bad population; secondly the cumulative distribution percentage of the total population is calculated, thirdly the cumulative distribution percentage of the bad population is plotted against the cumulative distribution percentage of the total population, the Gini-coefficient is the area between the “Lorenz curve” and the 45 degree line and then doubled to account for the total population. The higher the Gini-coefficient the better the scorecard is able to separate the bads from the goods (Siddiqi, 2006, p. 125).

In Figure 3.6, the 45-degree line indicates perfect equality (no predictive power) meaning that a scorecard is not able to separate the bads from the goods and perfect inequality (predictive power) is represented by the area above and below the 45 degree line indicating that the scorecard is able to separate the bads from the goods. Anderson (2007) mentioned that there is sensitivity around the calculation of the Gini-coefficient, firstly it can be exaggerated by an

increasing indeterminate population and secondly it can be sensitive to category definitions regarding contents, number and ordering (Anderson, 2007, p. 204). An advantage of the Gini-coefficient is that it is calculated over the entire range of the scores which is not the case for the KS-statistic. In retail application scoring an acceptable Gini-coefficient of above 50% is satisfactory while a Gini-coefficient below 35% is questionable and 30% unacceptable (Anderson, 2007, p. 205). These Gini-coefficient thresholds are dependent on information used when building the scorecard.

A similar statistic to the Gini-coefficient is known as the c-statistic or the area under the receiver operating characteristic (AUROC) represented in Figure 3.7. The AUROC is based on two concepts namely the ability to mark true positives (sensitivity) and ability to identify true negatives (specificity) (Anderson, 2007, p. 206). As with the Gini-coefficient, the AUROC is also determined over the entire score range of the scorecard. In Figure 3.7, the 45-degree line indicates an AUROC of 50% representing a "random model"; hence for a scorecard to be better than a "random model" the AUROC must be greater than 50%. An AUROC of 70% or above is seen as adequate (Siddiqi, 2006, p. 124). The Gini-coefficient is often referred to in the world of credit scoring and even in credit scoring software such as FICO model builder.

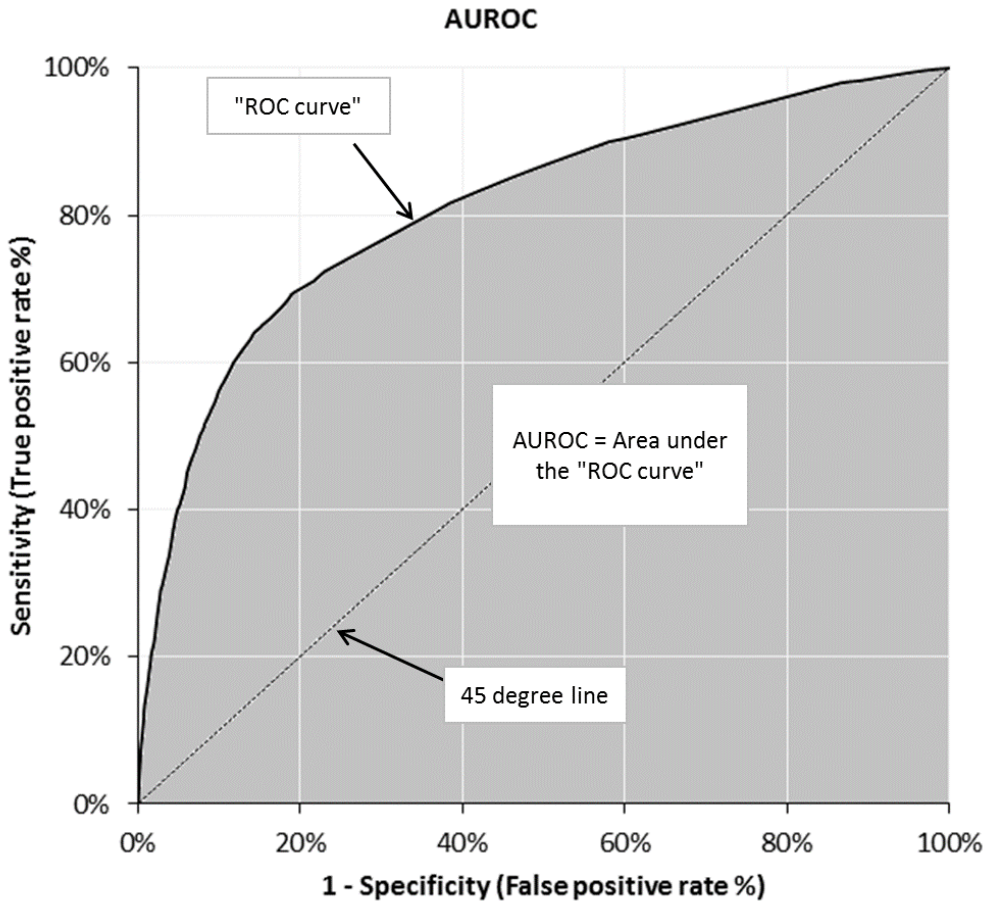


Figure 3.7: AUROC.

3.5 Data and methodology

3.5.1 Data

Retail banking data were used in the research study based on a bank in South-Africa. External credit bureau data were used where applicable and based on credit bureau data in South-Africa specific. Table 3.1 presents the frequency and source data used in more detail.

Table 3.1: Data.

Topic	Data requirement	Source
Scorecard statistical measures for development and post implementation	- Scorecard development data - Post implementation data	Retail bank in South Africa

Data availability for Table 3.1 includes data from January 2002 up to and including December 2014.

3.5.2 Methodology

In this section, a statistical measure to compare performance more accurately from development against post implementation data known as the "swap-set Gini-coefficient" is introduced. The methodology to obtain this measure is presented in Figure 3.8 and explained in the sections that follow.

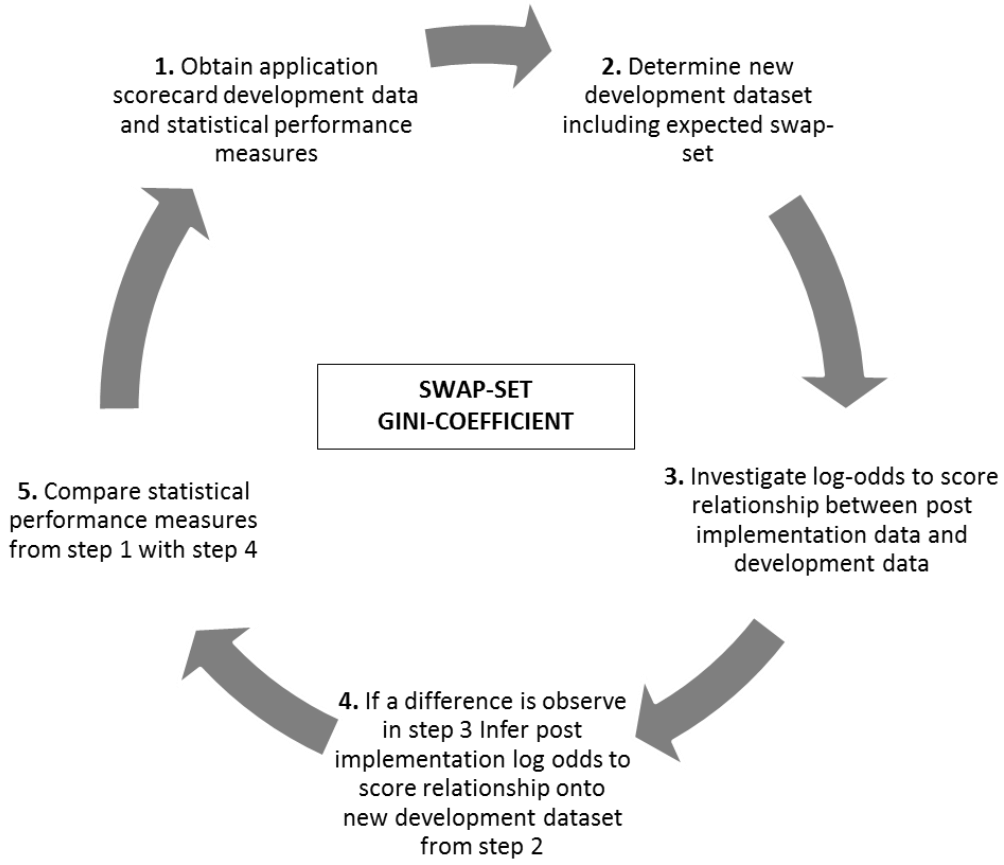


Figure 3.8: Methodology.

3.5.2.1 Development data

The first step of the methodology is to obtain the development data on which the application scorecard was build which will drive the new business decisions. Statistical measures such as the Gini-coefficient, the KS-statistic and the divergence can be computed from the development data for the accept population.

3.5.2.2 Swap-set

Application scorecards indicate approved applications and declined applications. Building a new application scorecard change the decision of a subset of the applications called the swap-set. To illustrate this more simply, see Figure 3.9.

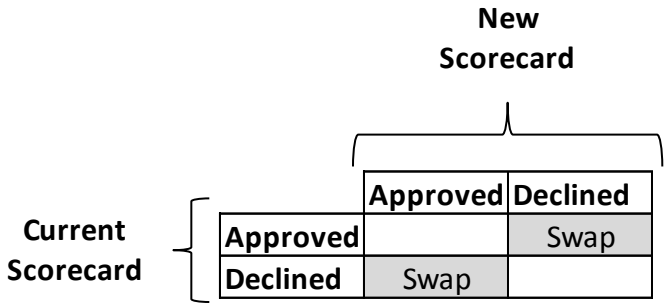


Figure 3.9: Swap-set.

In Figure 3.9 (comparing the current application scorecard with the new application scorecard) the swap-set indicates that under the current scorecard we have approved decisions yet under the new scorecard it would be declined and under the current scorecard we have declined decisions whereas under the new scorecard it would be approved.

The aim of this step is to re-create an accepts population for the development data which takes into account this swap-set as these accounts would be classified as approved or declined differently from the current scorecard when the new application scorecard will be implemented.

3.5.2.3 Score-to-Log Odds

A method to illustrate the monotonic increase in odds against increase in score is the log odds to score relationship. An important concept in scoring is that a scorecard measures the *propensity* to go bad and not the *probability* to go bad. However, under the advanced internal ratings based (IRB) approach under Basel II the point in time (PIT) probability of default (PD) component needs to be calibrated to the through the cycle (TTC) PD to ensure stable capital ratios TTC (Bonini & Caivano, 2014, pp. 41-42).

In September 2006 the Bank of international Settlements (BIS) indicated some divergence between IRB components and estimates used for internal purposes. An example of such divergence is pricing which is more likely to use estimates based on the life of the asset (Basel Committee on Banking Supervision, 2006, pp. 1-2). In this paper the focus is primarily on credit scoring from an acquisitions point of view for secured retail banking, meaning that the estimates of the IRB PD component for capital purposes should be regarded as independent from the credit scoring point of view for acquisitions which uses lifetime outcome of the asset (in other words measuring the propensity to go bad).

A quick method to test whether an application scorecard works is to assess whether there is a monotonic increase in log odds as the score increases, however the magnitude of the odds could be dependent on the macroeconomic environment, i.e. the score to log odds relationship in a downturn period could be different to a benign or upturn period.

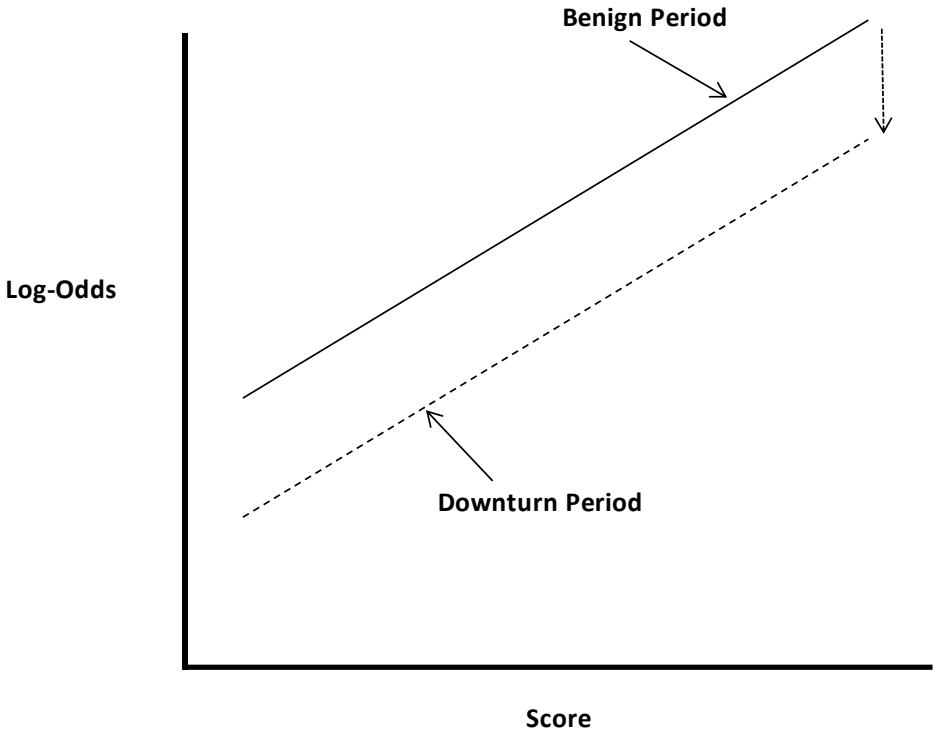


Figure 3.10: Log-odds to score relationship.

Figure 3.10 provides a simple illustration that can be expected when comparing the log odds to score relationship for a downturn period against a benign period. Firstly, both log odds to score relationships have a monotonic increase in odds as the score increases. Secondly the log odds to score relationship for the downturn period is on a lower level when compared to the benign period score to log odds relationship (expected in difficult economic times).

3.5.2.4 Inference

The aim of this step is to infer the accept population which was obtained from step 2. The accept population in step 2 was created by taking the swap-set into account which is to be expected once the new application scorecard is implemented. Taking the accept population from step 2 inference needs to be applied such that the macroeconomic performance from the log odds to score relationship are taken into account which is experienced post implementation of the scorecard.

3.5.2.5 Compare statistical performance measures

In step 5 of the methodology presented in Figure 3.8 one is able to compare the statistical performance measures from development against the measures post implementation for the accept population. These results should indicate the performance of the application scorecard which can be used for monitoring purposes.

3.6 Results

The aim of this section is to use the methodology presented in Section 3.5 to obtain statistical performance measures that can be used for monitoring the accept population of an application scorecard post implementation more effectively (taking scorecard implementations and the macro-environment into account within the statistical performance measure).

3.6.1 Development data

In this section we carry out step 1 from the methodology presented in Figure 3.8. Figure 3.11 presents the various application scorecards development periods and implementation dates for a retail bank in South Africa.

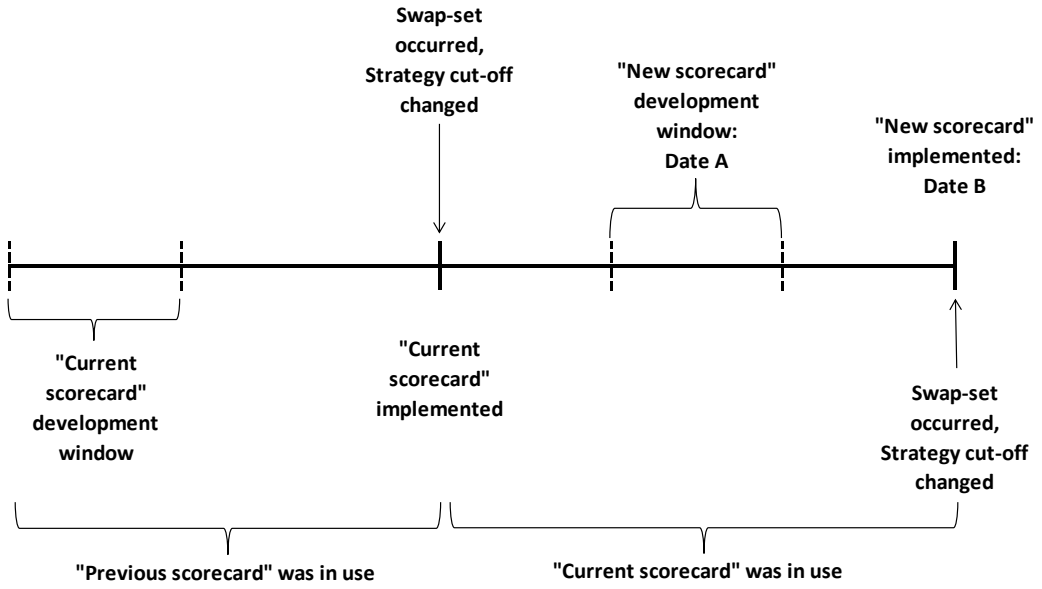


Figure 3.11: Scorecard timeline.

In this section, focus will be on the development window for the “New Scorecard” at “Date A” and post implementation at “Date B”. The statistical performance measures for the application scorecard for the bank are first presented for both the development window for the “New scorecard” as well as for data one year post implementation (see Table 3.2).

Table 3.2: Statistical performance measures (accepts population).

	Development	Post-Implementation
Actual bad rate %	0.08	0.04
Gini-coefficient	0.36	0.48
KS-statistic	0.29	0.38
Divergence	0.54	0.91

Table 3.2 presents performance measure from development and post implementation of the new application scorecard. The Gini-coefficient is shown in Figure 3.12.

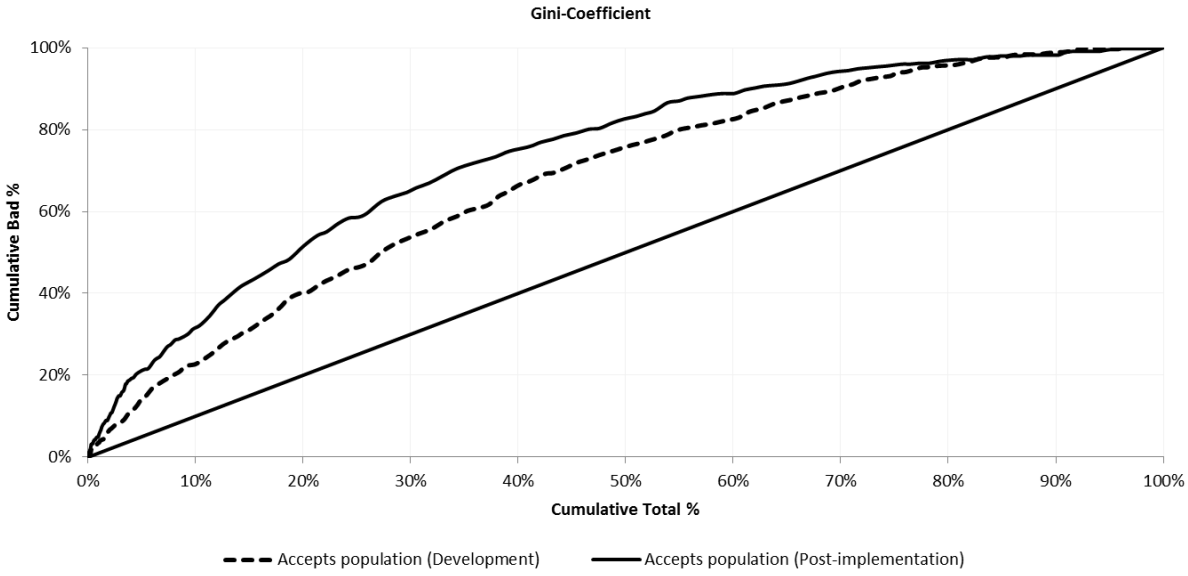


Figure 3.12: Gini-coefficient.

It is evident from Table 3.2 that a significant difference between the performance measures for the accept population exists. From Figure 3.11 it is evident that the development window for the "New scorecard" was selected in a period when the "Current scorecard" was in use. Huang and Scott (2007) observed a loss in power when comparing the development sample (selected prior when "Current scorecard" went live) with the out-of-time (OOT) window (selected post when "Current scorecard" went live) with reject inference not the root cause, but population shift. The retail bank indicates an increase in predictive power for the "New scorecard" comparing development against post-implementation, with the "Current scorecard" in use at the time of development.

A possible reason explaining the increase in predictive power is that the application scorecard in question is a *new bureau-based* application scorecard which is a different version to the bureau scorecard that was used at the point of development. The aim of this section is to create statistical performance measures to compare the predictive performance of the new scorecard post implementation against development.

3.6.2 Swap-set

In this section, step 2 from the methodology presented in Figure 3.8 is carried out. Risk categories are used by the South African retail bank to indicate the credit risk level. These risk categories drive approval and decline decisions. Table 3.3 provide the risk categories which were used during the development window which effectively determined the accept population under the current scorecard in use at the time and is captured under “Old risk categories”. In Table 3.3, the “New risk categories” is based on the new application scorecard which was built in the development window and drives the approval and decline decisions post implementation which effectively determines the final accepts population post implementation.

Table 3.3: Risk categories.

Old risk categories	New risk categories															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1	578	738	342	243	100	25	12	19	.	.	.	.	.	.	.	.
2	303	497	342	379	466	199	62	12	.	.	6	12	.	.	.	.
3	1575	3096	1967	2197	1518	840	555	261	324	37	25	25	12	.	.	.
4	205	1057	951	1126	1493	932	647	472	311	87	19	25	37	12	.	.
5	100	1250	1355	2060	2314	1792	1188	822	940	341	224	105	137	199	.	6
6	37	616	841	1780	2438	2408	2074	1959	1997	803	423	255	230	472	37	50
7	.	143	324	972	1374	1617	1507	1398	1674	716	684	367	324	733	131	50
8	.	87	236	603	1194	1297	1525	1543	1822	921	615	472	380	1050	205	87
9	.	37	75	336	958	964	1305	1691	2238	1249	832	547	603	1540	372	335
10	.	.	62	317	610	1025	1263	1703	2087	1654	1268	940	926	2635	739	572
11	.	.	.	75	199	173	268	410	397	435	428	155	341	942	348	149
12	.	.	.	37	50	162	193	392	528	460	323	193	385	1496	334	440
13	.	.	19	118	236	416	502	780	1027	812	1028	544	687	3889	1318	1615
14	.	.	.	43	30	166	203	430	694	571	578	320	399	1491	362	252
15	.	.	.	6	24	18	56	111	154	147	165	172	166	969	521	957
16	.	.	.	6	18	79	165	269	599	752	764	611	899	6075	2872	4017

From Table 3.3 it is clear that a swap-set occurs between old and new risk categories. As an example, an application that was in risk category 10 on the old risk category could now be part of the new risk Category 3 up to new risk Category 16. In essence, the swap-set takes into account the change in distribution of the accepted population which can have an impact on the predictability of the new application scorecard on the accept population.

3.6.3 Log odds to score relationship

In this section, step 3 from the methodology presented in Figure 3.8 is carried out. In Figure 3.13, the 12-month default rate is plotted over time, firstly the development window of the

new application scorecard was in a period where there were still elements of a downturn period as indicated by the high default rate % for the development period.

The implementation of the new application scorecard occurred in 2012 where, from Figure 3.13, it is evident that the default rate % was at a lower level than the development window.

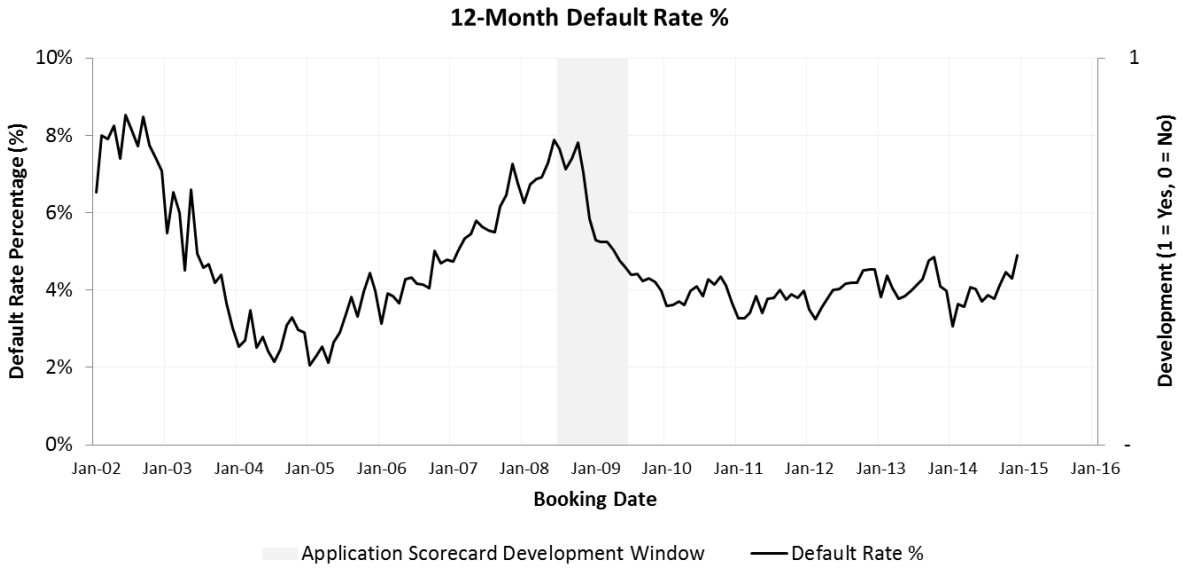


Figure 3.13: 12-Month default rate %.

Given the difference in the level of default rate between the development window and post implementation it is necessary to plot the log odds to score relationship for the accept population post implementation and for the accept population for the development window after the swap-set was taken into account as described in Section 3.6.2 to obtain a like for like comparison of accepts populations. In Figure 3.14 the log odds to score relation is plotted for both the development window after the swap-set and the post implementation data.

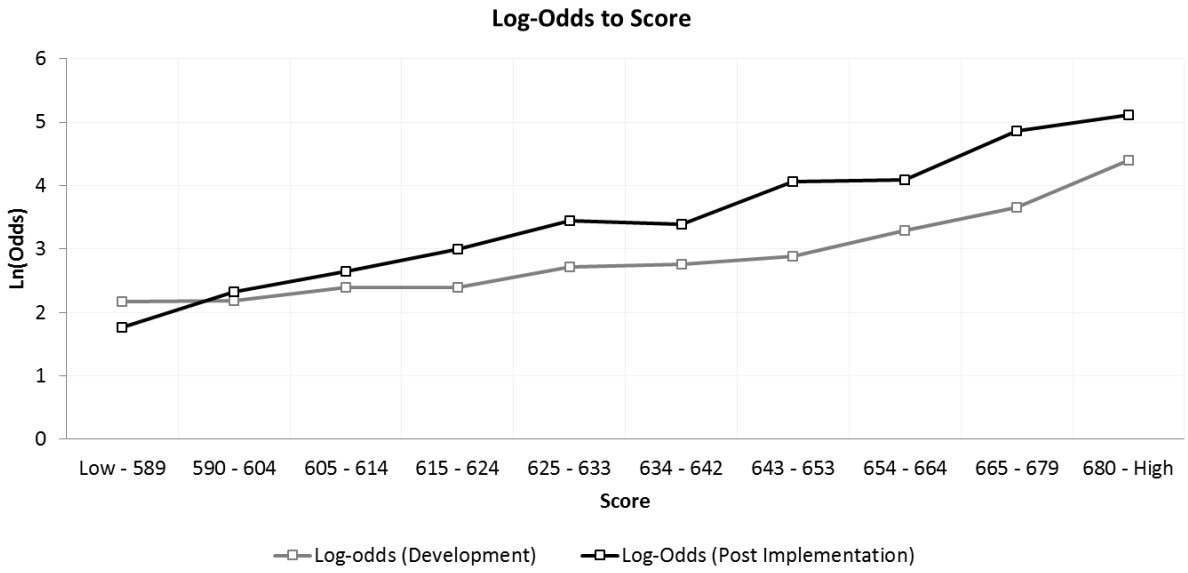


Figure 3.14: Log-odds to score.

The log odds to score relationship are monotonic for both the development and post implementation data indicating that the scorecard ability to measure propensity to go bad holds true. However, the level of log-odds between development and post implementation is different as a result of the cycle the bank is finding itself in and which is indicated by the 12-month default rate in Figure 3.13.

3.6.4 Inference

In this section step 4 from the methodology presented in Figure 3.8 is carried out. It was illustrated in Figure 3.13 that the development window for the new application scorecard was in the downturn period with the post implementation indicating a benign period with lower default rates.

It was confirmed in Section 3.6.3 that the new application scorecard is working as the log odds to score relationship is monotonic, however on different levels due to the macroeconomic environment. To obtain a statistical performance threshold for the development accept population it is necessary to infer the development data using the scorecard's post implementation performance, this will ensure that a statistical performance threshold is obtained for the development window which can then be used to monitor the accept population after implementation.

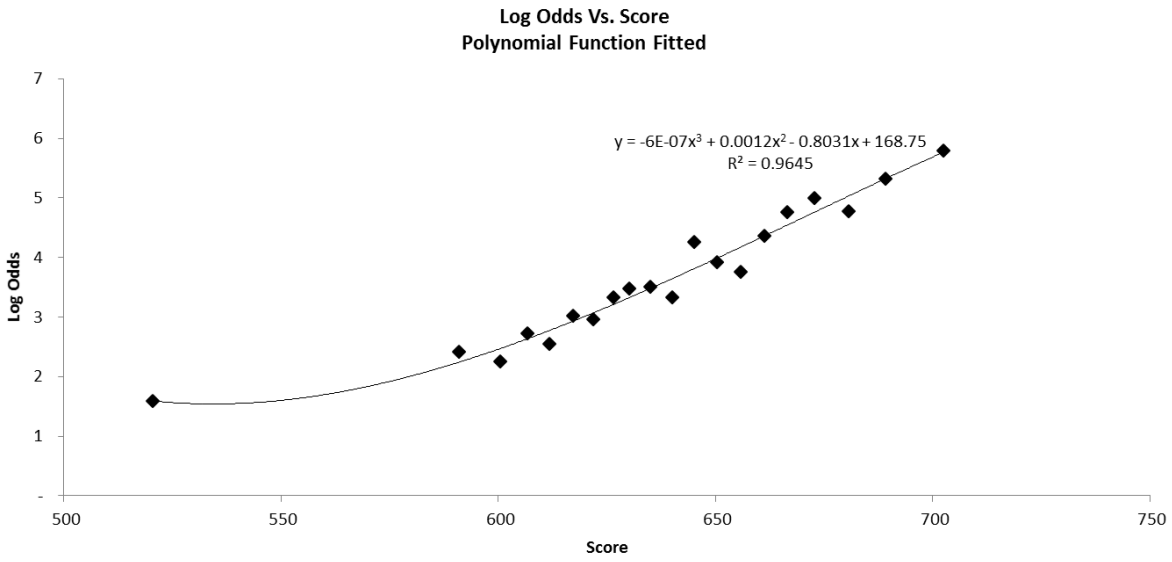


Figure 3.15: Fitted curve (Post-implementation).

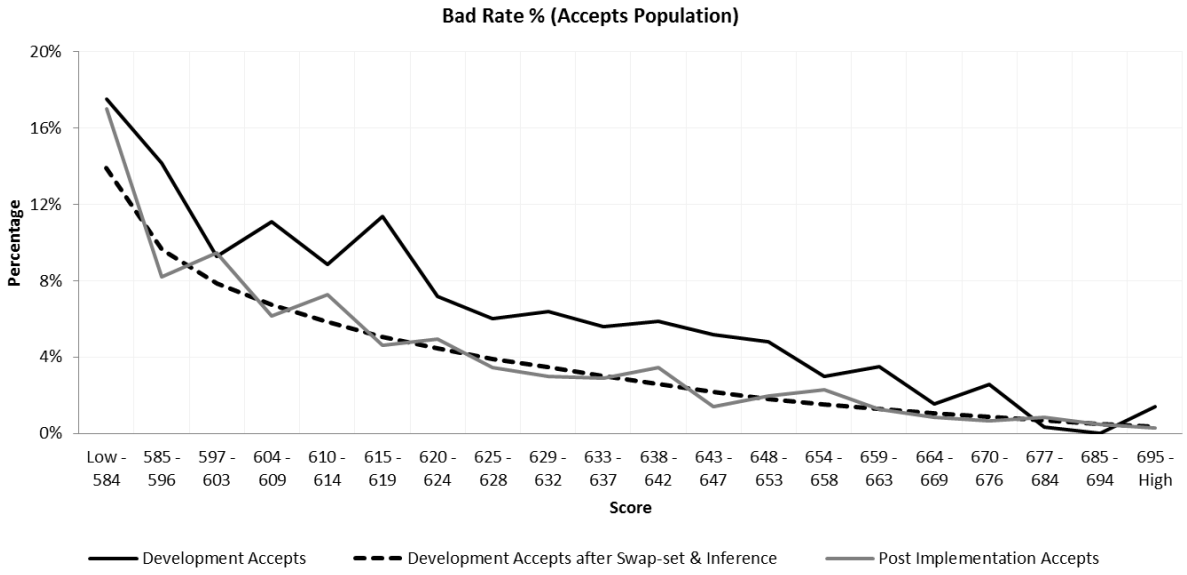


Figure 3.16: Inference.

Figure 3.15 presents the log-odds to score relationship for the accept population based on post implementation data. Given this log-odds to score relationship a polynomial curve was fitted to this relationship which was used to infer the development data from Section 3.6.2. Figure 3.16 illustrates three different line graphs, firstly the "Development Accepts" line graph which is the accept population from development with a 24-month bad rate of 7.90% as presented in Table 3.2, secondly the "Post implementation Accepts" line graph which is the accept population post implementation with a 24-month bad rate of 4.00% as presented in Table 3.2, and lastly the "Development Accepts after Swap-set & inference" line graph which is

the “Development Accepts” line graph after the swap-set described in Section 3.6.2 and after inference is applied. It is evident from Figure 3.16 that the inference was successful as the pattern for the “Development Accepts after Swap-set & inference” line graph follows the same level of bad rate as the “Post implementation Accepts” line graph.

3.6.5 Statistical performance measures comparisons

In this section we carry out the final step 5 from the methodology presented in Figure 3.8. In Section 3.6.2 the accept population from the development window was adjusted to take the swap-set into account which is expected after post implementation. In Section 3.6.3 and 3.6.4 the accept population from the development window after taking the swap-set into account was inferred using the log-odds to score relationship of the post implementation data as the development window represented a downturn period and the post implementation data represents a benign period. The statistical performance measures are presented in Table 3.4. The Gini-coefficient graphs are presented in Figure 3.17.

Table 3.4: Statistical performance measures (Accepts population).

	Development	**Post-Implementation**	**Development (After swap-set and Inference)**
Actual bad rate %	0.08	0.04	0.04
Gini-coefficient	0.36	0.48	0.45 (Swap-set Gini-coefficient)
KS-statistic	0.29	0.38	0.35
Divergence	0.54	0.91	0.77

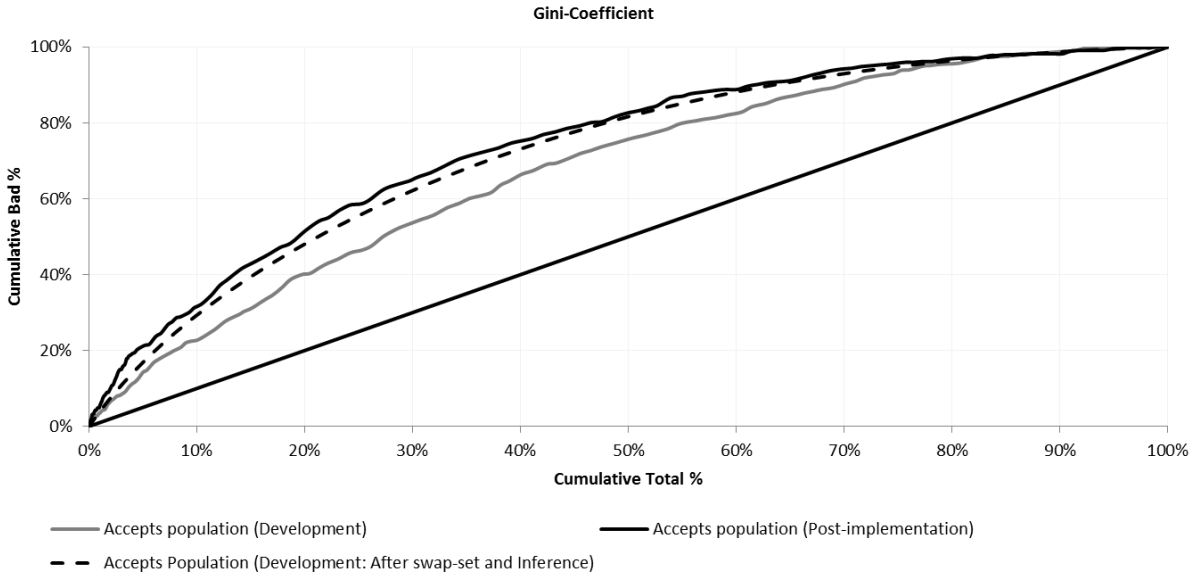


Figure 3.17: Gini-coefficient.

In Table 3.4 comparing the statistical performance measures for the accept population between the post implementation data and the development data after the swap-set and inference the results appear to be significantly closer than when comparing to the original accepts population from the development data. This makes intuitive sense as it is known that a swap-set occurred post implementation and that the scores from the new application scorecard performs differently although monotonically between a downturn period and a benign period. The statistical performance measures for the development data after swap-set and inference can be used to monitor the accept population post implementation. In Table 3.4 the swap-set Gini-coefficient of 45% can be compared to the post-implementation Gini-coefficient of 48%. This indicates that the scorecard separates good from bad accounts to a greater extent than at development.

A disadvantage of this approach is that performance data need to build up post implementation to carry out the inference process, however dependent on the outcome the application scorecard was built a regression between shorter outcome performance data and scorecard outcome performance data can be carried out to hasten the process. The advantage of the methodology described in this paper is the ability to compare statistical performance measures more effectively for the accept population at development against the post implementation data.

The key element soon after post implementation is to test the monotonic log odds to score relationship to observe whether the application score rank credit risk, once sufficient performance data is obtained post implementation the effectiveness of the application scorecard to separate good from bad accounts can be monitored comparing the statistical performance measures of the new benchmark thresholds determined after the swap-set and inference for the development data against the post implementation statistical performance measures. It is also important to note that if statistical performance measures are compared for the through the door population (which contains all applications) from development data against post implementation data it is also necessary to infer post implementation performance of application scores onto development data if there is a difference in the macro economy.

3.7 Conclusion

The acquisition of new business is critical for any bank which drives the futuristic quality of the book. This is where the application scorecard plays an essential role to separate good accounts from bad accounts. A problem that exists within application scoring is the possible deterioration of the predictive power of the application scorecard when performance measures are compared between development data and post implementation data for the accept population making the monitoring of the scorecard complex post implementation of the scorecard.

In this paper a literature review was provided which detailed the various statistical performance measures that are used to determine how well a scorecard separate good accounts from bad accounts. These measures of separation, power or ranking ability include the divergence statistic, the misclassification matrix, the KS-statistic and the Gini-coefficient. A methodology was presented in this paper to address the problem stated consisting of 5 steps; obtaining development data, determining the swap-set, investigating the log odds to score relationship, inference and the comparison of statistical performance measures.

The contribution from this paper is a methodology that can be used to compare and monitor statistical performance measures more accurately and effectively for an application scorecard between the development data accepts population and post implementation data accepts population. By having more accurate comparisons for performance the effectiveness of the application scorecard can be thoroughly realised. It is important to realise that an application scorecard for acquisition purposes measures the propensity to be a good or bad client and

not the probability to be a good or bad client. Changes in the economic cycle have an effect on the level of bad rate for the application scorecard and this was illustrated in this paper for the retail bank data in question.

Results used in this paper originate from a South African retail bank. The development data for the application scorecard were used to obtain the accept population performance measures and was compared against the accept population of the post implementation data. These comparisons indicated that there is a significant difference in the performance measures and were investigated. After applying the methodology introduced in this paper, the new threshold of statistical performance measures for the development accepts population was determined and gave intuitive results when compared to the post implementation accepts population. These new thresholds for comparison are a significant change in what was generally used when monitoring an application scorecard for the accept population.

Possible future research includes methods to obtain the level of bad rate for post implementation data early such that thresholds for statistical performance measures can be determined as soon as possible for the development population for monitoring purposes. Other possible research areas include the effect of statistical performance measurements comparisons for the through the door population between development and post implementation data.

Bibliography

Abdou, H. & Pointon, J., 2011. Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance & Management,* 18(2-3), pp. 59-88.

Anderson, R., 2007. *The credit scoring toolkit, theory and practise for retail credit risk. Management and decision automation..* United States: Oxford University Press Inc..

Basel Committee on Banking Supervision, 2006. *The IRB Use Test: Background and Implementation,* s.l.: Bank for International Settlements.

Bonini, S. & Caivano, G., 2014. Probability of Default: A Modern Calibration Approach. In: C. Perna & M. Sibillo, eds. *Mathematical and Statistical Methods for Actuarial Sciences.* Switzerland: Springer International Publishing, pp. 41 - 44.

Engelman, B. & Rauhmeier, R., 2011. *The Basel II Risk Parameters: Estimation, Validation, and Stress Testing - with Applications to Loan Risk Management.* Second ed. Berlin: Springer.

FICO, F. I. C., 2011. *Explanation of Divergence,* s.l.: FICO.

Gini, C., 1910. 'Indici di Concentrazione e di dipendenza'. Atti della III Riunione della Societ'a Italiana per il Progresso delle Scienze. Reprinted in his 1955 'Memorie di methodologia statistica, I, Variabiliá e Concentrazione'. Libreria Eredi Virgilio Veschi: Rome.. pp. 3-120.

Huang, E. & Scott, C., 2007. *Credit Risk Scorecard Design, Validation and User Acceptance - A Lesson for Modellers and Risk Managers,* s.l.: University of Edinburgh Business School.

Kolmogorov, A. N., 1933a. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer.

Kolmogorov, A. N., 1933b. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell Istituto Italiano degli Attuari*, pp. 483-491.

Lim, M. K. & Sohn, S. Y., 2007. Cluster-based dynamic scoring model. *Expert Systems with Applications,* Volume 32, pp. 427-431.

Lorenz, M. O., 1905. Methods of Measuring the Concentration of Wealth. *Publications of the American Statistical Association,* Volume 9, pp. 209-219.

Pareto, V., 1896. *Cours d'économie politique,* s.l.: Université de Lausanne, 3 volumes, 1896-1897.

Siddiqi, N., 2006. *Credit risk scorecards, developing and implementing intelligent credit scoring.* United States of America: John Wiley & Sons, Inc..

Thomas, L. C., 2000. A survey of credit and behavioural scoring: forecastingfinancial risk of lending to consumers. *International Journal of Forecasting,* Volume 16, pp. 149-172.

Van Gestel, T. & Baesens, B., 2009. *Credit Risk Management - Basic concepts: financial risk components, rating analysis, models, economic and regulatory capital.* New York: Oxford University Press.

Chapter 4
Non-capital calibration of bureau scorecards

Nico Kritzinger[1] and Gary van Vuuren[2]

ABSTRACT

Application scorecards play a critical part in determining the credit worthiness of applicants for acquisition purposes. However, the level of bad rate in a downturn period and upturn period although monotonic are different across the scores due to procyclicality. This paper investigates the performance of a bureau scorecard and compares the performance between a downturn period and upturn period. The comparison indicates that calibration is essential to account for procyclicality. The aim of this paper is to present a calibration model that can be used to adjust an application scorecard to a level of expected future bad rates.

Keywords: Credit risk, application scoring, calibration, credit risk management

JEL classification: C11, C20, G32

4.1 Introduction

Financing assets or consumer lending goes back long into history, making the credit-worthiness determination of a client critically important. The modelling of this credit worthiness is known as credit scoring which guides the classification of a bank's clients to reduce the propensity of a client to be a future bad debtor (Abdou & Pointon, 2011, p. 60). The most general form of credit scoring is known as application scoring which models the credit-worthiness of a client at the point of application to determine the propensity of the client to go bad in the future (Lim & Sohn, 2007, p. 427).

The credit bureau score which predicts whether a debtor will be bad in future can also be related to application scoring which is based on data provided by various credit organisations to the credit bureau. The credit bureau score takes into account the performance of borrowers with his/her various credit borrowing obligations. In South Africa four major credit bureaus exist which are known as TransUnion, Experian, Compuscan and XDS which build their

[1] PhD-student at the Faculty of Natural Sciences, North-West University.
[2] Visiting professor, Faculty of Natural Sciences, North-West University.

bureau scorecards on data received from the credit lenders that supply the data regularly to the bureaus. In general, the data collected by the credit bureaus are similar, however there could be differences in how the data are captured, displayed and stored (TransUnion, 2015). At TransUnion the bureau score is referred to as the Empirica score which has seven objectives namely evaluating a client's risk with expected performance, assessing risk at origination stage, assessing clients with no historical credit history, managing clients from a collection point of view, identifying cross-sell opportunities, growing the client base and ranking clients according to credit risk. TransUnion uses five categories of consumer information to build the Empirica score to assess credit risk which are demographics, client judgments, client default experiences, client enquiries and client payment profiles (TransUnion & FICO, 2009, p. 2).

As credit scoring aids as a tool for the banking industry to classify credit worthiness of clients, economic upturns and downturns could drive the level of this credit worthiness which in essence determines the level of impairments and profitability. The Bank for International Settlements (BIS) introduced three approaches for identifying appropriate downturn periods which are related to GDP growth, default rates and recovery rates (Basel Committee on Banking Supervision, 2005, p. 3). Figure 4.1 illustrates the GDP growth of South Africa from 2003 to 2015: negative GDP growth was realised in the latter part of 2008 and the first half of 2009.

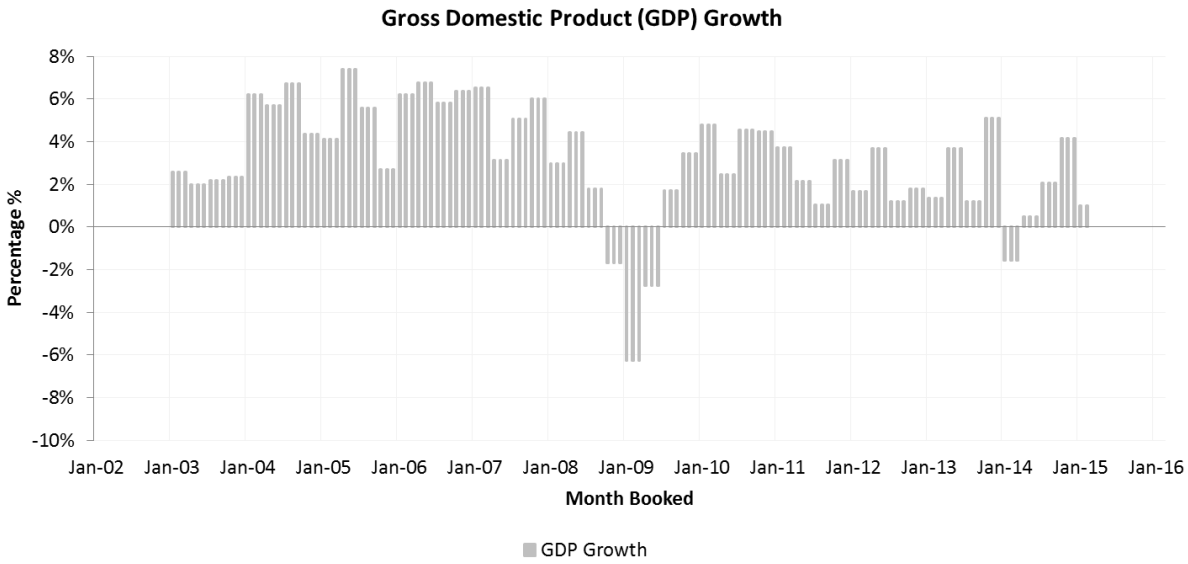


Figure 4.1: South African GDP Growth.

Anderson (2007) asked whether a credit scorecard can be built in one point of an economic cycle and used in another with the answer usually yes. Credit scorecards are in general robust and do not have to be discarded with changes in the economy (Anderson, 2007, p. 84). The level of credit-worthiness in economic upturns and downturns can be related to a term called procyclicality. The problem of procyclicality surfaced during the introduction of the capital accord which aims for capital stability. Banks responsiveness to higher default risk leads to higher capital during a recession and a decrease in capital during benign periods hence producing the procyclicality problem (Gordy & Howells, 2004, p. 1). Procyclicality is the fluctuation of financial characteristics around a trend in an economic cycle. An increase in procyclicality indicates broader amplification in fluctuations (Landau, 2009, p. 1).

Banks have endured crises in the past making it crucial to have more reliable risk management tools. Under Basel II, the aim of the parameter probability of default (PD) is to calibrate the point in time (PIT) PD to the through the cycle (TTC) PD to ensure stable capital ratios TTC. In this way capital should be kept in positive economic stages which can then be used in negative economic stages (Bonini & Caivano, 2014, pp. 41-42). This paper does not aim to focus on capital, but to use this concept of calibration combined with application scorecards for future acquisition purposes. The power of a scorecard is the extent to which defaults are avoided when classifying good borrowers. However, even though a scorecard can have strong power, calibration is needed to match actual default rates (Bohn & Stein, 2009, p. 362). In this paper, the procyclicality effect on a bureau scorecard is investigated and using a calibration technique to adjust to future expected default rates. Given this approach the bureau score cut-offs can be adjusted given the macroeconomic stage such that it can be used for acquisition purposes.

4.2 Problem statement and objective

The aim of this article is to illustrate the effect of a changing economy on a bureau scorecard in South Africa and investigating techniques for calibration which can then be used to calibrate the internal rating model to an expected future bad rate level for acquisition purposes. The primary objectives of this study are to improve application scoring for acquisition purposes and to enhance credit risk measurement and management. The focus includes credit

risk, application scoring, credit risk management and calibration. Methods to identify downturn periods and calibration techniques are investigated and the effect of the proposed primary objectives on credit risk management is assessed.

4.3 Literature review

4.3.1 Scenarios affecting scorecards

In building a credit scorecard, the base assumption is made that the past will be repeated in the future. Although this assumption is never true, it usually suffices when building a scorecard (Anderson, 2007, p. 83). Mark Twain observed that "History never repeats itself, but it does rhyme" (Collins, 2016, p. 1). Scorecards can be affected firstly by the economy with changes in interest rates, unemployment rates, GDP and other factors causing downturn and upturn periods. Secondly, scorecards can be affected by the market where changes in product features and customer demographics play a role. Thirdly, changes to processes, calculations, forms or systems affects scorecards operationally. The manner in which the behaviour of people changes over time can also affect scorecards and lastly changes witnessed that cannot be explained except for the age of the scorecard can affect a scorecard (Siddiqi, 2006, pp. 167-168; Anderson, 2007, pp. 83-90).

Business cycle and downturn period

A business cycle is defined as a period comprising expansions in economic activity followed by a period of a recession or downturn, contractions and revivals merging into the expansion phase of the next cycle (Burns & Mitchell, 1946, p. 3). The business cycle was more clearly defined by Beveridge and Nelson (1981) and Nelson and Plosser (1982) as the residual part of a time series after the extraction of the permanent component and while the trend represents the non-stationary component the business cycle is stationary constrained (Turri & Salis, 2010, p. 4).

In June 2004, the BIS introduced a new framework for capital measurement and standards (commonly known as Basel II). The credit capital requirement formula for other retail exposures is:

$$K = LGD \cdot N\left(\frac{1}{\sqrt{1-R}} \times G(PD) + \sqrt{\frac{R}{1-R}} \times G(0.999)\right) - PD \cdot LGD \qquad (1)$$

where:

K represents the capital requirement, LGD is the loss given default, $N(x)$ denotes the cumulative distribution function for a standard normal random variable, $G(x)$ denotes the inverse cumulative distribution function for a standard normal random variable, PD is the probability of default and R the correlation.

A problem with (1) is that it does not incorporate correlations between PD and LGD leading to the underestimation of capital, hence the BIS requirement that the LGD should reflect downturn period conditions from a business cycle to be more conservative (Miu & Ozdemir, 2006, p. 44). Under paragraph 468 of the capital framework from BIS economic downturn conditions was referred to when estimating LGD (Basel Committee on Banking Supervision, 2004; Basel Committee on Banking Supervision, 2006, p. 103). In July 2005, the BIS introduced a guide to paragraph 468 of the capital accord framework document which presented three possible approaches when identifying appropriate downturn conditions (Basel Committee on Banking Supervision, 2005, p. 3). Firstly, periods of negative GDP growth and elevated unemployment rates, secondly, periods of elevated default rates observe from historical default rates and lastly, periods where drivers that influence default rates and recovery rates are distressed (Basel Committee on Banking Supervision, 2005, p. 3).

4.3.2 Calibration

Under the BIS capital framework, capital should reflect TTC to ensure stability, i.e. in a benign period capital should be kept for negative economic conditions. This requires that the PD parameter in the capital formula reflect long run historical default rates. Application scorecards (or PD rating models) are typically based on client individual characteristics and information regarding the product and bureau information. No macroeconomic information is generally used when assembling an application scorecard. Calibration is thus required to adjust PD rating models to reflect long-run historical defaults rates under the regulatory capital framework. Calibration applies an “addendum” or adjustment to the PD rating model itself (Bonini & Caivano, 2014, pp. 41-42).

Glößner (2003) introduced a simple seven step methodology to estimate and calibrate PD for Basel II purposes by counting relative defaulting frequencies of a typical retail portfolio. In step one a risk horizon of one year is defined to observe a random default process. In the next

step a rating system characteristics and credit scores are assessed at the beginning of the risk horizon defined. In step three every single borrower that defaulted is observed within the risk horizon period defined. The next step uses histograms in which Glößner distinguished between two possibilities at the end of the risk horizon. The first possibility is the projection method in which for each credit score for the borrowers at the start of the risk horizon a histogram is drawn indicating the borrowers that defaulted and the total borrowers in the risk horizon. The second possibility is a multivariate method in which for every characteristic for the borrowers at the start of the risk horizon a histogram is drawn indicating defaulted borrowers and the total borrowers in the risk horizon.

In step 5 the PD is determined from the histograms by dividing the number of defaulted borrowers by the total number of default borrowers. This result in a PD function which is either dependent on the credit score or the characteristics. This function is only valid for the risk horizon defined with the supposition that the defaults that occurred are not just random, however these default data can look totally different in the following new risk horizon defined, but represent a stable structure over time or at least slowly changing.

In the remaining two steps the PD function is used to predict the future. In step 6 when a new borrower arrives, his characteristics and credit score from the rating system are assessed. Based on either the characteristics or credit score the borrower is assigned a PD obtained from the PD function. Afterwards (step 7) the borrower is taken as part of the cohort of borrowers and contributes to the PD function which is calibrated over the next time interval following the risk horizon. After the investigation of the histograms, Glößner (2003) indicated the shortcomings of this simple methodology where the scarcity of data (especially default data) is a problem.

Glößner (2003) indicated one method to overcoming this is to use suitable continuous densities instead of histograms. Glößner (2003) further stated that modelling data histograms as continuous densities means to postulate a structure calibrated by the sample data, in essence claiming to know the default structure of borrowers. However, the histograms observed indicated how problematic it is to assume this structure. To overcome this Glößner (2003) referred to the modelling of the Lorenz curve (Glößner, 2003, pp. 7-11). The Lorenz curve may be used for measuring the discriminative power of a rating system, however Glößner (2003) constructed the Lorenz curve to use it to calculate PD (Glößner, 2003, p. 16). Glößner (2003)

states that the borderline between structure and randomness is much more stable at the Lorenz curve (Glößner, 2003, p. 17). Glößner (2003) constructed a numerical Lorenz curve from the univariate histogram and fits it with a parameterised function. When a fit from the numerical Lorenz curve is found it is differentiated, combined with the cumulative distribution of all borrowers and multiplied by the average PD of the portfolio to move from the Lorenz curve to a PD. Glößner (2003) mentions that even though the Lorenz curve is a suitable vehicle to separate randomness from structure, it relies on the scoring mechanism which projects the characteristics of the borrower onto the credit score. These ideas led to a setup of a valid and stable PD estimation system, (Glößner, 2003, pp. 28-31). The fitting procedure of the Lorenz curve still has elements of trial-and-error. This can be improved by obtaining a clearer view of the structure of the invariancy group of the Lorenz curve which would require a more comprehensive definition of the notion of invariancy. Glößner (2003) still claims that the Lorenz curve approach exhibits an advantage over "automatic" approaches such as neural networks (Glößner, 2003, p. 64).

Bonini and Caivano (2014) summarised the main findings from previous limited research on the calibration topic focussing on large corporate portfolios. Pluto and Tasche (2005) estimated PD using the most prudent estimation principle based on upper confidence intervals ensuring PD ordering that reflects differences in credit quality. Although this methodology is easy to apply, the underlying assumption requires independent default events. Kiefer (2008) made use of the Bayesian approach for PD estimation. Firstly, using an expert rank method for default probabilities to construct a prior distribution and secondly using Bayes theorem computing a posterior distribution. The expected value of the posterior distribution is then used for the PD estimation. Iqbal and Ali (2012), by calculating Bayesian and real probability for each rating class, proposed a convolution methodology. Under this methodology, an implied distribution is generated using convolution techniques. Tasche (2013) demonstrated that upper confidence bounds can be presented as quantiles of a Bayesian posterior distribution in the case of independent default events. Tasche (2013) suggested a constrained uninformed Bayesian estimator as opposed to the upper confidence bound estimators (Bonini & Caivano, 2015, pp. 4-5).

To ensure consistent meaning across scorecards Anderson (2007) presented two ways on which calibration can be achieved namely banding and scaling. Banding identifies score

ranges of common risk and generally ranges between five and 25 groups. Anderson (2007) presented three approaches to obtain the optimal number of groups in the banding process. Firstly the Calinski-Harabasz statistic which is an algorithm generally used for determining the optimal number of clusters which can be used to compare different grouping options. Secondly benchmarking where the scores are mapped to a predefined set of groups. This requires the identification of breakpoints to provide the best possible fit. Lastly marginal risk boundaries which is similar to the benchmarking approach except that upper and lower boundaries for each risk group are set and not based on the average (Anderson, 2007, pp. 419-420).

Although banding is an effective way to make scorecards consistent many lenders dislike the loss of granularity. As one can always create more bands, more is often not good enough. Scaling transform scores onto a consistent scale and has the advantage of allowing maximum granularity. Before scaling is carried out, Anderson (2007) indicates that a linear shift needs to occur where scores are aligned to a common definition. The most demanding feature from scaling is that final scores must always (or usually) be positive. Anderson (2007) provided two scaling equations. The *log-reference equation* is used when a reliable odds estimate already exist. The *linear* transformation equation is used when probabilities for each score are not available, but available for reference scores towards upper and lower ends of the score range, dependent on a linear relationship between credit quality and score (Anderson, 2007, pp. 419, 424-427).

Medema, Koning and Lensink (2008) proposed a validation methodology that can be used to validate credit scoring models under the Basel II accord. The methodology was contextualised by investigating and applying it to mortgage loans of a commercial bank (Medema, et al., 2009, pp. 701-702). Medema et al. (2008) indicated that a model is well calibrated if the fraction of events which actually occur, is unbiasedly estimated by the estimated probability of these events (Medema, et al., 2009, p. 704). Medema et al. (2008) briefly described a refitting technique to determine calibration and a graphical tool called the calibration plot to also determine the calibration. In the calibration plot the estimated default probabilities is plotted against the actual outcome and is useful to determine in which region of the estimated probabilities the model provide a good fit (Medema, et al., 2009, p. 705). As the estimate of the calibration from Medema et al. (2008) was based on a non-parametric smoothing method, no

straight forward numerical summary was presented for calibration (Medema, et al., 2009, p. 707).

Bohn and Stein (2009) emphasised that the first objective when building a credit scorecard is to obtain a high degree of predictive power. A poorly calibrated PD model can still rank obligors since credit scorecard has predictive power. Calibration of PD models (credit scorecards) leads to PD models which work well in portfolio risk estimation and valuation. A PD model that is well calibrated is one in which the estimated PD is an unbiased estimate of the actual PD. Although most models such as logit or probit produce natural probability estimates, most practitioners still find it necessary to calibrate these probabilities further. The first reason for this is because credit data (especially default data) tend not to follow regular distributions which often result in the underlying assumptions of many econometric approaches not being met in practice. Often it is still the case that the ranking of the model probabilities is highly correlated with the ranking of the empirical probabilities, therefore if one can transform the mapping from the ranking to the probabilities while still preserving the ranking, the output from the model will better reflect the correct probabilities.

A second reason why calibration is necessary is because the development data used to build the model (credit scorecard) are not fully representative of the data which will be used by the model. Bohn and Stein (2009) indicated that calibration typically consists of two steps. The first requires mapping the scores of a model to historical empirical probabilities. The second step requires an adjustment for the difference between historical empirical default rates and actual default rates (i.e. the probability needs to be adjusted to reflect the true prior distribution). Bohn and Stein (2009) suggested simple calibration techniques in situations when one needs to map model probabilities into empirical probabilities when the assumptions of the econometric model are not met. This technique is called nonparametric density estimation which can improve the alignment between model predictions and actual default probabilities. The flexible form of the density estimation takes care of the nonlinearity relationship between model probability estimate outputs and empirical default frequencies related with those outputs. This cost of alignment leads to complexity and makes the analysing of driving factors difficult. This result in calibration in general being performed after the model is fitted instead of in parallel with this process (Bohn & Stein, 2009, pp. 215-217).

Bohn and Stein (2009) indicated that calibrating to agency (or internal) ratings is not similar to calibrating to default probabilities, as ratings are not directly observable such as default events. Ratings are assigned by rating agencies though their own analysis and serve multiple purposes with the default probability being only a single rating attribute. Several approaches exist to calibrate to ratings all typically involving mapping a model output to a rating class. When calibrating to ratings, information is discarded because PDs offer higher resolution. However many organisations require calibration to ratings as staff are more familiar with agency (and internal) rating scales and policies are frequently based on ratings. Bohn and Stein (2009) advises that when a model produces a PD it is best advised to use PD-based policies especially in valuation and risk management.

Whenever rating mappings are used, users must also be aware of how such mappings are constructed and which information is embedded in the ratings. Bohn and Stein (2009) presented three approaches when calibrating to ratings. In the first approach, mapping is done using historical default rates. Bohn and Stein (2009) indicated that it is most common in practice for users to map PDs to historical default rates. To achieve this one needs historical default probabilities for each rating class after which cut-offs are determined for each rating class based on the default probabilities of the respective rating class. Typically the upper cut-off of a rating class is the lower cut-off of the next higher rating class.

There are three methods to determine the cut-offs, firstly actual cut-offs make use of historical probabilities of each rating class as lower bounds for each rating class. The second method is the arithmetic mean where the midpoint between the current and the next lower rating is used as a lower cut-off, and the midpoint between the current and the next higher rating is used as the upper cut-off. The third method is the geometric mean which follows the same approach as the arithmetic mean, however the geometric mean is used instead of the mean for two adjacent rating classes.

The most conservative method is the actuals followed by the geometric mean and then the arithmetic mean. This mapping through historical default rates approach equates the probabilities predicted by a model to the default rate associated with a specific rating. This approach, however, is an indirect mapping as it does not directly map from the model to the rating. This approach equates the model probability estimate to a rating scale attribute and

relies strongly on both probability estimates and PDs. Since PD is only one attribute of a rating, this approach could miss important ratings characteristics (Bohn & Stein, 2009, pp. 227-230).

A second approach of Bohn and Stein (2009) to calibrate to ratings is the mapping to quantiles of rating distribution. In this approach PDs are mapped to ratings where the distribution of the ratings is used as benchmark and making no validity or distribution assumptions of the model PDs. This is an attractive method when policies or business practise can relate to the dispersion of the rating distribution. However, in the case of financial institutions involved with portfolio management, distortions are created in the decision making process as the institutions attempt to manage against available opportunities in the market. This approach is, however, less sensitive to credit cycle changes as with the mapping through historical default rates approach.

The third approach in calibrating to ratings is the mapping to rating class with closest average PD. This approach again takes the PD as a given and examines the ratings PD attributes. This approach starts off by calculating PDs for all rated entities after which a central tendency is calculated for each rating class. If this quantity is known this approach proceed in following a similar process than the mapping through historical default rates approach, however measures of PD central tendency are used instead of historical default rates. Calculating the central tendency over longer periods will make the mapping more sensitive to credit cycle changes. When calculating the central tendency it is generally preferred to use measure less sensitive to outliers and large differences in PD such as the median, however the geometric mean is also feasible (Bohn & Stein, 2009, pp. 230-233).

Bohn and Stein (2009) presented a type of calibration technique whereby a dynamic measure such as the equity based PD is bucketed. Sometimes a continuous metric such as PD spanning a large range of values needs to be converted to smaller, discrete buckets. In this instance, buckets should be spread evenly with as little concentration as possible. A considerable concern raised by Bohn and Stein (2009) was assigning a homogeneous group for each bucket. Using a linear scale to produce these buckets is not efficient as a PD can range from one to 1 000 basis points. Bohn and Stein (2009) address these concerns with an approach by mapping to a rating class with the closest average PD. As a guide, this approach makes use of agency ratings or bank ratings. This approach is also appropriate if one wants to maintain comparability to another rating system. In another approach to address the concerns Bohn and Stein

(2009) present a convenient standard for the breakpoints in each grade. These are generally taken as historical default rates of agency rating categories mapped through historical default rates, however this does not have to be the case. A function is used allowing the lower grades to build gradually (Bohn & Stein, 2009, pp. 335-336).

Bohn and Stein (2009) emphasise that although the power of a model measure different things compared to calibration, power and calibration are related. Using calibration curves (model score versus the PD) the power of a model was found to be a limiting factor on how a high resolution may be reached through calibration. Four calibration curves were presented, each with a different power. Bohn and Stein (2009) indicated that the more powerful a model is, the greater the resolution it can achieve though probability estimation. Although there is a mathematical relationship between discriminatory power and probability of default estimates, a model may be calibrated well, but can still not be powerful (Bohn & Stein, 2009, pp. 389 - 391).

Due to the relationship between the discriminatory power of a model and the probabilities of default Bohn and Stein (2009) state that a weaker model will generally have a flatter calibration curve than a more powerful model. This results in the weaker model systematically assigning higher (lower) probabilities of default for saver (riskier) segments resulting in higher (lower) regulatory capital requirements for that segment. It is therefore possible to have two acceptable, unbiased models, each comprising different power characteristics resulting in different regulatory or economic capital requirements due to the difference in probability estimates. Bohn and Stein (2009) highlight the importance of calibration for some institutions. If an institution's risk characteristics are remarkably different than the norm due its lending practices, then based on the riskiness of the institution's book profile and lending practices, the institution may prefer either the powerful or weaker model for capital purposes (Bohn & Stein, 2009, pp. 394-395).

Ingolfsson and Elvarsson (2010) proposed a practical variable scalar methodology to transform PIT PDs to TTC PDs (Ingolfsson & Elvarsson, 2010, p. 374). Firstly, the PIT rating model used was a logistic regression model based on retail data. Secondly, a credit cycle model was estimated by using the Kalman filter. The cycle model which uses write-off data is a structural decomposition of the credit cycle. To obtain an appropriate representation of PD, the estimated cycle including a LGD component was regressed on an arrears data series as the PD

series was considered too short. The calibrated variable scalar for PD was obtained from the loss cycle by firstly formulating the average cross-section PD of the portfolio, secondly the first-order Taylor expansion relationship between loss-based credit cycle and PIT PD was estimated using linear regression. Lastly, from this estimate, the recalibrated distance between the PIT estimate and the long-term average PD (at any given point in time) the average cross-section PD is derived.

The remaining task from the methodology would then be to apply the calibrated scaling factor to the PIT model PD for the time period of interest to a TTC PD (Ingolfsson & Elvarsson, 2010, pp. 375-377). The variable scalar approach addresses the need to obtain TTC estimates for regulatory compliance and for long-term strategic planning. Implementing the variable scalar approach will differ between banks which depends on business objectives and historical data availability, however this approach might give interest even if just comparing with own solutions (Ingolfsson & Elvarsson, 2010, p. 379). This paper's aim is to use the Bayesian approach for calibration. It is, however, important to understand the origin of Bayes theorem. The section that follows derives Bayes theorem and presents it in an explicit form.

4.3.2.1 Bayesian Theorem

Let Ω denote a certain event or sample space, let *A* indicate that the event *A* occurs, let A^c denote that the event *A* does not occur, let *B* indicate that the event *B* occurs, let $A \cup B$ denote that either or both of A and B occur and let $A \cap B$ denote that both events *A* and *B* occur. Figure 4.2 presents the notation by the use of Venn diagrams as follows:

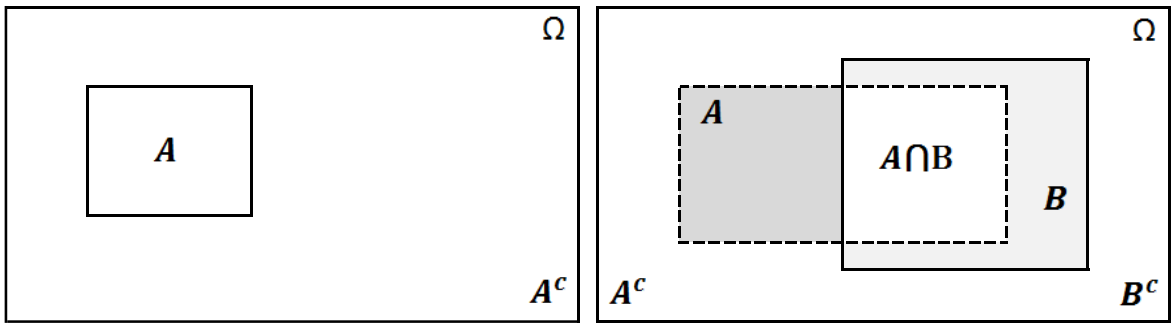


Figure 4.2: Venn diagram.

Let $P(A)$ denote the probability of event A with $0 \leq p \leq 1$ (Stirzaker, 1999, p. 32). From the left hand side of Figure 4.2 we have $1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$ (Stirzaker,

1999, p. 44). Using the right hand side of Figure 4.2 the $P(A)$ can also be expressed as follows in (2):

$$P(A) = P(A \cap B) + P(A \cap B^c) \tag{2}$$

Let $P(A|B)$ denote the conditional probability of event A given that event *B* occurs (Stirzaker, 1999, p. 49). From the right hand side of Figure 4.2 given that we know that $P(A \cap B)$ must lie in *B* we have:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{3}$$

$$\therefore P(A \cap B) = P(A|B)P(B) \tag{4}$$

Similarly, for $P(A \cap B^c)$ we have:

$$P(A|B^c) = \frac{P(A \cap B^c)}{P(B^c)} \tag{5}$$

$$\therefore P(A \cap B^c) = P(A|B^c)P(B^c) \tag{6}$$

(4) can also be written as:

$$\therefore P(A \cap B) = P(B \cap A) = P(B|A)P(A) \tag{7}$$

Substituting (7) into (3) we have the simple form of Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{8}$$

Substituting (4) and (6) into (2) we have:

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

$$\therefore P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

$$\therefore P(B) = P(B|A)P(A) + P(B|A^c)P(A^c) \tag{9}$$

Substituting (9) into (8) we obtain an explicit form of Bayes theorem presented by (10) as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \tag{10}$$

(Stirzaker, 1999, p. 57).

The explicit form of Bayes theorem expressed in (10) can be used for calibration purposes. The process for calibration using (10) to adjust current realised bad rates per scorecard rating model to expected future bad rates will be described and illustrated later in this paper.

4.4 Data and methodology

4.4.1 Data

South African economic data together with retail banking data were used in this research study based on a retail bank in South-Africa. External credit bureau data were used where applicable and based on credit bureau data in South-Africa specific. Table 4.1 presents the frequency and source data used in more detail.

Table 4.1: Data.

Topic	Data requirement	Source
Identifying downturn periods	Economic data	Retail bank data
Bureau scorecard	Bureau scorecard data in downturn and upturn	Bureau data Retail bank data
Calibration	Bureau scorecard data	Bureau data Retail bank data

Data availability for Table 4.1 includes data from January 2002 up to and including December 2014.

4.4.2 Methodology

In this section, a methodology is presented to investigate the procyclicality effect on a bureau scorecard and using the Bayesian approach for calibration purposes to adjust to future expected default rates for acquisition purposes. The methodology is presented in Figure 4.3 as follows:

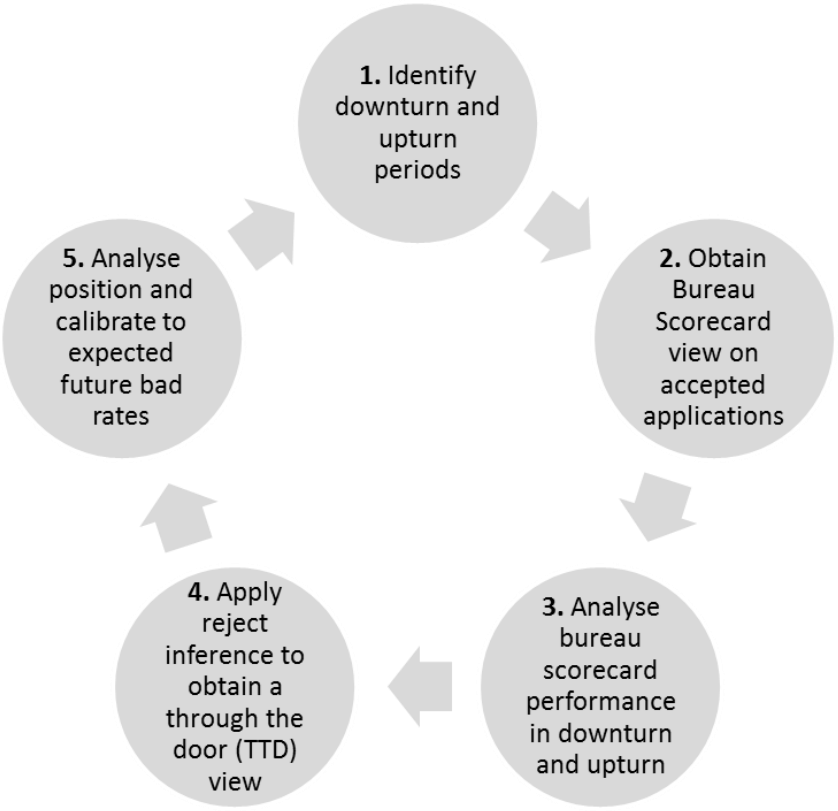


Figure 4.3: Methodology.

4.4.2.1 Identifying downturn and upturn periods

Using the guidelines of the BIS the following scenarios were investigated to identify downturn periods:

- periods of negative GDP growth,
- periods of elevated bad rates observed from historical bad rates, and
- periods where recovery rates are distressed

4.4.2.2 Bureau scorecard

Using historical bureau information, the performance of the bureau scorecard can be obtained for the accepted applications. This illustrates how effectively the scorecard distinguishes between good and bad clients and measures such as score trend performance can be used to observe this.

4.4.2.3 Analyse bureau scorecard performance in downturn and upturn

Acquisition scorecards aim to distinguish between good and bad clients which relates to the propensity of a client to be bad in the future. The first question to ask is whether the scorecard has a monotonic "bad rate" shape across the scores with the low scores giving high bad rates

and the high scores giving lower bad rates. However, the level of bad rate could differ when comparing the scorecard performance in a downturn period to the scorecard performance in an upturn period. The aim is to analyse this step more closely.

4.4.2.4 Reject inference

Scorecards used for acquisition purposes should be based on the through the door (TTD) applications i.e. all applications. If only the accepted applications are used and the rejected application ignored there would be bias to the accepted applications from the past (Anderson, 2007, p. 65). The aim in this phase of the methodology is to infer the rejected applications from step 2 in Figure 4.3 to obtain a TTD application view.

4.4.2.5 Calibration

Based on the findings from step 3 if differences are observed for scorecard performance it is necessary to calibrate current realised bad rates depending on risk appetite. In this step the Bayes approach would be followed to adjust current observed bad rates to various scenarios of future expected bad rates and analysing the effect of each scenario.

4.5 Results

4.5.1 Data

In this section, the methodology presented in Section 4.4 is applied to a South African retail bank to obtain an understanding of bureau scorecard performance in down and upturn periods. Analysing the performance in the down and upturn periods calibration is applied using the Bayes approach to various scenarios of expected future bad rates.

4.5.2 Downturn and upturn periods

Figure 4.4 presents the GDP growth and bad rate percentages based on a 12-month outcome (A longer outcome can be used based on the definition used for acquisition purposes, e.g. a 24-month outcome). The bad rate percentage reaches a peak in June 2008 at the point where

the GDP growth decreases significantly up to June 2008. Also the bad rate percentage starts stabilising from January 2010.

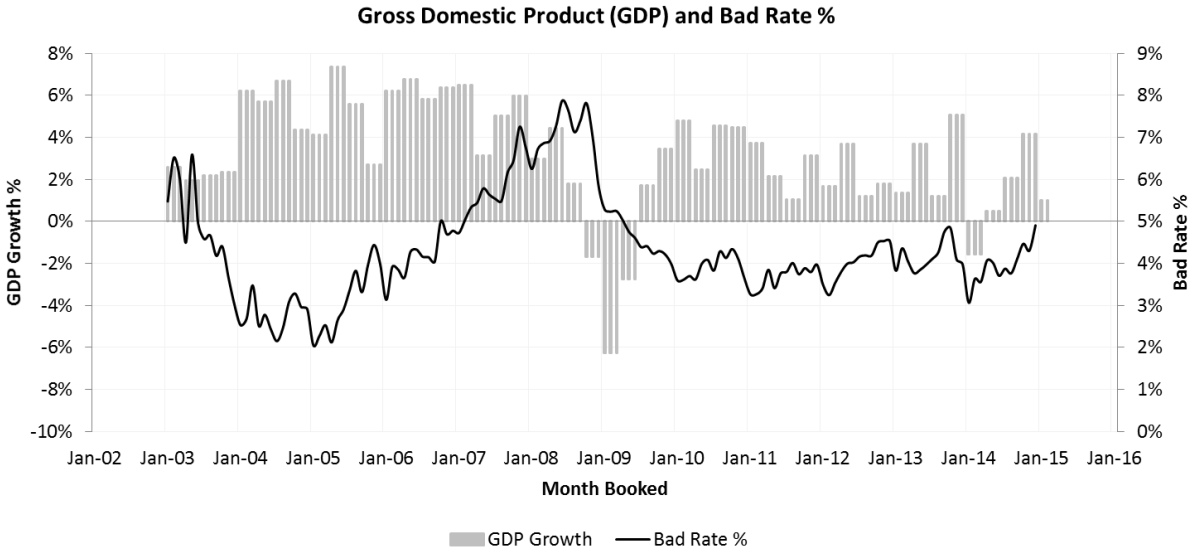


Figure 4.4: GDP growth and bad rate %.

Figure 4.5 presents the GDP growth and new bad inflow percentages. The new bad inflows illustrate the percentage of new bads realised which can be used as a measure for elevated bad rates.

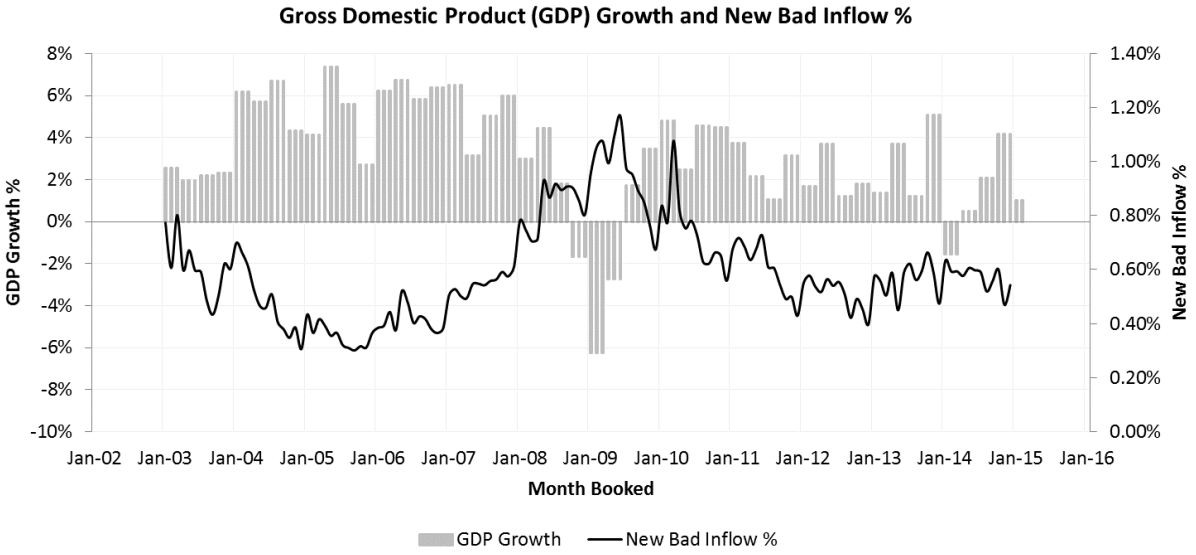


Figure 4.5: GDP growth and new bad inflow %.

In Figure 4.5 the GDP growth is at its lowest from October 2008 up to and including June 2009. The new bad inflow percentage is elevated at its highest from May 2008 up to and including

September 2009. The GDP Growth also stabilise from January 2010. Figure 4.6 presents GDP growth plotted against the recovery percentage.

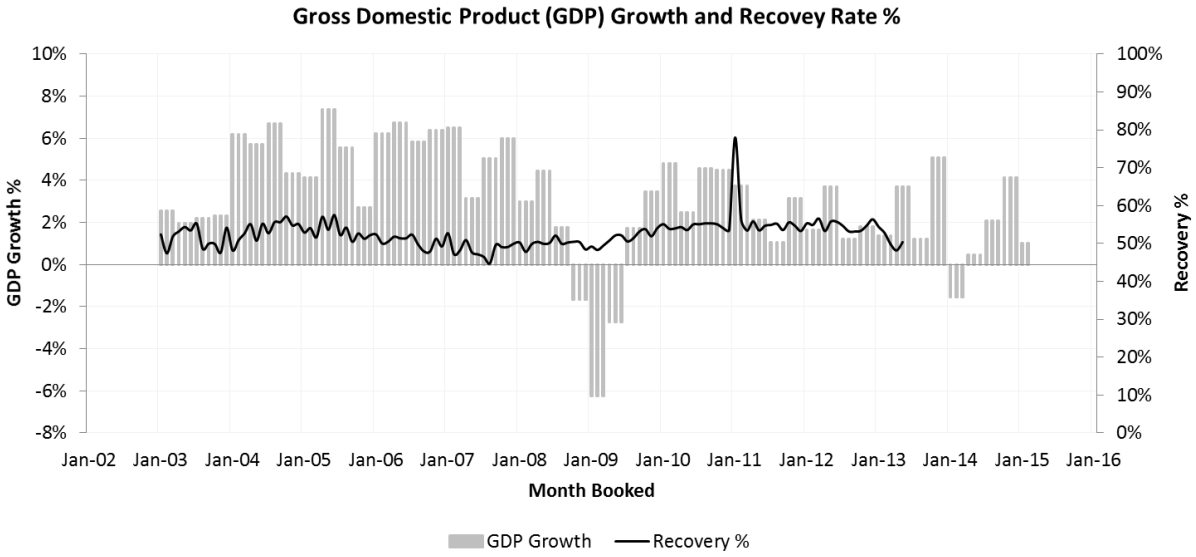


Figure 4.6: GDP growth and recovery %.

In Figure 4.6 the recovery percentage for the period October 2008 up to and including September 2009 is also low when comparing to periods after September 2009 and also stabilising around the period of January 2010. From Figure 4.4, Figure 4.5 and Figure 4.6 the downturn and upturn periods chosen for step 3 in the methodology presented in Figure 4.3 are presented in Table 4.2.

Table 4.2: Downturn and upturn periods.

	Start date	End date	Reason
Downturn period	October 2008	September 2009	Negative GDP Elevated defaults
Upturn period	January 2010	December 2010	Positive GDP Stable default rates Stable recovery rates
Calibration model build and model test period	January 2010	December 2010	Use upturn period and calibrate to different scenarios
Calibration Scenarios **(Out of Sample Model build Scenarios)**	January 2011	December 2011	Test calibration results after calibration period

4.5.3 Analyse bureau scorecard performance in downturn and upturn

This section presents the performance of the bureau scorecard on accepted applications in the down and upturn periods. The bureau scorecard is analysed on the downturn period and upturn period from Table 4.2.

With the score trend performance, the bureau scores from the downturn period were grouped into 10 score intervals such that the score intervals is equally distributed as far as possible. The same score intervals were used for the upturn period to have a like-for-like comparison. Figure 4.7 presents the score trend performance for the bureau scorecard for both the downturn and upturn period.

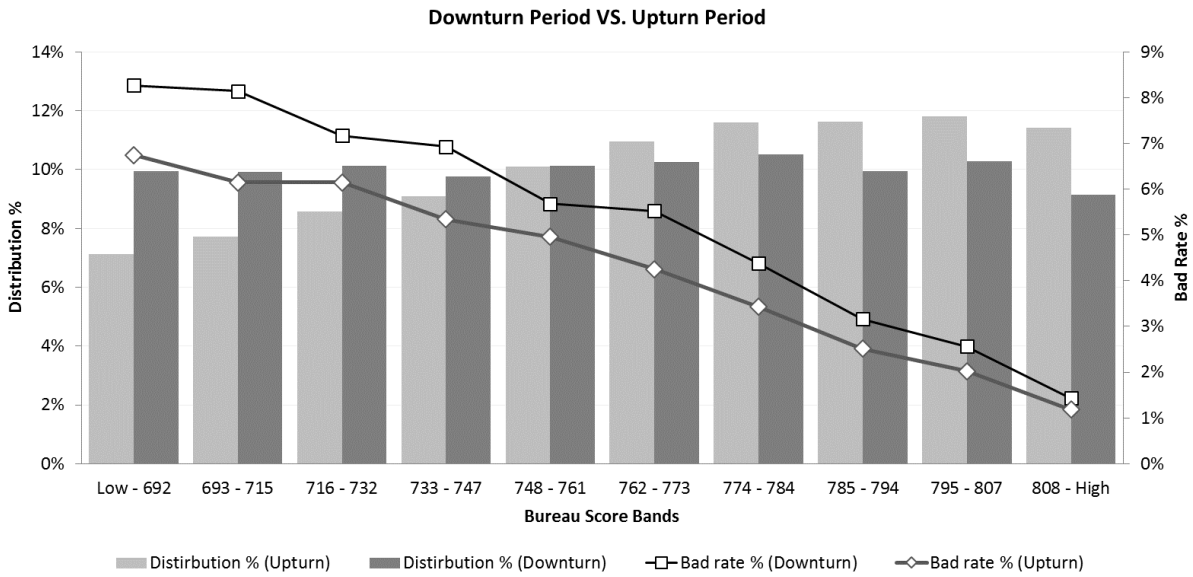


Figure 4.7: Score Trend Performance.

The most important step to observe whether a scorecard separate good clients from bad clients is to analyse the monotonicity of the bad rates across the score intervals as presented in Figure 4.7. In Figure 4.7, the bad rates have a monotonic shape for both the downturn and upturn period, i.e. high bad rates for the lower scores monotonically decreasing to the high scores with the lowest bad rates. Given that the bureau scorecard separates the good clients from the bad clients the level of bad rate across all the score intervals are on a higher level for the downturn period than the upturn period. In Figure 4.7, the population shift is also evident between the downturn and upturn period. This necessitates that given a certain point in the economy that an adjustment or calibration should take place dependent on credit risk appetite.

The significance of the paper is to build a calibration model that can be used to calibrate the bureau scorecard to expected future bad rates for acquisition purposes. The order of flow for the sections that follows is firstly to obtain a TTD application view of bad rates for the upturn period, secondly to build and test the calibration model on upturn period data and thirdly to present scenarios of expected future bad rates that are comparable to actual bad rates observed.

4.5.4 Reject inference

Using the performance of the accepted applications from the upturn period on the bureau scorecard the rejected applications are inferred. Figure 4.8 presents the log-odds to score relationship for the accepted applications from the upturn period.

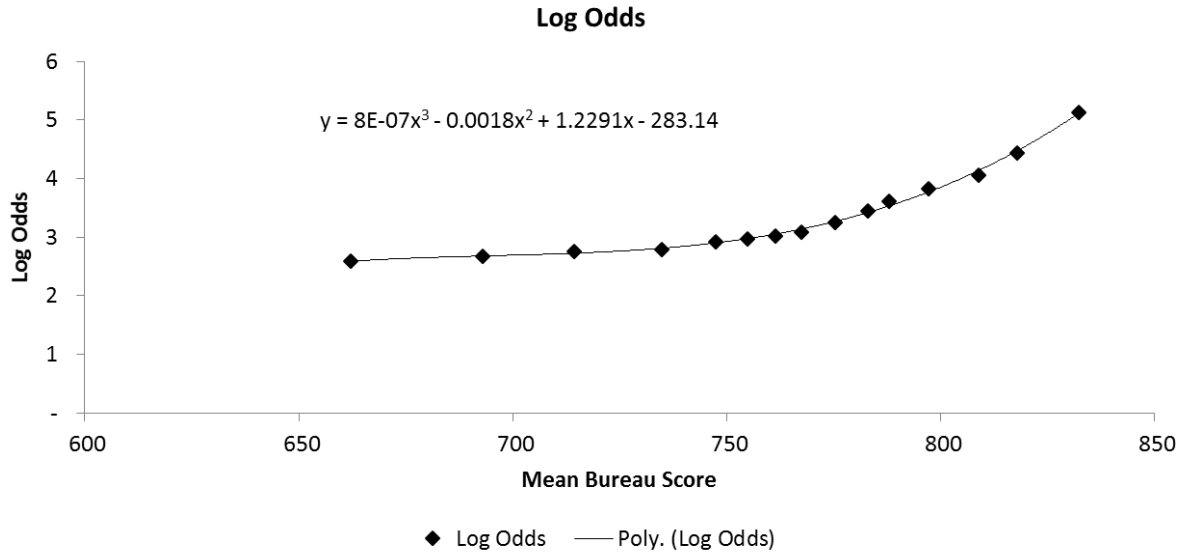


Figure 4.8: Log-Odds (Upturn Period).

Using the polynomial fit as presented in Figure 4.8 the rejected applications from the upturn period are inferred. The Gini-coefficient (a measure on how a scorecard separates good clients from bad clients) is shown for the bureau scorecard for the upturn period in Table 4.3 for the TTD population, i.e. all applications which are accepted and rejected.

Table 4.3: Gini-coefficient (All applications).

	Bad Rate (Accepts)	Bad Rate (Rejects)	Gini-coefficient
Upturn period	4.00%	9.09%	47.8%

4.5.5 Calibration

4.5.5.1 Calibration model

The aim of this section is to present the steps taken in building the calibration model to be used for acquisition purposes. Figure 4.9 presents the calibration model:

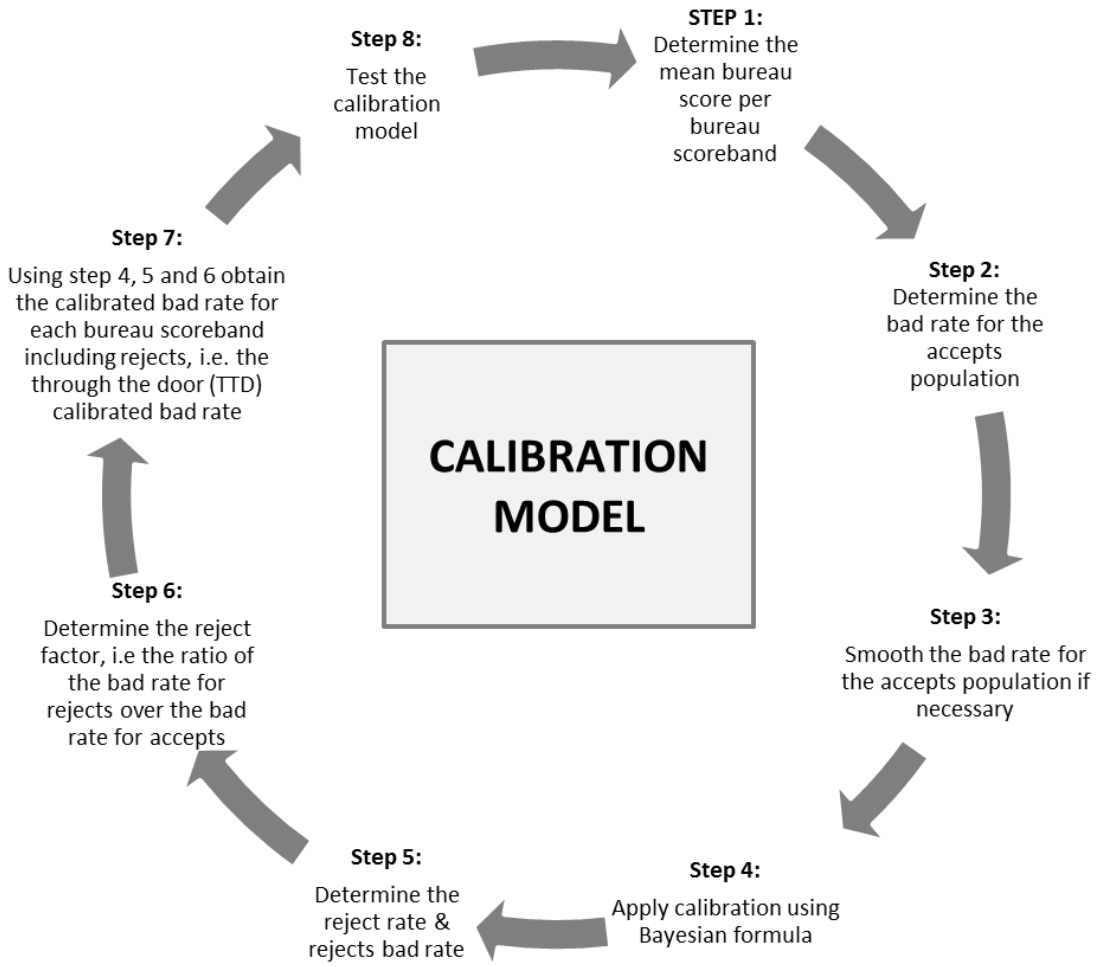


Figure 4.9: Calibration Model.

4.5.5.1.1 Calibration Model steps 1 - 3: Smoothing of bad rate

The first three steps for the calibration model is to obtain a smooth bad rate sloping for the accepted applications. This is done using a polynomial fit between the average bureau score and log of the bad rate across the bureau score intervals. Using the same score intervals as determined in the downturn period in Section 4.5.3 the first step is to obtain the average bureau score for each bureau score interval, secondly the actual bad rates needs to be obtained and lastly a function is fitted to obtain a smooth bad rate across the bureau score intervals. Figure 4.10 and Table 4.4 presents the smoothed bad rate for the upturn period against the average bureau scores for each bureau score interval.

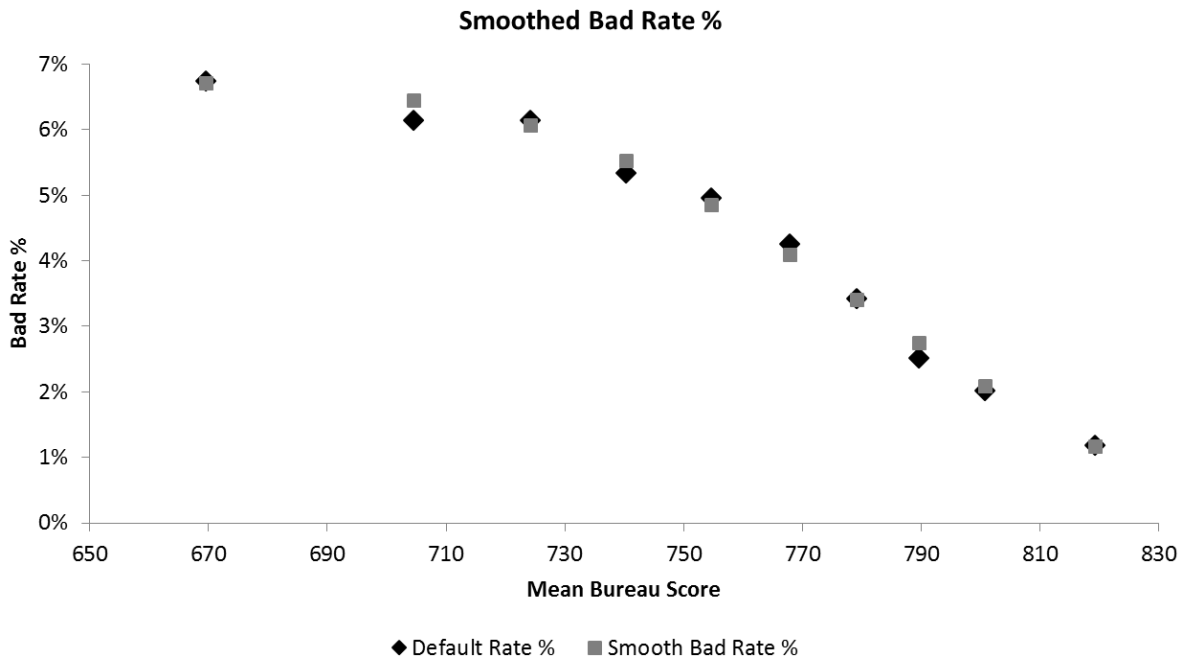


Figure 4.10: Smoothed Bad Rate % (Upturn Period).

Table 4.4: Upturn Period.

	Actual Bad Rate	Smoothed Bad Rate
Low - 692	6.74%	6.72%
693 - 715	6.14%	6.45%
716 - 732	6.15%	6.07%
733 - 747	5.34%	5.53%
748 - 761	4.95%	4.86%
762 - 773	4.25%	4.10%
774 - 784	3.43%	3.41%
785 - 794	2.51%	2.75%
795 - 807	2.02%	2.09%
808 - High	1.18%	1.17%

4.5.5.1.2 Calibration Model step 4: Calibration

In Section 4.3.2.1 the origin of the Bayesian theorem was presented. Using the explicit form of the Bayesian theorem presented by (10) in Section 4.3.2.1 the smoothed bad rate per bureau score interval in Table 4.4 can be calibrated. It is important to define concepts in relation to the explicit form of the Bayesian formula in this context: let A represent bads observed within 12 months in bureau score interval i, let B represent an adjustment made on the population, let A/B represent bads observed within 12 months in bureau score interval i given that an adjustment B has been made and let B/A represent an adjustment B has been made given that bads within 12 months in bureau score interval i has been observed. In this context (10) in Section 4.3.2.1 can be written as:

$$ABR_i = \frac{\left(\frac{Adjustment}{BR_{Upturn}}\right) \times SBR_i}{\left(\frac{Adjustment}{BR_{Upturn}}\right) \times SBR_i + \left[(1 - SBR_i) \times \frac{1 - Adjustment}{1 - BR_{Upturn}}\right]} \quad (11)$$

where:

ABR_i represents the adjusted bad rate % of bureau score bucket i, SBR_i represents the smooth bad rate % of bureau score bucket i, $Adjustment$ represents the input to solve for an overall level of bad rate and BR_{Upturn} represents the smoothed bad rate % for the upturn period. In (11) one can interpret as follows: $P(A) = SBR_i$, $P(B|A) = \left(\frac{Adjustment}{BR_{Upturn}}\right) =$ ratio between adjustment and overall bad rate observed for upturn period (i.e. overall adjustment level). ABR_i is the expected bad rate adjusted per bureau score interval i. The $Adjustment$ in (11) can then be iteratively changed such that the overall bad rate across all bureau score intervals achieves the desired expected bad rate.

For the calibration model the upturn data will be used and each bureau score interval will be calibrated such that the overall calibrated bad rate is the same level as the upturn period for testing purposes. These calibration results will then be tested against actual bad rates for the upturn period in step 8 after rejected applications have been added to account for the TTD applications. Using the smoothed bad rate for each bureau score interval and (11) the calibrated bad rate for each bureau score interval is presented in Table 4.5.

Table 4.5: Calibrated bad rate %.

	Actual Bad Rate	Smoothed Bad Rate	Calibrated Bad Rate %
Low - 692	6.74%	6.72%	6.60%
693 - 715	6.14%	6.45%	6.33%
716 - 732	6.15%	6.07%	5.96%
733 - 747	5.34%	5.53%	5.43%
748 - 761	4.95%	4.86%	4.77%
762 - 773	4.25%	4.10%	4.02%
774 - 784	3.43%	3.41%	3.34%
785 - 794	2.51%	2.75%	2.69%
795 - 807	2.02%	2.09%	2.05%
808 - High	1.18%	1.17%	1.14%
Total	**4.00%**	**4.45%**	**3.96%**

4.5.5.1.3 Calibration Model steps 5-7: Include rejected applications

Scorecards used for acquisition purposes should include rejected applications to avoid bias to historical accepts. In Section 4.5.4 the rejected applications from the upturn period were inferred. Using these results the reject rate, rejects bad rate and reject factor for each bureau score interval can be determined. The reject rate represents the percentage applications rejected and the reject factor represents the ratio of the rejects bad rate over the accepts bad rate. (12) presents the reject rate formula.

$$RR_i = \frac{RA_i}{A_i} \tag{12}$$

Where:

RR_i represents the reject rate for bureau score interval i, RA_i represents the number of rejected applications for bureau score interval i and A_i represents the number of applications in bureau score interval i. (13) presents the reject factor formula.

$$RF_i = \frac{BR_i}{BA_i} \tag{13}$$

Where:

RF_i is the reject factor for bureau score interval i, BR_i represents the bad rate of rejected applications for bureau score interval i and BA_i represents the bad rate of accepted applications in bureau score interval i. Table 4.6 presents the reject rate for each bureau score interval for the upturn period.

Table 4.6: Reject rate.

	Accepted Applications	**Rejected Applications**	**Reject Rate**
Low - 692	9 062	387 206	97.71%
693 - 715	9 814	73 782	88.26%
716 - 732	10 885	47 259	81.28%
733 - 747	11 560	35 301	75.33%
748 - 761	12 816	34 102	72.68%
762 - 773	13 900	34 261	71.14%
774 - 784	14 737	36 762	71.38%
785 - 794	14 768	33 916	69.67%
795 - 807	15 005	35 900	70.52%
808 - High	14 496	38 311	72.55%
Total	127 043	756 800	85.63%

Table 4.7 presents the reject factor for each bureau score interval for the upturn period.

Table 4.7: Reject factor.

	Accepts Bad Rate	Rejects Bad Rate	Reject Factor	Final Reject Factor
Low - 692	6.74%	13.68%	2.03	2.03
693 - 715	6.14%	6.29%	1.02	1.02
716 - 732	6.15%	5.95%	0.97	1.00
733 - 747	5.34%	5.49%	1.03	1.03
748 - 761	4.95%	4.86%	0.98	1.00
762 - 773	4.25%	4.13%	0.97	1.00
774 - 784	3.43%	3.42%	1.00	1.00
785 - 794	2.51%	2.74%	1.09	1.09
795 - 807	2.02%	2.03%	1.01	1.01
808 - High	1.18%	1.11%	0.94	1.00
Total	**4.00%**	**9.09%**	**2.27**	**2.27**

In Table 4.7 for some of the bureau score intervals the rejects bad rate is lower than the accepts bad rate resulting in a reject factor below 1. These reject factors below 1 will be capped at 1 as we expect that the rejects bad rate will be higher than accepted clients. The final calibrated bad rate for each bureau score interval for the TTD population is obtained using (13):

$$FCBR_i = [ABR_i \times (1 - RR_i)] + [ABR_i \times RR_i \times RF_i] \quad (13)$$

Where:

$FCBR_i$ represents the final TTD calibrated bad rate for bureau score interval i.

(13) comprises two components: the calibrated bad rate for accepts and the expected calibrated bad rate for rejected applications. Table 4.8 present the TTD calibrated bad rates for each bureau score interval.

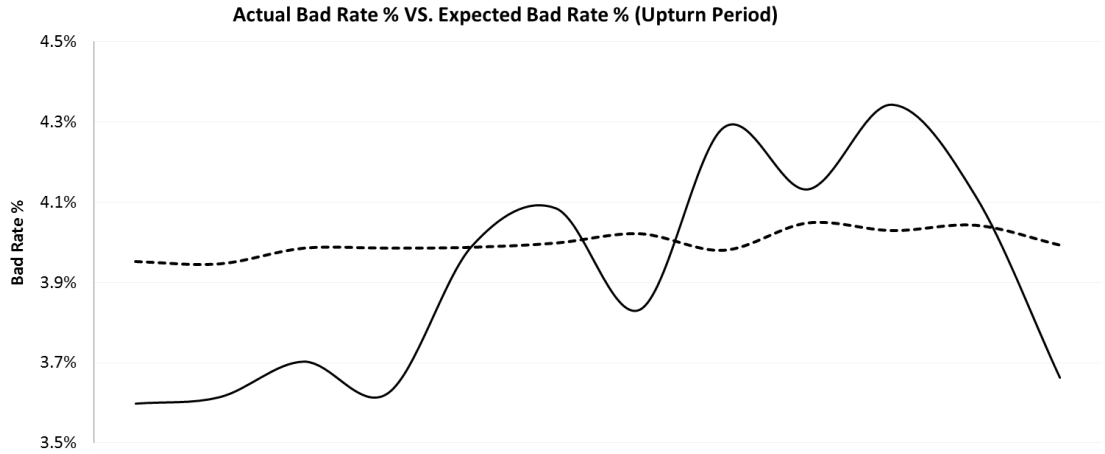
Table 4.8: Final calibrated bad rate.

	Accepts Smoothed Bad Rate	Rejects Bad Rate	Reject Rate	Final Reject Factor	Final Calibrated Bad Rate
Low - 692	6.72%	13.68%	97.71%	2.03	13.23%
693 - 715	6.45%	6.29%	88.26%	1.02	6.47%
716 - 732	6.07%	5.95%	81.28%	1.00	5.96%
733 - 747	5.53%	5.49%	75.33%	1.03	5.55%
748 - 761	4.86%	4.86%	72.68%	1.00	4.77%
762 - 773	4.10%	4.13%	71.14%	1.00	4.02%
774 - 784	3.41%	3.42%	71.38%	1.00	3.34%
785 - 794	2.75%	2.74%	69.67%	1.09	2.86%
795 - 807	2.09%	2.03%	70.52%	1.01	2.06%
808 - High	1.17%	1.11%	72.55%	1.00	1.14%
Total	**4.45%**	**9.09%**	**85.63%**	**2.27**	**8.28%**

In Table 4.8, the calibrated bad rate for each bureau score interval is presented which can be used on the TTD applications. It should be noted that the lowest bureau scores in bucket “Low – 692” have a reject rate of 97.71% which means that 2.29% of applications are booked in this bureau score interval. These 2.29% of applications that are accepted are known as 'low end override' accounts where the expected bad rate should be lower that the TTD calibrated bad rate of 13.23% [as presented in Table 4.8]. In this case to test the calibration model the low end bucket containing the majority of low end overrides will be allocated the expected calibrated bad rate of the accepted clients from Table 4.5.

4.5.5.1.4 Calibration Model step 8: Test the calibration model

Using the reject rate for each bureau score interval from Table 4.6 and the final calibrated bad rate for each bureau score interval from Table 4.8, an overall expected bad rate for accepted applications for the upturn period using the TTD applications can be determined. The overall expected bad rate for accepted applications determined using the calibration model from the TTD applications is plotted against actual bad rates for accepted applications in Figure 4.11.



	Jan-10	Feb-10	Mar-10	Apr-10	May-10	Jun-10	Jul-10	Aug-10	Sep-10	Oct-10	Nov-10	Dec-10
Expected Bad Rate	3.95%	3.95%	3.99%	3.99%	3.99%	4.00%	4.02%	3.98%	4.05%	4.03%	4.04%	3.99%
Actual Bad Rate	3.60%	3.61%	3.70%	3.62%	3.99%	4.08%	3.83%	4.29%	4.13%	4.34%	4.11%	3.66%

Figure 4.11: Expected bad rate vs. actual bad rate (upturn period).

The Student's t-test can be used to determine whether the two populations are likely to have come from the same two underlying populations that have the same mean of bad rate. Table 4.9 presents the hypothesis for this test:

Table 4.9: Hypothesis.

Hypothesis	The difference between the actual bad rate and expected bad rate = 0

Performing the Student's t-test between the expected bad rate from the calibration model and the actual bad rate from Figure 4.11 the p-value obtained is illustrated in Table 4.10.

Table 4.10: Student's t-test.

Hypothesis	The difference between the actual bad rate and expected bad rate = 0
p**-value**	0.29
Conclusion	Since 0.29 > 0.05, the hypothesis not rejected

In Table 4.10 since the hypothesis is not rejected it can be concluded that the difference between the expected bad rate from the calibration model and the actual bad rates = 0.

4.5.5.2 Calibration Scenarios

In this section the calibration model as described in Section 4.5.5.1 is used to calibrate to different scenarios. Suppose the bank is at the end of 2010 and wants to obtain an outlook of expected bad rates for the year 2011. The three scenarios are presented in Table 4.11.

Table 4.11: Calibration scenarios.

	Scenario	Reason	Bad Rate Level
1	Calibrate the bureau scorecard to the same level as the bad rates experienced in 2010	It is felt that the economy for the foreseeable future will remain stable	4%
2	Calibrate the bureau scorecard to the worst overall bad rate experience in 2008	It is felt that the economy for the foreseeable future will deteriorate to the level of 2008	8%
3	Calibrate the bureau scorecard to the lowest overall bad rate experience in 2005	It is felt that the economy for the foreseeable future will improve to the level of 2005	2%

The application distribution percentage per bureau score interval is presented in Figure 4.12.

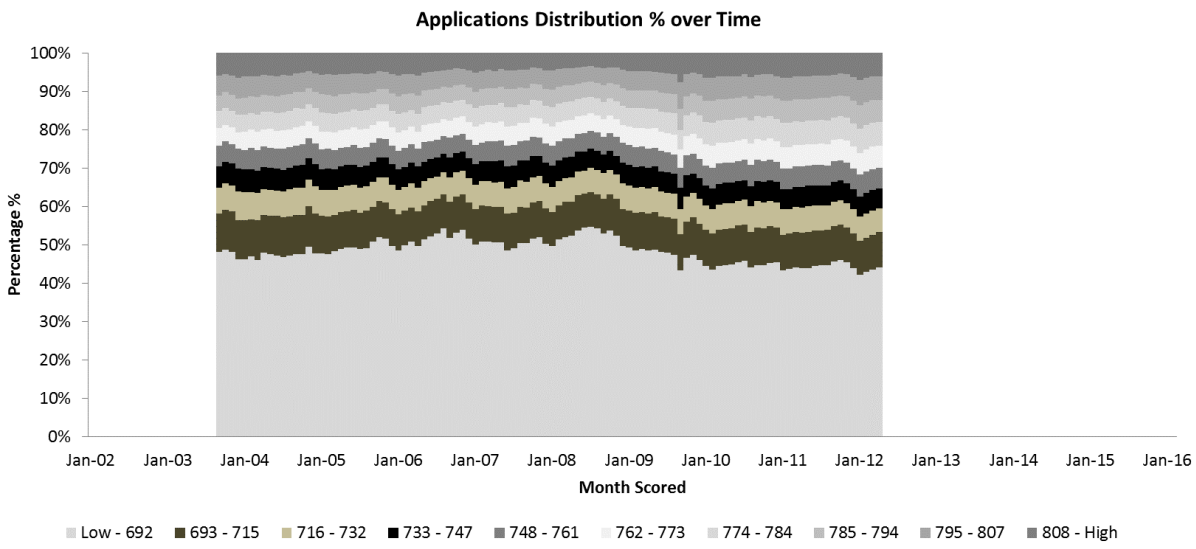


Figure 4.12: Application distribution % per bureau score interval.

By using the TTD applications as presented in Figure 4.12, the reject rate for each bureau score interval from Table 4.6 and the final calibrated bad rate for each bureau score interval from Table 4.8, Figure 4.13 presents the expected bad rates against the actual bad rates for the three scenarios given in Table 4.11.

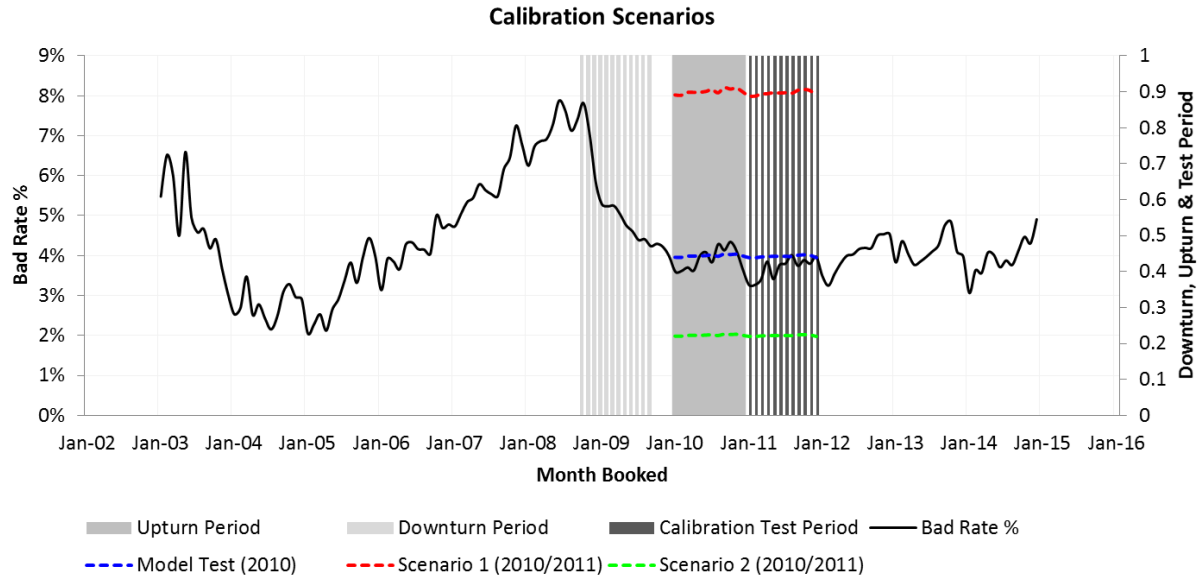


Figure 4.13: Calibrated scenarios.

The scenarios presented above for the year 2011 represent an out of sample test for the calibration model. Scenario 1 is closely related to the actual bad rates experienced with scenario 2 and 3 presenting the level of bad rates should the retail bank feel it is necessary to calibrate to such bad rate levels experienced in the past. As an example, if the retail bank felt (at the

end of 2010) that scenario 3 could be expected for 2011 the expected future bad rates should reflect lower bad rates if the cutoff strategy for accepted applications remains unchanged. In this scenario, if the retail bank wanted to maintain the same level of bad rate from 2010 the cutoffs for accepted clients can be reviewed such that applications can be accepted to maintain the 2010 bad rate level.

4.6 Conclusion

Application scorecards play an essential and critical role to determine the credit worthiness of applicants for acquisition purposes. Application scorecards separate good applicants from bad applicants and the applicants that are eventually accepted drives the quality of the retail banks book for the future. A problem that exist for any retail bank is the overall state of the economy in the business cycle which influence the level of bad rate for applicants. This in effect have an impact on application scorecards.

In this paper, a literature background was provided on scenarios affecting application scorecards. Methods for identifying downturn periods were also presented and a background on calibration was given with more emphasis put on the Bayesian theorem that can be used to adjust bad rates.

This work indicated that although the bureau scorecard is monotonic across the bureau scores during both the downturn and upturn period that the level of bad rate is higher for the downturn period. The contribution from of this paper is a methodology and calibration model that can be used to adjust or calibrate current experienced bad rates per score interval to expected future bad rate levels. Calibrating to expected future bad rates can result in lower future bad rates, lower future impairments and less future capital to hold against unexpected losses.

Data from a retail bank in South Africa were used together with bureau information from South Africa specific to build a calibration model. The model was built using data from the upturn period selected and was also tested by comparing actual bad rates against the bad rates produced from the calibration model. Using the methodology and calibration model three scenarios were carried out in an independent out of sample to perform comparisons between predicted expected bad rates and actual bad rates. These comparisons indicated that there is a significant relationship between actual bad rates and expected bad rates from the calibration model.

Possible future research includes other methods to determine downturn and upturn periods apart from those discussed in this paper and the age effect of an application scorecard during an economic cycle.

Other possible research areas include the procyclical effect on non-bureau application score-cards or on application scorecards using a longer outcome period.

Bibliography

Abdou, H. & Pointon, J., 2011. Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance & Management,* 18(2-3), pp. 59-88.

Anderson, R., 2007. *The credit scoring toolkit, theory and practise for retail credit risk. Management and decision automation..* United States: Oxford University Press Inc..

Basel Committee on Banking Supervision, 2004. *International Convergence of Capital Measurement and Capital Standards,* s.l.: Bank for International Settlements.

Basel Committee on Banking Supervision, 2005. *Guidance on Paragraph 468 of the Framework Document,* s.l.: Bank for International Settlements.

Basel Committee on Banking Supervision, 2006. *International Convergence of Capital Measurement and Capital Standards,* s.l.: Bank for International Settlements.

Basel Committee on Banking Supervision, 2006. *The IRB Use Test: Background and Implementation,* s.l.: Bank for International Settlements.

Beveridge, S. & Nelson, C. R., 1981. A New Approach to Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the Business Cycle. *Journal of Monetory Economics,* Volume 7, pp. 151-174.

Bohn, J. R. & Stein, R. M., 2009. *Active Credit Portfolio Management in Practice.* First ed. New Jersey: Wiley.

Bonini, S. & Caivano, G., 2014. Probability of Default: A Modern Calibration Approach. In: C. Perna & M. Sibillo, eds. *Mathematical and Statistical Methods for Actuarial Sciences.* Switzerland: Springer International Publishing, pp. 41 - 44.

Burns, A. F. & Mitchell, W. C., 1946. *Measuring Business Cycles,* s.l.: National Bureau of Economic Research.

Collins, M., 2016. *The Global Credit Cycle,* s.l.: Prudential Financial.

Glößner, D. P., 2003. *Calculating Basel II Risk Parameters for a Portfolio of Retail Loans,* Oxford: Thesis submitted in partial fulfilment of the requirements for the MSc in Mathematical Finance.

Gordy, M. B. & Howells, B., 2004. *Procyclicality in Basel II: Can We Treat the Disease Without Killing the Patient?.* [Online]
Available at: https://www.bis.org/

Ingolfsson, S. & Elvarsson, B., 2010. Cyclical adjustment of point-in-time PD. *Journal of the Operational Research Society,* Volume 61, pp. 374-380.

Iqbal, N. & Ali, A., 2012. Quoted in Bonini & Caivano, 2014. Probability of Default: A Modern Calibration Approach. In: C. Perna & M. Sibillo, eds. *Mathematical and Statistical Methods for Actuarial Sciences.* Switzerland: Springer International Publishing, pp. 41 - 44.

Kiefer, N. M., 2008. Quoted in Bonini & Caivano, 2014. Probability of Default: A Modern Calibration Approach. In: C. Perna & M. Sibillo, eds. *Mathematical and Statistical Methods for Actuarial Sciences.* Switzerland: Springer International Publishing, pp. 41 - 44.

Landau, J.-P., 2009. *Procyclicality - what it means and what could be done,* s.l.: Bank for International Settlements.

Lim, M. K. & Sohn, S. Y., 2007. Cluster-based dynamic scoring model. *Expert Systems with Applications,* Volume 32, pp. 427-431.

Medema, L., Koning, R. H. & Lensink, R., 2009. A practical approach to validating a PD model. *Journal of Banking & Finance,* Volume 33, pp. 701-708.

Miu, P. & Ozdemir, B., 2006. Basel Requirements of Downturn Loss Given Default: Modelling and Estimating Probability of Default and Loss Given Default Correlations. *Journal of Credit Risk,* 2(2), pp. 43-68.

Nelson, C. R. & Plosser, C. I., 1982. Trends and Random Walks in Macroeconomic Time Series. *Journal of Monetary Economics,* Volume 10, pp. 139-162.

Pluto, K. & Tasche, D., 2005. Quoted in Bonini & Caivano, 2014. Probability of Default: A Modern Calibration Approach. In: C. Perna & M. Sibillo, eds. *Mathematical and Statistical Methods for Actuarial Sciences.* Switzerland: Springer International Publishing, pp. 41 - 44.

Siddiqi, N., 2006. *Credit risk scorecards, developing and implementing intelligent credit scoring.* United States of America: John Wiley & Sons, Inc..

Stirzaker, D., 1999. *Probability and Random Variables - a beginners guide.* 1st ed. Cambridge: Cambridge University Press.

Tasche, D., 2013. Quoted in Bonini & Caivano, 2014. Probability of Default: A Modern Calibration Approach. In: C. Perna & M. Sibillo, eds. *Mathematical and Statistical Methods for Actuarial Sciences.* Switzerland: Springer International Publishing, pp. 41 - 44.

TransUnion, 2015. *Credit Bureau, Credit Reporting Companies. Learn about your credit..* [Online]
Available at: https//www/mytransunion.co.za
[Accessed 2 4 2015].

TransUnion & FICO, 2009. *Empirica. Minimise your credit risk. Increase your profitability.,* Johannesburg: TransUnion.

Turri, A. & Salis, F., 2010. *Downturn LGD: Flexible and Regulatory Consistent Approaches.* [Online]
Available at: http://www.greta.it/credit/credit2010/PAPERS/Posters/Turri_Salis.pdf
[Accessed 11 7 2016].

Chapter 5
Conclusions and recommendations

5.1 Summary and conclusions

Ever since the late 1950s, credit scoring has been widely adopted to guide credit decisions. Growing pressures for increased revenue, increased competition and a growing economy have led financial institutions to establish new effective ways to attract credit worthy clients. Credit scoring plays an important role ultimately affecting future impairments, capital requirements and profits, however literature in the world of credit scoring have been limited. In light of the limited research and the important role of credit scoring to guide credit decisions, the need to enhance and augment existing techniques and develop new techniques to improve credit risk measurement and management are paramount.

Within this thesis, three important areas of credit scoring were explored namely statistical techniques for modelling credit scorecard models, the monitoring of credit scorecards and the necessity of calibration. Three significant problems in each of these areas were addressed and solutions proposed.

5.1.1 Credit scoring Matrix Model (CSMM)

In Chapter 2, the problem of no optimal technique that exists to build a credit scoring model and the disagreement on an appropriate cut-off score was investigated. A CSMM was developed to optimise the separation between good and bad clients. The CSMM consists of an internal application scorecard (internal organisation information) component and the Empirica score (External bureau information) component. Firstly, the internal application model was developed by following a 12 step approach. In step one, the “bad”, “indeterminate” and “good” definition were stated on which the internal application scorecard will be built with the “bad” defined as including accounts that are ever three-plus in arrears with balance greater than ZAR100 or in legal department or written-off. In step 2, maturity analysis was performed to obtain a suitable outcome to build the internal application scorecard. It was decided to build the scorecard on a 24-month outcome based on two reasons; firstly to capture as much bad accounts as possible and secondly too much history is undesirable as applications at development could not represent the applications of today. In steps 3 and 4, the development windows were presented and the exclusions listed, respectively. Using a simple

random sampling technique in the statistical analysis software (SAS) a sample was created from which the internal application scorecard was built in step 5.

The bureau score at outcomes that provided a Gini-coefficient of 81% was used to infer rejected applications in step 6. Thirty-two borrower or transaction type characteristics were used under step 7. Business logic, together with the FICO MB7 software, was used to bucket characteristics under steps 8 and 9 and the "Scorecard Divergence" method in FICO MB7 was used to carry out the regression in step 10. Under step 11, scorecard tests were applied. The internal application scorecard (which excludes bureau related information) gave a Gini-coefficient of 30.5%, a divergence of 0.32 and a KS-statistic of 21.84 for the development sample, no correlation was observed between the characteristics, and the score trend performance indicated a monotonic decrease in bad rate from the low scores to the high scores (as expected). The PSI also indicated stability in the internal application scorecard.

Secondly using the same "bad", "indeterminate" and "good" definitions and 24-month outcome, the Empirica bureau scorecard gave a Gini-coefficient of 42% with the score trend performance indicating a monotonic decrease in bad rate from the low scores to the high scores. By using the internal application scorecard and the Empirica bureau scorecard the CSMM was constructed. The Gini-coefficient of the CSMM was calculated as 46% which gave a relative percentage uplift in Gini-coefficient of 12% from the Empirica score optimising the separation between good and bad clients. This will lead to fewer clients to be initially selected that would have gone bad affecting impairments, capital requirements and profits. With the CSMM, clients with a low bureau (Empirica) score, but high application score can be considered, clients with a low application score, but high bureau score (Empirica) can be considered, negative effects of application form manipulation for internal application scorecard are reduced because the bureau (Empirica) hedges this manipulation. By comparing a one-dimensional credit scorecard model to the CSMM, it was illustrated how the CSMM provides a greater dimension of granularity to determine the appropriate cut-off score.

5.1.2 Swap-set Gini-coefficient

The aim of monitoring is to observe whether a credit scoring model is working or not. The specific monitoring of an application scorecard accepts population is of special importance due to the effort risk mangers apply regarding risk strategies and setting cut-off decisions. In Chapter 3, the problem regarding the difference in statistical performance levels between

development and post implementation data was investigated. A methodology was presented to enable the comparison of statistical performance measures more effectively between the development sample and post implementation data for the accepts population. Comparing the statistical performance measures for the accept population between development and post implementation data for the bureau scorecard indicated significant differences. The Gini-coefficient at development was 36% with the post implementation data indicating a Gini-coefficient of 48%. Using the methodology presented, a new development dataset was created that takes the swap-set into account which is to be expected when the scorecard is implemented. The next step in the methodology was to compare the log-odds to score relationships for both the development data after swap-set and the post implementation data.

The log-odds–to-score relationships indicated the monotonicity of the scorecard in both the development data after the swap-set and the post implementation data, however the bad rates across the score intervals were on different levels due to the macroeconomic environment. Using a polynomial curve of the log-odds–to-score relationship of the post implementation data, the development data after the swap-set was inferred to obtain a statistical performance threshold based on post implementation performance. In the final stage of the proposed methodology the new statistical performance measures from development were compared to the post implementation data. The swap-set Gini-coefficient was calculated as 45% which is a significant change from the original 36% and which compares more intuitively with the 48% Gini-coefficient of the post implementation data.

The methodology and the introduction of the swap-set Gini-coefficient adds to the contribution of this paper. The methodology together with the swap-set Gini-coefficient and other measures gave intuitive results and the new thresholds are a significant change to what were generally used for application scorecard monitoring purposes for the accepts population.

5.1.3 Calibration

In Chapter 4, the procyclicality problem was addressed whereby credit scorecard models giving different bad rate levels when comparing the performance in a downturn period versus the performance in a benign or upturn period. By using the method of calibration, a credit scorecard can be adjusted. Calibration is a requirement whenever the scores from the credit scorecard model cannot directly be associated with the required probability estimates as a

result of the statistical modelling technique used, the passing of time or the difference between model's target variable and metric to be estimated.

A methodology was presented to analyse the effect of procyclicality on a credit scorecard model and illustrating the use of a calibration model to adjust to expected future bad rate levels. Firstly, methods for identifying downturn periods were investigated. By using GDP growth, elevated bad rate levels and recovery levels the downturn period was identified as the period October 2008 up to and including September 2009. The upturn period chosen was the period January 2010 up to and including December 2010.

In the next step of the methodology, the performance of the bureau scorecard was investigated. This analysis indicated that the bureau scorecard is monotonic in both the downturn and upturn periods. However, the level of bad rate is on different levels across the score intervals with the downturn period on a higher level. This indicated the necessity of the calibration of the credit scorecard model. An eight step calibration model was presented to calibrate the credit scorecard model to expected future bad rate levels.

In the first step the average bureau score is determined across the bureau scorecard score intervals. Secondly the bad rates for the accepts population is determined for each bureau score interval. Under the third step the bad rates of the accepts population is smoothed with step 4 carrying out the calibration technique using the explicit form of the Bayesian theorem. To obtain a TTD application view, the bad rates for the rejected applications were also determined for each bureau score interval. This was done using a polynomial function of the upturn periods accepts population performance to infer the rejected applications.

Using the reject rate and reject factor determined in steps 5 and 6, a calibrated bad rate is obtained for each bureau score interval under step 7. In the final step of the methodology the calibration model is tested by firstly plotting the actual bad rates against the expected bad rates produced by the calibration model. Using the Student's t-test, the hypothesis that the difference between the actual bad rate and expected bad rate $= 0$, gave a p-value of 0.29 which is > 0.05 (the threshold indicating that the hypothesis is not rejected).

This paper contributes a methodology and calibration model to calibrate PIT bad rates per score interval to expected future bad rate levels. This paper combined the probability origins of the Bayesian theorem with a real-world problem to facilitate calibration. Three scenarios

were carried out in an independent out-of-sample to compare predicted expected bad rates against actual bad rates. Comparisons indicated a significant relationship between actual bad rate levels and expected predicted bad rate levels.

5.2 Recommendations

5.2.1 Credit Scoring Matrix Model (CSMM)

The CSMM presented in this paper combines an internal application scorecard with a bureau scorecard. Possible future research could explore a similar matrix model, but basing the bureau scorecard on specific bureau related characteristics that can prove to be more relevant to the financial institution in question. The construction of the CSMM can also be explored to simplify and simulate the determination of the score intervals.

5.2.2 Swap-set Gini-coefficient

The new statistical performance measure called swap-set Gini-coefficient that was presented in Chapter 3 provides a more efficient way to compare development data with post implementation data for the accepts population. As with any new credit scorecard implementation, time is required to observe bad rate performance.

Future research could include the detection of the bad rate performance post implementation as quickly as possible to observe the performance of the credit scorecard. Exploring more efficient ways to monitor the TTD population to detect the performance of the credit scorecard used for acquisition purposes should be considered.

5.2.3 Calibration

Methods for identifying downturn and upturn periods were illustrated. The BIS provided guidelines on methods to obtain such periods in the business cycle. This could be explored in more detail whereby different methods to identify downturn periods could potentially be obtained.

Scenarios affecting scorecards was discussed in Chapter 4 which include the age of the scorecard. This is another area open for research that can have an effect on the performance of a credit scorecard during the business cycle.

In Chapter 4 calibration was applied on a bureau scorecard with less focus on non-bureau scorecard, future research should concentrate on scorecards that can contribute in the guidance on credit decisions that are non-bureau information related.

5.3 Contribution

The three topics explored in this thesis contribute to credit risk measurement and management in the following ways summarised in Table 5.1.

Table 5.1: Summary of thesis contributions.

Attribute	Problem statement	Analysis	Results
Optimise credit scoring and determine appropriate cut-off score	Address the issues raised especially obtaining an optimal model and determining the appropriate cut-off score	Submitted for publication in *South African Journal of Economics and Management Sciences*	CSMM gives uplift in the Gini-coefficient and provides a more granular level to determine the appropriate cut-off score
Enhance application scorecard monitoring for the accepts population	Statistical performance measures differ between development and post implementation data for the accepts population	Submitted for publication in *Applied Economics*	Methodology presented and introduced swap-set Gini-coefficient provides significant changes to the threshold for use in monitoring the accepts population
Bureau scorecard performance during the business cycle	A changing economy affects bureau scorecards. An adjustment is needed	Submitted for publication in *Applied Economics*	Bureau scorecard is monotonic in both downturn and upturn periods, however on different bad rate levels. The explicit form of Bayes theorem calibrated to expected future bad rates indicates a significant relationship with actual bad rates

Within each topic, a literature review was conducted that includes non-parametric and parametric statistical techniques to build credit scoring models, statistical performance measures, scenarios affecting scorecards, detect downturn periods and identify calibration techniques which include the probability origins of the Bayesian theorem.

Literature in credit scoring is limited with no optimal technique existing to build a credit scorecard model. Chapter 2 explored this and proposed a CSMM that both optimises the credit scorecard model and provides a matrix that enables greater granularity for setting the appropriate cut-off score.

Credit scorecards are usually monitored on the accept population due to the effort risk managers apply on risk strategies and making cut-off decisions. However, statistical performance measures appear to differ with post implementation data for the accepts population. Chapter 3 introduces a methodology and the swap-set Gini-coefficient to monitor credit scorecards more effectively between the development and post implementation data for the accepts population.

Procycllicality affects bureau scorecards. Chapter 4 illustrated that the monotonicity holds during the downturn and upturn period, however the level of bad rates is higher in the downturn period across the score intervals. This paper contributes by connecting the probability origins of the Bayesian theorem with a real world problem that provides a significant relationship between predicted bad rates and actual bad rates.

5.4 Final statement

Credit scoring play an important and critical role in the finance industry. In 1956 Engineer Bill Fair and mathematician Earl Isaac pioneered a credit scoring technique which has been widely adopted ever since (Anderson, 2007, p. 40). Vast areas of research in credit scoring are still possible with the limited research that are available. It is paramount that financial institutions have credit scorecards in place to guide credit decisions which affects future impairments, capital requirement for unexpected losses and profits.

The studies conducted in this thesis indicate the effectiveness of credit scorecards to separate good and bad clients. It is important to optimise, validate and constantly monitor credit scorecards to ensure clients credit risk is assessed as accurately as possible. It is also important to calibrate credit scorecards when necessary to account for expected future bad rate levels. Significant contribution to enhance credit risk measurement and management can be and has been made through the studies covered in this thesis.

Bibliography

Abdou, H. & Pointon, J., 2011. Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance & Management,* 18(2-3), pp. 59-88.

Allen, L., De Long, G. & Saunders, A., 2004. Issues in the credit risk modelling of retail markets. *Journal of Banking and Finance,* Volume 28, pp. 727-752.

Anderson, R., 2007. *The credit scoring toolkit, theory and practise for retail credit risk. Management and decision automation..* United States: Oxford University Press Inc..

Basel Committee on Banking Supervision, 2004. *International Convergence of Capital Measurement and Capital Standards,* s.l.: Bank for International Settlements.

Basel Committee on Banking Supervision, 2005. *Guidance on Paragraph 468 of the Framework Document,* s.l.: Bank for International Settlements.

Basel Committee on Banking Supervision, 2006. *International Convergence of Capital Measurement and Capital Standards,* s.l.: Bank for International Settlements.

Basel Committee on Banking Supervision, 2006. *The IRB Use Test: Background and Implementation,* s.l.: Bank for International Settlements.

Basel Committee on Banking Supervision, 2012. *Core Principles for Effective Banking Supervision,* s.l.: Bank for International Settlements.

Beveridge, S. & Nelson, C. R., 1981. A New Approach to Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the Business Cycle. *Journal of Monetory Economics,* Volume 7, pp. 151-174.

Bohn, J. R. & Stein, R. M., 2009. *Active Credit Portfolio Management in Practice.* First ed. New Jersey: Wiley.

Bonini, S. & Caivano, G., 2014. Probability of Default: A Modern Calibration Approach. In: C. Perna & M. Sibillo, eds. *Mathematical and Statistical Methods for Actuarial Sciences.* Switzerland: Springer International Publishing, pp. 41 - 44.

Burns, A. F. & Mitchell, W. C., 1946. *Measuring Business Cycles,* s.l.: National Bureau of Economic Research.

Collins, M., 2016. *The Global Credit Cycle,* s.l.: Prudential Financial.

Durand, D., 1941. *Risk elements in consumer instalment financing, studies in consumer instalment financing.* New York: National Bureau of Economic Research.

Engelman, B. & Rauhmeier, R., 2011. *The Basel II Risk Parameters: Estimation, Validation, and Stress Testing - with Applications to Loan Risk Management.* Second ed. Berlin: Springer.

F., 2014. *FICO Model Builder.* United States of America.

FICO, 2014. *Building powerful, predictive scorecards. An overview of scorecards module for FICO Model Builder,* s.l.: FICO.

FICO, F. I. C., 2011. *Explanation of Divergence,* s.l.: FICO.

Fisher, R. A., 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics,* 7(2), pp. 179-188.

Galton, F., 1889. Quoted in Anderson, R., 2007. *The credit scoring toolkit, theory and practise for retail credit risk. Management and decision automation..* United States: Oxford University Press Inc..

Gini, C., 1910. 'Indici di Concentrazione e di dipendenza'. Atti della III Riunione della Societ'a Italiana per il Progresso delle Scienze. Reprinted in his 1955 'Memorie di methodologia statistica, I, Variabiliá e Concentrazione'. Libreria Eredi Virgilio Veschi: Rome.. pp. 3-120.

Glößner, D. P., 2003. *Calculating Basel II Risk Parameters for a Portfolio of Retail Loans,* Oxford: Thesis submitted in partial fulfilment of the requirements for the MSc in Mathematical Finance.

Gordy, M. B. & Howells, B., 2004. *Procyclicality in Basel II: Can We Treat the Disease Without Killing the Patient?.* [Online]
Available at: https://www.bis.org/

Government Gazette, 2006. *National Credit Act,* Republic of South Africa: The Presidency.

Hand, D. J. & Henley, W. E., 1997. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 160(3), pp. 523-541.

Harrell, F. E., 2015. *Regression Modeling Strategies.* Second ed. Switzerland: Springer International Publishing.

Huang, E. & Scott, C., 2007. *Credit Risk Scorecard Design, Validation and User Acceptance - A Lesson for Modellers and Risk Managers,* s.l.: University of Edinburgh Business School.

Ingolfsson, S. & Elvarsson, B., 2010. Cyclical adjustment of point-in-time PD. *Journal of the Operational Research Society,* Volume 61, pp. 374-380.

Iqbal, N. & Ali, A., 2012. Quoted in Bonini & Caivano, 2014. Probability of Default: A Modern Calibration Approach. In: C. Perna & M. Sibillo, eds. *Mathematical and Statistical Methods for Actuarial Sciences.* Switzerland: Springer International Publishing, pp. 41 - 44.

Johnson, R. W., 2004. Legal, social, and economic issues in implementing credit scoring in the United States. In: *Readings in credit scoring: Recent develpments, advances, and aims.* s.l.:Oxford University Press, pp. 5-15.

Kiefer, N. M., 2008. Quoted in Bonini & Caivano, 2014. Probability of Default: A Modern Calibration Approach. In: C. Perna & M. Sibillo, eds. *Mathematical and Statistical Methods for Actuarial Sciences.* Switzerland: Springer International Publishing, pp. 41 - 44.

Kolmogorov, A. N., 1933a. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer.

Kolmogorov, A. N., 1933b. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell Istituto Italiano degli Attuari*, pp. 483-491.

Landau, J.-P., 2009. *Procyclicality - what it means and what could be done,* s.l.: Bank for International Settlements.

Lewis, E. M., 1992. *An introduction to credit scoring.* 2nd ed. San Rafael CA: The Athena Press.

Lim, M. K. & Sohn, S. Y., 2007. Cluster-based dynamic scoring model. *Expert Systems with Applications,* Volume 32, pp. 427-431.

Lorenz, M. O., 1905. Methods of Measuring the Concentration of Wealth. *Publications of the American Statistical Association,* Volume 9, pp. 209-219.

Medema, L., Koning, R. H. & Lensink, R., 2009. A practical approach to validating a PD model. *Journal of Banking & Finance,* Volume 33, pp. 701-708.

Miu, P. & Ozdemir, B., 2006. Basel Requirements of Downturn Loss Given Default: Modelling and Estimating Probability of Default and Loss Given Default Correlations. *Journal of Credit Risk,* 2(2), pp. 43-68.

Nelson, C. R. & Plosser, C. I., 1982. Trends and Random Walks in Macroeconomic Time Series. *Journal of Monetary Economics,* Volume 10, pp. 139-162.

Pareto, V., 1896. *Cours d'économie politique,* s.l.: Université de Lausanne, 3 volumes, 1896-1897.

Pluto, K. & Tasche, D., 2005. Quoted in Bonini & Caivano, 2014. Probability of Default: A Modern Calibration Approach. In: C. Perna & M. Sibillo, eds. *Mathematical and Statistical Methods for Actuarial Sciences.* Switzerland: Springer International Publishing, pp. 41 - 44.

Sears, 1950. Quoted in Anderson, R., 2007. *The credit scoring toolkit, theory and practise for retail credit risk. Management and decision automation..* United States: Oxford University Press Inc..

Siddiqi, N., 2006. *Credit risk scorecards, developing and implementing intelligent credit scoring.* United States of America: John Wiley & Sons, Inc..

South African Reserve Bank, 2007. *Introduction to the South African Reserve Bank.* [Online]
Available at: http://www.resbank.co.za
[Accessed July 2007].

Stirzaker, D., 1999. *Probability and Random Variables - a beginners guide.* 1st ed. Cambridge: Cambridge University Press.

Tasche, D., 2013. Quoted in Bonini & Caivano, 2014. Probability of Default: A Modern Calibration Approach. In: C. Perna & M. Sibillo, eds. *Mathematical and Statistical Methods for Actuarial Sciences.* Switzerland: Springer International Publishing, pp. 41 - 44.

Thomas, L. C., 2000. A survey of credit and behavioural scoring: forecastingfinancial risk of lending to consumers. *International Journal of Forecasting,* Volume 16, pp. 149-172.

Thomas, . L. C., Edelman, D. B. & Crook, J. N., 2002. *Credit scoring and its applications.* Philadelphia: SIAM.

TransUnion, 2015. *Credit Bureau, Credit Reporting Companies. Learn about your credit..* [Online]
Available at: https//www/mytransunion.co.za
[Accessed 2 4 2015].

TransUnion & FICO, 2009. *Empirica. Minimise your credit risk. Increase your profitability.,* Johannesburg: TransUnion.

Turri, A. & Salis, F., 2010. *Downturn LGD: Flexible and Regulatory Consistent Approaches.* [Online]
Available at: http://www.greta.it/credit/credit2010/PAPERS/Posters/Turri_Salis.pdf
[Accessed 11 7 2016].

Van Gestel, T. & Baesens, B., 2009. *Credit Risk Management - Basic concepts: financial risk components, rating analysis, models, economic and regulatory capital.* New York: Oxford University.

Wells, H., 1992. Quoted in Lewis, E. M., 1992. *An introduction to credit scoring.* 2nd ed. San Rafael CA: The Athena Press.

Wonderlic, E. F., 1946. Quoted in Johnson, R. W., 2004. Legal, social, and economic issues in implementing credit scoring in the United States. In: *Readings in credit scoring: Recent develpments, advances, and aims.* s.l.:Oxford University Press, pp. 5-15.