MCOM Mini-Dissertation

Comparison of Logistic Regression and Classification Trees to Forecast Short Term Defaults on Repeat Consumer Loans

By

Keeland Naicker (NCKKEE005)

Department of Finance and Tax

University of Cape Town, South Africa, 7700

SUBMITTED TO THE UNIVERSITY OF CAPE TOWN

In partial fulfilment of the requirements for the degree

Masters of Commerce specialising in Corporate Finance and Valuations

Supervisor: Professor Kanshukan Rajaratnam

School for Data Science and Computational Thinking

University of Stellenbosch

Abstract:

This dissertation highlights the performance comparison between two popular contemporary consumer loan credit scoring techniques, namely logistic regression and classification trees. Literature has shown logistic regression to perform better than classification trees in terms of predictiveness and robustness when forecasting consumer loan default events over standard twelve-month outcome periods. One of the major shortcomings with classification trees is its tendency to overfit data eroding its robustness, making it vulnerable to underlying population characteristic shifts. Classification trees remains a popular technique due to its ease of application (algorithm machine learning basis) and model interpretation. Past research has found classification trees to perform marginally better than logistic regression with respect to predictiveness and robustness when modelling short term consumer credit default outcomes related to previously unseen new customer credit loan applications. This dissertation independently tested this finding on reloan consumer loan data, repeat customers who renewed loan facilities at a significant South African micro lender. This dissertation tests the finding if the classification tree technique would outperform logistic regression when modelling this new type of loan data. Credit scoring models were built and tested for each respective technique across identical data sets with the intent to eliminate bias. Robustness tests were constructed via careful iterative data splits. Performance tests measuring predictiveness and robustness were conducted via the weighted sums of squared error evaluation approach. Results reveal logistic regression to outperform classification trees on predictiveness and robustness across the designed uniform iterative data splits, which suggests that logistic regression remains the superior technique when modelling short term credit default outcomes on reloan consumer loan data.

Contents

List of Figures

List of Tables

List of Equations

1. Introduction

1.1 Introduction to Credit Scoring

Credit scoring is and will continue to be the pivotal procedural process in debt origination (Hand & Henley, 1997: 524). Credit originators the world over use techniques, ranging in sophistication used in the loan origination decision process. Over time these techniques have moved from subjective rudimentary to mathematically sophisticated. These sophisticated tools can incorporate complex information, which can be applied on large volumes of applications at high speed automation accompanied with low costs. This dissertation compares two highly sophisticated techniques currently used in practice, namely, logistic regression and classification trees. The comparison will specifically investigate the suitability of each technique when applied to a pre-defined data set, exploring topics on predictive ability and robustness. The literature has explored this topic, with evidence indicating that within a credit scoring application, logistic regression is more predictive and robust compared to classification trees. This dissertation however differs to the literature by the nature of the outcome events modelled. Outcome events are the specific measures being modelled by the decisioning tool, for credit scoring it is predicting the propensity of a client to default on their loan repayment commitments (simply, propensity to default). In the general literature, the propensity to default is measured over long outcome periods, generally twelve months after origination. This dissertation explores a much shorter outcome period of five months. This shorter outcome period brings into consideration the classification tree technique, which literature has shown to be predictive in modelling shorter outcome periods (Chang et al., 2016: 6814). The important context to modelling short term default outcomes (Chang et al., 2016: 6814) is that time to default is very important in profit consideration, separating from merely considering default outcomes over a traditional outcome period. Profit losses incurred are larger when loans default faster.

In addition to the modelling of a shorter default period, this study will analyse repeat consumer reloan data (customers who have settled and originated repeat credit loans within a credit institution). Modelling credit risk on repeat consumer reloan's in contrast to new consumer loan data follows the same modelling principles, however, repeat consumer reloan's are lower risk

and lower cost, with organisations making better credit decisioning through retained customer behaviour records, offering repeat credit agreements to good performing clients at a lower cost acquisition in comparison to new customers (Churchill., 2000: 9). The same modelling principles incorporate, modelling technique selection, feature selection and modelling default outcomes.

Credit is defined as money lent to a client by a financial institution which is to be consequently repaid with interest according to a mapped amortization schedule drawn up prescribing the repayment amounts (Hand & Henley, 1997: 524). The lent monies are defined as an asset by the financial institution and correspondingly as a liability by the counterparty. Sophisticated credit scoring modelling is the formal statistical process identifying the probability by which a prospective client will repay their loans. This probability is interchangeably known as the propensity of the client to default. Effectively it is estimating the probability to which a client will default on their future repayments on borrowed credit liabilities. This dissertation will focus on modelling consumer credit, where the underlying borrower is a human entity within a private consumer capacity. Statistical credit scoring models, called scorecards or classifiers are built using predictor variables to estimate the probability of default on a prospective client. Predictor variables are typically the client's past payment behaviour across the client's commitments housed at the originating institutions and other third-party credit providers (Hand & Henley, 1997: 524). Typical providers of consumer credit are, but not limited to:

- Banks
- Private credit providing institutions
- Building societies
- Retailers
- Mail order companies
- Utilities

Over time, the logistic regression scorecard building technique has been seen to be the relied credit scoring model build process, used by credit institutions the world over in creating automated predictive credit scoring tools. It has proved to be reliable in its predictiveness, be it, that it to suffers degradation over time, like any other asset depreciation. However, with the

transition into an age with increased volume in data depth and width accompanied with increasing computing ability to process this higher volume of data, rival approaches have emerged, with an inherent challenge to the traditional logistic scorecard building techniques. Hand & Henley (1997: 524) list the various types of credit scoring techniques used in industry:

- Linear discriminant analysis
- Linear regression
- Logistic regression
- Classification and Regression Trees
- Probit analysis (type of parametric linear regression)
- Non-parametric smoothing methods
- Mathematical programming
- Markov chain models
- Recursive partitioning
- Expert systems
- Genetic programming algorithms
- Neural networks
- Conditional independence models
- Rough Sets

Huang et al. (2007: 848) list the following additional methods available to credit scoring model builds:

- K-nearest neighbour
- Support vector machines (SVM)

This dissertation will focus in on the comparison between the parametric logistic regression model building technique against the machine learning, non-parametric classification tree technique.

Blanco et al. (2013: 356) highlight the importance of credit scoring to business in that it reduces cost of credit analysis, improves cash flow, enables faster credit decisions, reduces losses (particularly bad debt losses), results in the closer monitoring of existing accounts and the

prioritisation of repayment. Hand & Henley (1997: 523) note, apart from the obvious reduction in bad debt losses, revenue is an important consideration. Profitability and revenue need not be monotonically related to risk. They point to an example, where low risk clients who pay off their credit card liabilities will be regarded low risk though will also generate low revenues for the business. Interest charges on these low risk clients are low. On the other hand, higher risk clients who make minimum required payments, accrue interest on their outstanding balances, allowing for the generation of interest revenue accruing to the loan provider. Credit providers attempt to maximise profits by choosing the low risk clients (low bad debt losses) accompanied with high revenue streams. Mortgages are typical of this type of product, though during the 2008 US mortgage crisis businesses went deeper into higher risk markets which were accompanied by high bad debt losses.

The underlying data modelled is of a large South African micro lender, disbursing fixed term annuity paydown based unsecured personal loans. This dissertation will focus specifically on this micro lender's existing customer reloans portfolio. Thomas (2000: 149) defines behavioural credit scoring the same as the existing customer reloan credit scoring concept in this dissertation. Thomas (2000: 149) defines new loan credit granting to be a judgmental focused analytical process whereas behavioural scoring (reloan) is defined more as a behavioural focused analytical process, though using the same statistical modelling approaches. Thomas (2000: 149) also shows that behavioural credit models typically use an outcome performance period between 12 and 18 months, to allow more time for default behaviour to mature and precision in accurate default forecasting, paramount for behavioural credit risk profit modelling. Horng et al. (2010: 1) showed that backpropagation neural network techniques to outperform linear discriminant analysis and support vector machine techniques when modelling behavioural credit card default data for clients based in Taiwan. Loans are fixed term amortised products, with a defined repayment period. The South African consumer credit market and banking system is well developed. The total balances outstanding in the South African credit consumer market as at the end of 2019 was R1.95 trillion (TransUnion [TU], 2020: 2). Table 1 and accompanying bullet points represent a snapshot of the South African personal loans market as the end of 2019 (TU, 2020: 9), (StatsSA, 2020: 10):

Number of loans	13.1m
Value of Loans	R343bn
Ave. balance per account	R26.4k
Proportion of personal loans as percentage of total consumer credit	17.6%

Table 1 - Summary of South African Personal Loans Market (December 2019)

- Total outstanding South African consumer credit balances as at the end of 2019 as a percentage of 2019 full year Gross Domestic Product – 61.9%
- Total outstanding South African personal loans consumer credit balances as at the end of 2019 as a percentage of 2019 full year Gross Domestic Product – 10.9%
- Balance delinquency rate of South African personal loans as at the end of 2019 (3+ months in arrears) – 28.3%

Personal loans comprise a sizable portion of the total South African consumer credit market. Total consumer credit includes other big balance items such as housing bonds, vehicle finance, revolving and credit card facilities and retail credit facilities. An important consideration in the statistics above is the high balance delinquency rate incurred on South African personal loans. This bad debt cost incurred is countered by the good revenue margins South African personal loans credit providers enjoy, via high interest rates and auxiliary account charges, set by regulatory authorities.

1.2 Research Question

The purpose of this dissertation will be to compare the performance in the logistic regression scorecard building technique against one of the challengers, the classification tree technique. Two specific questions will be interrogated in the context of shorter credit default outcome periods and the underlying data modelled being reloans (repeat existing client loans):

- Which technique proves to be more *predictive*, tested on pre-defined data sets?
- Which technique proves to be more *robust regarding stability*, tested on pre-defined data sets?

1.3 Aims and Objectives

Literature has shown that when modelling standard twelve-month default outcome periods, classification trees have performed, marginally below par relative to logistic regression in relation to the two research questions described in Section 1.2. Classification trees however remain a popular challenger to logistic regression and other techniques, with the reason being that studies have shown classification trees to give favourable results when modelling short term outcomes (Chang et al., 2016:6814). This dissertation aims to show that classification trees could match or better logistic regression techniques in terms of predictiveness and robustness, when specifically modelling shorter credit default outcome periods with underlying data being existing customer repeat loans. Given this is a mini masters dissertation, the development of new theory and models are out of scope. This dissertation aims to evaluate the quantitative comparison between existing models.

1.4 Structure of the Dissertation

The remainder of this dissertation comprises five chapters. Chapter 2: Literature Review, introduces the credit scoring concepts moving onto results of sampled credit modelling studies comparing the various credit scoring techniques, finally laying the platform for this dissertation's research question. Chapter 3: Methodology, details the mathematical underpinnings for both the classification tree algorithm and logistic regression modelling techniques. Chapter 4: Data and Model Construction, introduces the underlying credit reloan data to be modelled which includes defining the credit default outcome and understanding the predictor model variables, concluding with the details of the models created. Chapter 5: Analysis of Results, introduces the weighted sums of squared error measure, being the primary tool used in answering the research questions progressing to the analysis of the model results in context to the research questions. Chapter 6: Conclusion and Future Research, recaps key findings of the dissertation in relation to the research questions and suggests an extension of this dissertation to different applications.

2. Literature Review

The evolution of credit scoring has progressed from subjective experienced based methods to current day objective statistical models employed. The subjective expert based methods or traditional manual methods were when bank managers typically used credit reports, customer personal histories and experience judgement to make credit decisions (Mester, 1997: 5). The so called "4 C's" of credit (Altman & Saunders, 1998: 1972) where bankers used information on various borrower characteristics, borrower *character* (reputation), *capital* (leverage), *capacity* (volatility in earnings) and *collateral* were the cornerstone features to the subjective credit decisioning. The current day objective credit decisioning models are statistical and computational models. These typically include discriminant models (for example, linear discriminant models (LDA), quadratic), regression models (for example, logit, probit) and inductive models (for example, neural networks, genetic algorithm) (Hand & Henley, 1997: 531), which attempt to link the elements of the "4 C's" and other predictive factors to the likelihood of default through an automated standard credit decisioning process. Regardless of the technique used to build the credit scoring models, Galindo & Tamayo (2000: 114) note that in economics and finance, good classification and predictive models should have the following five properties:

1. Accuracy: having low generalization error rates
2. Parsimonious: representing and generalizing the relationships in succinct manner
3. Non-Trivial: results are interesting
4. Feasible: in terms of time and resources
5. Transparent and interpretable: being able to provide high level representations of the relationships inferred from the underlying data

Ibtissem (2014: 10) notes two key features required to be satisfied in a good credit scoring model, namely, transparency and accuracy. Accuracy enables correct credit assessment by optimizing losses. Transparency refers to the good interpretability of model, the mapped credit scoring process between inputs and outputs should be clear and concise. Gurny & Gurny (2013: 163) note that the incorrect estimation of probability of defaults will lead to an unreasonable rating, incorrect pricing of financial instruments, incorrect calculation of economic capital or regulatory

capital required to be held for Basel III requirements for a deposit taking financial institution. At a portfolio level aggregating to industry secular levels, Altman & Saunders (1997: 1722) note that incorrect credit risk estimation can lead to increase in the number of structural bankruptcies, declining value of real assets (and thus collateral), dramatic growth of off-balance sheet instruments with inherent default risk exposure.

The different credit scoring models should be considered as different routes and techniques to all arrive at the same point, used in the decisioning process to either grant or fail the prospective credit applicant. The power of credit scoring models is critically dependent on the data used to build them, called features, predictor or independent variables. The quality of the credit scoring tool used to accurately predict outcomes, depends heavily on the functional inputs, hence quality feature selection is arguably the most important ingredient to building powerful credit scoring models. Hand & Henley (1997: 528) suggest feature selection to be performed via expert knowledge, using stepwise statistical procedures and using information value as a measure of separation. Hand & Henley (1997: 524) go deeper on feature selection, noting features used in the model development differs greatly by product. Mortgage loans scoring features will differ to credit card features which will differ to that of mail order products. In tough economic times, client credit repayment hierarchies change according to needs, particularly when a client has multiple credit products in their portfolio. Feature selection also changes according country specific legislation, like for example race and gender cannot be used in South African credit scoring models (*National Credit Act* , *Act No. 34 of 2005*, 2005: Vol. 489). Hand & Henley (1997: 524) note credit bureau characteristics to be important to the credit scoring process, for example the number and details of loan accounts, details of slow payments, bankruptcies, number of requests for new credit and so forth on prospective scored credit clientele. Kocenda & Vojtek (2009: 1) highlight socio-demographic features being powerful for a credit scoring study performed on European Union markets, going further onto show that socio-demographic and behavioural features evolve over time in a stable manner. Khandani et al. (2010: 2772) highlight that intuitive feature selection is required for a parsimonious feature input selection set.

As time evolves population characteristics and distributions forming the makeup of features change, called population drift, the key factor contributing to probability of default scorecard

instability. Hand & Henley (1997: 525) highlights population drift as an important facet in scorecard robustness. Population drift can also be described as the tendency of performance within populations to evolve with time. This is perfectly expected due to changing consumer environments with different economic pressures and booms but the magnitude of the shift tests the predictive robustness and stability of the built scorecards. Population drift is a pertinent concept tested in this dissertation, as highlighted in the objective to test robustness as more data is added to model build. Hand & Henley (1997: 525) noted that techniques which are robust against negatives affects from population drift and which are less cost intensive (time and money resources) would be favourable. Model power increases would result from more intelligent predictive characteristics.

Numerous studies referenced in this dissertation have compared the various modelling techniques over time. Blanco et al. (2013: 357) note that artificial neural networks represent the most powerful of the non-parametric statistical model's due to their non-linear and non-parametric adaptive learning properties. They demonstrate that non-parametric credit scoring models based on the multilayer perceptron approach, which is a type of neural network outperform three other traditional parametric techniques, linear discriminant analysis, quadratic discriminant analysis and logistic regression techniques. Tam & Kiang (1992: 942) found that neural networks are the best predictor of credit default, with logistic regression marginally trumping decision trees. They found that neural network models are most accurate, adaptive and robust over all other techniques.

Dong et al. (2012: 2463) highlight the most important reason for logistic regression being the most commonly used credit scoring technique in the banking industry due to its desirable features of robustness and transparency. Many techniques, particularly neural networks have shown superior prediction accuracy, though suffer the major problem of having difficulty of interpretability. Regulated banking institutions require high transparency of models as regulation requires, black box models such as neural networks lack that transparency.

Logistic regression allows for the features to be non-normal, overcoming the assumptions required by the linear discriminant models, it uses maximum likelihood estimation for predictor

parameters, provides the probabilities toward a definite class (good or bad) and can deal with categorized data. Through this dissertation, goods refer to observations who don't reach default with bads being vice versa, where observations reach a default status. Studies show where logistic regression outperforms discriminant models (Vojtek & Kocenda, 2006: 162). Charitou et al. (2004: 492) found the logit method superior to other methods. West (2001: 1131) and Baesons et al. (2003: 627) find that logistic regression to be more accurate and robust than classification trees. Ong et al. (2005: 41) show that, to increase accuracy and flexibility to the traditional binary logistic regression, models can be extended to include multinomial logistic regression models and logistic regression models for ordered categories.

Davis et al. (1992: 43) showed how classification tree models can be successfully applied to build credit scoring models. Ruonan (2013:1) showed that random forests (classification trees) were found to outperform on fit (better on Gini-coefficient, Kolmogorov-Smirnov and classification rate error) than neural networks, logistic regression and classification trees, for China consumer data. Galindo et al. (2000: 107) show that classification trees performed better in terms of lowest error rates with artificial neural networks coming in second and probit (logistic regression) coming in third. Khandani et al. (2010: 2777) highlight classification tree models overcome limitations of logistic regression models, where the dependent variable is forced to fit a single linear model throughout the entire input space.

Ong et al. (2005: 41) compared built credit scoring models across different techniques, with emphasis on genetic programming. They unpacked that genetic programming provides better performance as compared to credit models developed via logistic regression (parametric), artificial neural networks (computing) and decision trees (non-parametric). Genetic programming was better than its closest rival, artificial neural networks, because it was found to also work effectively on small data sets and exhibited better performance on validation data sets. Genetic programming is better able to select discriminant features in an un-supervised learning environment, particularly because artificial neural networks has been shown to choose irrelevant feature characteristics. Ong et al. (2005: 41) also found, setting genetic programming aside, that artificial neural networks and logistic regression outperformed the other methods, including classification trees.

Ibtissem (2014: 7) took a contrarian approach to credit scoring modelling, testing a hybrid fuzzy logic approach to credit scoring against the classification tree modelling approach. Classification trees were found inferior to the hybrid fuzzy logic approach, with respect to accuracy. Ensemble techniques, namely hybrid models are also entering the fray. For example, the cross technique between logistic regression and neural networks, which share the robustness of logistic regression value added with the predictive accuracy of neural networks.

Hand & Henley (1997: 536) could not find a large variance in techniques estimating the risk classifying good and bads. Galindo & Tamayo (2000: 115) show that machine learning algorithms offer the advantages of being more computational-based without the need of distribution assumptions, with the ability to model non-linear, non-normal, complex relationships however they can be dis-advantageously large, idiosyncratic and often difficult to interpret (artificial neural networks). Galindo & Tamayo (2000: 115) also note that there is no single method or algorithm which is perfect or guaranteed to work, hence users must be aware of the strengths and weaknesses in each technique.

Summarising the findings in the literature it is inconclusive that there emerges a single superior technique that convincingly satisfies all tests, being combined best in class for predictiveness, transparency and robustness. Each technique has its own strengths and weaknesses, hence the reason for all techniques still be being tested and applied in industry and academia.

This dissertation interrogates a specific question, between classification trees and logistic regression, on which technique proves to have better predictiveness and robustness when modelling a short-term credit default outcome (five months). The studies in literature predominantly model standard term credit default outcomes (twelve months or longer). Chang et al. (2016: 6814) found that classification trees perform superior to logistic regression techniques, when modelling short term credit default outcomes on new loans. This dissertation will test this finding, to possibly establish classification trees as the superior technique to logistic regression for modelling short term credit default outcomes on reloan data.

3. Methodology

Leading on from Chapter 2, with the research question clearly articulated, a just and fair comparison between the logistic regression and classification tree techniques should be performed on models prepared via robust modelling processes. This will give each technique the best possible presentation at comparison. This chapter unpacks the mathematical foundations of each technique, to highlight the strengths and shortcomings of each technique. Modelling tools which include the statistical computer software packages and various decision process measures leading up the final models are unpacked in detail. These measures are used through the modelling process in Chapter 4.

3.1 Classification and Regression Trees

First developed by Breiman (Breiman et al., 1984) the classification and regression tree (CART) algorithm has applications across various topics, with its first application academically used in credit scoring by Makowski (Makowski, 1985:30). CART is a computational technique which embodies the fitting of a model onto a data set, by classifying paths through combinations of independent variables and their attributes, correlated with a specific outcome. It's essence being a binary recursive partitioning algorithm, establishing well defined paths to a specific outcome. Khandani et al. (2010: 2773) describe classification trees as a supervised learning framework, where a learner is presented with input and output pairs from past data, in which the input data represent pre-identified attributes to be used to determine the output value. Machine learning attempts to map out the functional relationship between input and output. CART can define various paths arriving at the same outcome.

Classification decision trees are defined deep learning models with the following characteristics (Ertel, 2017: 198):

1. Employs data mining and non-statistical techniques to learning patterns on a large set of training data.
2. The model adapts as more data becomes available.
3. Classification trees can be easily displayed graphically for user ease of understanding and interpretation.

4. The one caution however is that there could be a tendency to overfit models making them unusable on different out of time data. Out of time independent data is non- model performance data.

CART is non-parametric and non-statistical. There are no assumptions required for the underlying data to follow a Normal distribution, no required independence amongst predictor variables, no assumptions around linearity required between feature predictor variables and dependent variable and can work with small sample sizes. It is a versatile methodology, essential to most data mining projects. Classification trees deal with binary dependent variables and regression trees deal with continuous dependent variables (Hui et al., 2010: 5895). In this dissertation, classification trees are used to model the dichotomous good bad outcome dependent variable. Good's are defined as loans that do not default in the outcome window and vice versa for Bad's.

The following section deals with the various elements integral to the CART model build methodology. Training and validation data sets are defined, then the sequential mathematical decisioning process inherent in the CART algorithm is unpacked, being entropy, information gain and optimal tree size. It concludes by introducing the software package used to build the CART models, Microsoft Visual Studio (MVS) and its accompanying model building evaluation tools.

3.1.1 Training and Validation Data Sets and Generalization

Models, regardless of technique, are typically created in the following manner. A large subset of the overall data set (usually 80%) is used to "train" the model or develop the decision tree (Sandro, 2018: 58). This resultant model will have associated error misclassification rates (error rates). This model is unique to the underlying training data and it could fit the data well (with an overtly complex structure) which assists in minimising error rates. These trained models are validated on the validation data (residual 20% of the overall data), to gauge if the model can generalize enough (Sandro, 2018: 58). Generalization is the ability for models to fit onto independent data with the desired results (predicted error rates). Going beyond the overall data set, these models should further fit new independent data, with the error rates as predicted, this makes for a good model, fit for purpose. Galindo & Tamayo (2000: 116) notes as sample sizes increase in the training data, the classification tree prediction error begins to grow, conversely,

they found that prediction error in probit models converge to an asymptotic error rate as training data sets get larger, hence they found adding more data to training data sets for logistic regression models don't add any more predictive capacity to the model.

3.1.2 Entropy

Entropy is a measure of information content. It can be measured for an entire training data set or individually per predictor variable. The entropy measure will be unpacked referencing (Ertel, 2017: 201) in the following mathematical construct:

For a data set containing *n* events, with corresponding probabilities p_i such that

$$\sum_{i=1}^{n} p_i = 1$$

Equation 1 - Sum of event probabilities

There are 2 extreme cases:

Case 1: When a specific event will occur with definite certainty and consequently the rest of the *n – 1* events will not occur. The probability distribution across the *n* events is: *p = (1,0,0, 0,…,0)*. The measure of uncertainty regarding the other *n-1* events is described as minimal with no uncertainty.

Case 2: When the *n* events follow a uniform distribution, *p = (1/n, 1/n, …, 1/n)*, the uncertainty is maximal because the likelihood of all events is the same, such that none of the events can be distinguished from the others, in terms of occurrence.

Claude Shannon (Ertel, 2017: 201) explored the number of bits required to encode each of these cases.

Case 1: Zero bits were required because event 1 will always occurs

Case 2: Where there are *n* equally probable possibilities with the events uniformly distributed, with $log_2 n$ bits needed. With each individual event probability being, $p_i = \frac{1}{n}$ then, $log_2(\frac{1}{p_i})$ bits are required for the encoding.

For the general case, $p = (p_1, p_2, \ldots, p_n)$, and the probabilities of the events are not uniformly distributed, then the ***expected value, H,*** for the number of bits required is:

$$H = \sum_{i=1}^{n} p_i \left(-log_2 p_i\right) = -\sum_{i=1}^{n} p_i log_2 p_i$$

Equation 2 - Expected value of number of bits required (H)

Generally, the larger the value for ***H*** the higher the number of bits are required, meaning higher uncertainty. Ertel (2017: 201) cites (Shannon & Holden, (1976)) to have developed the Entropy measure, ***H***, as a measure of uncertainty in relation to a probability distribution is defined as follows:

$$H(p) = H(p_1, \ldots, p_n) := -\sum_{i=1}^{n} p_i log_2 p_i$$

Equation 3 - Entropy Measure H

For case 1 (being maximum certainty) ***H* = 0**, as opposed to case 2 (maximal uncertainty) ***H* = 1**. When extended to an underlying data set *(D)*, the information content of a data set *[I(D)]* is defined as:

$$I(D) := 1 - H(D)$$

Equation 4 - Information content of data set

3.1.3 Information Gain and Feature Selection

The decision tree algorithm recursively evaluates the information gain on variables and their attributes, to create each new node in the tree. The mechanics of the decision tree algorithm is typically described as top down, where the algorithm iteratively partitions the data by evaluating the marginal information gain. Variables and their attributes that have the highest information gain (information content) at each node iteration will go into creating the node in the tree at that point. The data set is subdivided by each new tree node.

Ertel (2017: 203) outlines *information gain G (D, A) as:*

The **information gain *G (D, A)*** using the attribute *A* is determined by the difference of the average information content of the dataset $D = D_1 \cup D_2 \cup \ldots \cup D_n$ divided by the *n*-value attribute *A* and the information content *I(D)* of the undivided dataset, which yields:

$$G(D, A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} I(D_i) - I(D)$$

Equation 5 - Information gain at node

From this definition, we obtain:

$$G(D, A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} I(D_i) - I(D) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} \big(1 - H(D_i)\big) - \big(1 - H(D)\big)$$

$$= 1 - \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i) - 1 + H(D)$$

$$= H(D) - \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i)$$

Equation 6 - Information gain at node (restated)

Where ***H*** is a measure of uncertainty (defined previous). When calculating through the algorithm the information gain, *G (D, A)*, is calculated per variable (and its attributes) at each new node iteration. The variable and attribute which gives the highest information gain is selected as the splitter at that node and so the algorithm iteratively processes until no more information gain per variable is possible. When the variable and its attributes deviate further from the uniform distribution, the higher the information gain possible. When information gain at a particular node is 0 then the algorithm terminates node creation at that particular node, however continues recursively for the other nodes. Ertel (2017: 207) summarizes the mathematical logic of the recursive decision tree algorithm illustrated in figure 1.

***Generate Decision Tree*(*Data*, *Node*)**

$A_{max} = Attribute\ with\ maximum\ information\ gain$

If $G(A_{max}) = 0$

Then *Node becomes leaf node with most frequent class in Data*

Else *assign the attribute* $\boldsymbol{A_{max}}$ *to Node*

For each value $a_1, \dots, a_n$ *of* A_{max}, *generate*

a successor node: $K_1, \dots, K_n$

Divide $Data$ into $D_1, \dots, D_n$ with $D_i = \{x \in Data | A_{max}(x) = a_i\}$

For all $i \in \{1, \dots, n\}$

If all $x \in D_i$ belong to the same class C_i

Then generate leaf node K_i of class C_i

Else Generate Decision Tree (D_i, K_i)

Figure 1 - Generating a Node

Feature selection assists with identifying the best set of predictors which improves model accuracy and makes the process of modelling more efficient. It helps solve two problems, having too much data that is of little value or having too little data that is of high value. The goal is to identify the minimum number of predictors that are significant to building an accurate model. Variables or features and their attributes are selected through the entropy process or iterative information gain process highlighted above. In model development, the parsimonious principle of having the appropriate number of predictors, is key, even more so for algorithm derived machine learning decision trees. Where two models can predict an outcome at a specific error rate, the simpler model will always be preferred. This is Occam's Razor, (Ertel, 2017: 207).

3.1.4 Optimal Tree Size

This section details the optimal size of a decision tree. When allowed to grow without parameter restrictions, the tree will naturally overfit based on the underlying training data. Gaining the optimal tree size is through the cross-validation process of minimization of error expectations.

Ertel (2017: 213), summarizes the mathematical logic of the algorithm finding the optimal tree size parameter, Gamma illustrated in figure 2:

$$\boldsymbol{Cross\ Validation}\ (\boldsymbol{X}, k)$$

$$Partition\ data\ into\ k\ equally\ sized\ blocks\ \boldsymbol{X} = \boldsymbol{X_1} \cup \ldots \cup \boldsymbol{X_k}$$

$$\boldsymbol{For\ all}\ \gamma \in \{\gamma_{min}, \ldots, \gamma_{max}\}$$

$$\boldsymbol{For\ all}\ i \in \{1, \ldots, k\}$$

$$Train\ a\ model\ of\ complexity\ \gamma\ on\ \boldsymbol{X} \backslash \boldsymbol{X_i}$$

$$Compute\ the\ error\ E(\gamma, \boldsymbol{X}_i)\ on\ the\ test\ set\ \boldsymbol{X_i}$$

$$Compute\ the\ mean\ error\ E(\gamma) = \frac{1}{k}\sum_{i=1}^{k} E(\gamma, \boldsymbol{X}_i)$$

Choose the value $\gamma_{opt} = \text{argmin}, E(\gamma)$ with smallest mean error

Train the final model with complexity γ_{opt} on the whole data set $\boldsymbol{X}$

Figure 2 - Finding Gamma, Optimal Tree Size

Summarized, the cross-validation process subsets the training data into *k* evenly distributed subsets. *K* number of models are trained for a specific Gamma value (tree size) across all *K* blocks of data, each time with an associated average error rate. The optimal Gamma chosen would be that associated with the lowest average error rate. Hence it is described to be a numerical estimation method like Newton-Raphson, (Ertel, 2017: 213). Although data intensive, cross-validation is critical in the process of modelling decision trees, addressing the problem in finding the fine balance between error minimization and avoiding overfit models.

3.1.5 Visual Studio Classification Tree Algorithm

This dissertation employs the Microsoft Visual Studio (MVS) classification tree algorithm to build the classification tree models. The MVS algorithm follows the process as mapped out above in creating decision trees. Step one, calculating entropy in aiding feature selection (continuous variables are discretized by the software allowing for meaningful groupings used in the entropy process to calculate information gain measures), (Microsoft, 2020). Step two of growing trees in conjunction with an optimally calculated Gamma such that final model expected error rates are

minimized. This is done via cross validating within training data sets and across to independent data sets culminating in final trained models, (Microsoft, 2020). The MVS package has the following three main tools: tree lift chart, model score and classification matrix, who are used to aid in unpacking model strength and effectiveness described below. These tools are used in the iterative model builds that will follow in Chapter 4.

Figure 5 (Appendix A) and Table 3 (Appendix B) illustrate the concept of the MVS tree lift chart. The chart is a measure of model accuracy, particularly at different cumulative cut-off points of the overall population. Table 3 maps out the ideal model plotted against the proposed fitted model. The ideal model line is a diagonal, with the x-axis on the lift chart measuring overall population percentage covered and corresponding y-axis measuring the percentage of the population correctly classified. The diagonal makes sense, at 50% of the total population, 50% of the population is correctly classified, at 100% of the total population, 100% of the population is correctly classified. The fitted model is measured relative to the ideal model. At 20% of the cumulative population the ideal model classifies 20% of the population correctly, however the fitted model classifies 19.31% correctly. As the cumulative population increases the fitted model begins to diverge from ideal model. This divergence is the type 1 error or false positive rate (FPR) at a cumulative population point. Overall, referencing Table 4 (Appendix B), the example shows the total type 1 error or FPR is 8.68%. MVS is an interaction model developer, with applications in cut-off management when comparing models important in setting cut-off thresholds

The model score is a summary measurement depicting the model fit. Its main use is to compare different models generated via the MVS tree algorithm. It effectively is a relative measure where the population is normalized. The score is calculated as the geometric mean score of estimated probabilities relative to actual classification across the entire population (development and validation combined). The higher the score toward 1 signals a strong model, (Microsoft 2020).

The classification matrix displays the type 1 error (FPR). The lower the FPR the better the model can classify goods and bads. Table 4 (Appendix A) illustrates the FPR rate stated above.

3.2 Logistic Regression

Logistic regression, a generalized linear model (GLM), is a traditional well established statistical technique in creating credit scoring models. It is backed behind well-established statistical theory, GLM models are an extension of linear regression models with the dependent variable allowed to be a non-normal discrete binomial outcome variable (Peng et al., 2002: 4). Credit default events are binary outcomes. The technique performs well on data where a discrete non-normal dependent variable is linearly related to its independent characteristics and enjoys ease of use, encompassing implementation, interpretation and monitoring.

The biggest disadvantage to using logistic regression is its assumption that the dependent variable is linearly related to its independent characteristics where complex data typically may not follow the linear relation, breaking the linearity assumption of linearity (Peng et al., 2002: 4). If the linear relation does not exist, then the resultant models are typically weak in discriminant power.

Logistic regression models a binomial outcome variable (estimating the probability between 0,1) instead of a continuous outcome variable under ordinary least squares (OLS). There are two assumptions required for the method to work well, independent variables are exogenous and optimal for purpose in building linear functional relationships to the dependent variable. Predictor variables need to be unbiased with error terms (difference between expected value and actual values), serially uncorrelated and homoscedastic. In general, the technique works well with large sample sizes where variables are well separated (low correlation) (Peng et al., 2002: 4).

As a build-up to arriving at the final logistic regression models, it is important to unpack its mathematical foundations (logit function) and co-efficient estimations (maximum likelihood function), subsequently moving onto the mechanical process and associated toolkit used to building the logistic regression models.

3.2.1 Logit function

The mathematical formulation begins with the elements, that goes into the arithmetic operation. Peach (2016) constructs out the derivation of the logistic regression in the following steps.

The elements referencing the mathematical formulation of the logistic regression are:

Element 1: Consider outcome variable, Y, as a Bernoulli distributed outcome random variable which takes on two values, 0 and 1. The target group being 'bad' or default coded as 1 with the 'goods' or non-default coded as 0 in the credit risk definitional context.

Element 2: Each observed outcome is associated with a specific vector $\overline{\boldsymbol{X}}$ which is a set of m random variables. This is also called the feature selection set. These are specific independent predictor observations associated with the specific outcome. In any training data set, we have the observed outcomes each with their associated independent predictor observations. All these data points are independently and identically distributed (*iid*)

Element 3: Define the conditional probability $P(Y = 1|\overline{\boldsymbol{X}})$ as the probability of the outcome variable Y = 1 (probability of outcome being bad) given an associated vector $\overline{\boldsymbol{X}}$ of m independent predictor observations. $P(Y = 0|\overline{\boldsymbol{X}})$ is conversely the probability of the outcome being good, given an associated vector $\overline{\boldsymbol{X}}$ of m independent predictor observations.

Element 4: Consider the sigmoid function: $\sigma(z) = \frac{1}{1+e^{-z}}$

Where $z = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$

z is a linear combination of the vector $\overline{\boldsymbol{X}}$ of m independent predictor variables, with the beta values being associated coefficients. These beta coefficients are the maximum likelihood estimators building up the generalised linear model, in attempting to predict the probability $P(Y|\overline{\boldsymbol{X}})$, for a given set of observed predictor variables.

The crucial step in developing the logistic regression formulation is making the central assumption that the predicted probability from the GLM model, $P(Y|\overline{\boldsymbol{X}})$, is approximated by the sigmoid function above, i.e.

$$P(Y|\overline{\boldsymbol{X}}) = \sigma(z) = \frac{1}{1+e^{-z}}$$

Equation 7- Sigmoid Relation

All the elements and assumptions have been declared, the mathematical formulation commences. The natural logarithm is applied of both sides of Equation 7, with step by step subsequent arithmetic operation as follows:

$$ln(P) = ln\big(\sigma(z)\big) = ln\left(\frac{1}{1+e^{-z}}\right)$$

$$= ln(1) - ln(1+e^{-z}) = -ln(1+e^{-z})$$

$$ln(1-P) = ln\left(1 - \frac{1}{1+e^{-z}}\right) = ln\left(\frac{e^{-z}}{1+e^{-z}}\right)$$

$$ln(P) - ln(1-P) = -ln(1+e^{-z}) - ln\left(\frac{e^{-z}}{1+e^{-z}}\right)$$

$$= -ln(1+e^{-z}) - ln(e^{-z}) + ln(1+e^{-z}) = ln(e^{z}) = z$$

$$ln\left(\frac{P}{1-P}\right) = z = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

Equation 8 - Logit Function

Equation 8 is the logit function and illustrates how the probability estimation is linearly related to the underlying predictor variables. Logit(P) is the logarithm of the odds, where odds are the conditional probability of $\frac{"bads"}{"goods"}$

3.2.2 Co-Efficient estimation – Maximum Likelihood Estimators

With the logit function defined, the next task at hand is to find appropriate estimation coefficients to quantify the linear relationship mapped out in Equation 8. Relational coefficients $\overline{\boldsymbol{\beta}}$ are assigned to independent predictor variables via maximizing likelihood estimation. Peach

(2016) summarized the derivation of logistic regression maximum likelihood estimators succinctly in a step wise view as follows:

Step 1: The probability of one data point can be expressed as:

$$P(Y = y|X = \overline{x}) = \sigma(\beta^T \overline{x})^y . [1 - \sigma(\beta^T \overline{x})]^{(1-y)}$$

Where: $\beta^T \overline{\boldsymbol{x}} = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$

And $y \in (0,1)$

Step 2: Establish a likelihood function based on a sample with ***n*** independent observations. Since all ***n*** observations are independent we get the following likelihood function representing the entire data set:

$$L(\beta) = \prod_{i=1}^{n} P\left(Y = y^{(i)} \middle| X = \bar{x}^{(i)}\right)$$

Equation 9 - Summarised likelihood function

Where $P\left(Y = y^{(i)} \middle| X = \bar{x}^{(i)}\right)$ is the conditional probability outlined in step 1 per each observation *i* in the sample ***n***. Inserting the components from step 1, then Equation 9 can be rephrased as:

$$L(\beta) = \prod_{i=1}^{n} \sigma\left(\beta^T \bar{x}^{(i)}\right)^{y(i)} . \left(1 - \left(\beta^T \bar{x}^{(i)}\right)\right)^{\left(1-y(i)\right)}$$

Equation 10 - Expanded likelihood function

Taking the log transformation of Equation 10 we get the log likelihood function (for the entire sample):

$$LL(\beta) = \sum_{i=1}^{n} y^{(i)} log\boldsymbol{\sigma}\left(\beta^T \bar{x}^{(i)}\right) + \left(1 - y^{(i)}\right) log\left[1 - \boldsymbol{\sigma}\left(\beta^T \bar{x}^{(i)}\right)\right]$$

Equation 11 - Expanded likelihood function

Maximum likelihood estimation involves choosing the appropriate β^T which maximizes the value of the likelihood function in Equation 11. Unfortunately, we cannot find the local maxima or

minima in Equation 10 via differentiation, closed formed solutions do not exist for Equation 11, (Peach, 2016). The alternative is to solve for the local maxima and minima via. iterative numerical methods, particularly, Newton-Raphson. Peach (2016) summarized the iterative numerical process efficiently. The iterative process used is called gradient ascent. Effectively, the process claims that when small continuous steps are taken in the direction of the solution, with intra back and forth movements, then eventually the local maxima or minima will be reached.

Taking the partial derivative of the log likelihood in Equation 11 with respect to each parameter $\beta_{j.}$ We arrive at the following equation:

$$\frac{\partial LL(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \left[y^{(i)} - \sigma\left(\beta^T \bar{x}^{(i)}\right) \right] x_j{}^{(i)}$$

Equation 12 - Partial differential of log likelihood function with respect to β_j

The iterative numerical method chooses an arbitrary β_j, call this β_j^{old} then takes a small sized step, μ , which results in a new estimate, β_j^{new} via the following equation:

$$\beta_j^{new} = \beta_j^{old} + \mu . \frac{\partial LL(\beta^{old})}{\partial \beta_j^{old}}$$

$$= \beta_j^{old} + \mu . \sum_{i=1}^{n} \left[y^{(i)} - \sigma\left(\beta^T \bar{x}^{(i)}\right) \right] x_j^{(i)}$$

Equation 13 - Iterative co-efficient update equation

Where μ is the size of the small step. Iteratively when small incremental steps are taken, with β_j^{new} updating, then the solution will eventually converge to the optimal value of $\boldsymbol{\beta}$, which will be the *maximum likelihood estimator*.

3.2.3 R Statistical Package: Logistic Regression Build Process

The popular open source statistical package, R, was used to build the logistic regression models. The procedural steps and relevant model evaluation statistics used build the regression models will be elaborated.

R is an open source, free software under the terms of the Free Software Foundation's GNU Public License with the ability to execute various common statistical functions, such as regression modelling (ordinary least squares, maximum likelihood estimation), time series models, classification and clustering to name a few.

The stepwise procedural process to building logistic regression models is to (1) size the initial Information Value (IV) statistics for each independent variable and then exclude inappropriate variables based on these initial IV's (2) separating the remaining development data set into training and validation data sets (3) fine classing and then course classing variables using the weights of evidence analysis (4) invoking the stepwise logistic regression procedure with the Akaike Information Criterion (AIC) statistic optimized via stepwise inclusion and exclusion of predictor variables (5) correlation between variables is analysed in this intermediate model, where problematic variables are excluded with step 4 re-run (6) Final models performance ability is then evaluated using *ROC* curve, *AUC*, *KS* and *PSI* statistics.

Measuring the IV statistics per variable is the first step in the logistic regression build. Information value summarizes the power of a feature variable in separating goods and bads in the population. Goods are those observations who do not default over the performance period where as bads are those observations who do default. Table 5 (Appendix B) illustrates the IV associated to the predictive power of a variable. Hand & Henley (1997: 529) suggest features with IV >= 0.1 make for good predictors.

$$IV = \sum_{i=1}^{n} \left(\frac{good(i)}{good(i) + bad(i)} \right) * ln\left(\frac{good(i)}{bad(i)} \right)$$

Equation 14 - Information Value

Like the classification tree model build, the logistic model build used 80% of the total data as training data and remaining 20% as validation data. A key step in the process of building logistic regression models is the classing process, which is the meaningful grouping of categories within a variable. Classing is the process of creating categorical statistically significant splits of a variable by maximizing the Weight of Evidence (WOE) whilst ensuring a monotonic linear relationship is established between the independent and dependent variable.

$$WOE = ln\left(\frac{Percentage\ of\ Goods}{Percentage\ of\ Bads}\right)$$

Equation 15 – Weight of Evidence

Classing is traditionally one of the most important aspects in the logistic regression scorecard building process, as essentially each class receives a score or assignment of value (Zhang et al., 2013: 870). At an observation level, each class score is added together to form the total score assigned to the observation. It is critical to have meaningful monotonic splits of the independent variable relational to the outcome variable for scores to rank order risk. The higher the WOE the greater or more meaningful the monotonic splits of independent variable are in relation to the outcome or dependent variable. Fine classing are the categories created for each variable by maximizing the WOE per variable by tested category splits. The fine classes are produced by R through procedural invocation. Course classing is then the super user (data scientist or modeler) subjective judgement, grouping the automatically created fine classes. Course classing is required to achieve two outcomes, firstly, the resultant course classed categories within the variable must have sufficient volumes whilst also holding the monotonic relationship to the outcome variable, fine classed categories whilst being predictive, might not have enough volumes for modelling and secondly, resultant classes are realistic or meet business sense, when grouped by the super user having specific judgmental heuristic knowledge as compared to a purely statistical view produced by the statistical algorithm. The final course classed variables will then be passed through the GLM procedure, building the linear relationship between the classes and the output variable.

Executing the stepwise GLM procedure in R builds the linear relationship between the categorically grouped independent variables and the binary dependent outcome variable. The GLM procedure regresses the categorically grouped independent variables on the binary

dependent outcome variable. The GLM procedure assigns coefficients via a stepwise iterative process including and removing variables, seeking the maximum AIC value. The AIC evaluates the overall model fit, like the R^2 statistic is used to evaluate the model fit in a simple linear regression model (Glatting et al., 2007: 4285). The AIC is a relative model fit measure. The final model AIC measure is a maximum relative to the other stepwise iterations. The AIC statistic is measured against the Chi-Square distribution with a 5% level of significance. Through this iterative process a model is reached, with certain variables likely removed. Each variable in the intermediate model receives an explanatory coefficient (maximum likelihood estimator) requiring a level of significance greater than 99% to be included.

Once there is a viable model, it is imperative to test for correlation between regressors. The variance inflation factor (VIF) procedure is invoked in R to calculate multicollinearity between regressors. The VIF represents the amount the standard deviation of an estimated coefficient is increased due to correlation with other variables, (O'Brien, 2007). This dissertation model development used a VIF > 5 to indicate a regressor is correlated with another variable.

The proposed model is next performance tested (testing for fit) using the Kolmogorov-Smirnov (K-S) statistic and area under the receiver operating characteristic curve (AUC) measure. The K-S value is a measure of the degree of separation between good and bad distributions, falling between 0 and 1 where 0 indicates no degree of separation between good and bad distributions and 1 indicating perfect separation between good and bads. Generally, the higher the K-S measure the better the model does in separating goods and bads. Model fit measures are required on both the training data and validation data sets. The receiver operating characteristic (ROC) curve is the second diagnostic measure used in evaluating resultant model fit. The 2-dimensional curve plots out the false positive rate (FPR) on the X-axis against the true positive rate on the Y-axis, as a visual representation of the confusion matrix. The less 'confused' the model in minimizing the false positive rate, the steeper the ROC curve is, with the ultimate model will having a vertical ROC curve with zero FPR. Measuring the AUC indicates the FPR associated with the model. The steeper the ROC curve, the greater the AUC becomes. Generally, the models with AUC between 50 and 100 have good ability to minimize the FPR, with AUC = 100 being a

vertical ROC curve (ultimate model). AUC is generated for both training and validation models respectively within the development data.

Once the model has been decided to be viable post going through the detailed model building process, the final step is to translate the model to be user friendly for human interpretation, (Leontyev et al., 2016: 1701). The user-friendly version is the model translated to a scorecard, with the goal to score each new credit applicant. In South Africa scores are usually calibrated to an industry norm which is: At a score of 660 an observation has odds ratio of 1/15. With every 50-point increase from 660 the odds ratio is doubled and vice versa with every 50-point decrease down from 660 the odds ratio is halved. Scorecards are calibrated to this standard in this dissertation.

4. Data and Model Construction

This chapter introduces the underlying data modelled, covering topics relating to the loan data, origination company credit decisioning processes, data cleansing, high level volume and bad rate statistics, reject inference and the division of the data into each respective iterative development data sets, encompassing test, validation and performance data sets. The predictor feature variables are unpacked in detail, defining each variable and elementary statistics relating correlations to the outcome variable.

4.1 Introduction to the Data

This dissertation models disbursed reloan credit data of a prominent South African micro-lender spanning over a 5-year period, from January 2014 to December 2018. A micro-lender is defined as a certified credit provider underwriting unsecured retail loans in mass to approved South African citizens. Reloans are defined as loans disbursed to existing clients, who have a loans history with the micro-lender. Effectively the existing client is applying for a new term loan by settling an existing loan or taking out a new loan after some period being dormant. The approved clients then begin to make monthly annuity based payments, which include interest and service fees and amortizes the principal down to Zero Rands over the agreed term of the loan. Depending on business rules, clients will recursively reloan over a span of time, hence a single client can appear with several loans. This dissertation will focus on loan level data. The credit vetting process of the company in question meets the standards of the South African National Credit Act, as governed by the South African National Credit Regulator (NCR). Credit vetting procedures exclude insolvent applicants, those in financial rehabilitation and clients who are financially overextended with no headroom to afford further credit commitment, (*National Credit Act, No. 34 of 2005*, 2005: 171(1))

A fundamental driver to meaningful parametric and non-parametric credit scoring models is the cleanliness of the underlying model development data, (Khandani et al., 2010: 2772). In this dissertation, the machine learning algorithm will have full power in dictation of the makeup of the model and consequently a fair amount of time was spent cleaning out the noise from the underlying data to generate meaningful models. Cleansing of data included missing data

frequency analysis, where variables with high frequency of missing data were excluded. The key consideration is that classification trees does not require a rigorous model step by step process as compared to the logistic regression approach. Boosting techniques, such as reweighting scarcer and abundant observations were not performed as there was an ample supply of scarce bad observations through the development data sets. When data is unbalanced, in there having a scarce supply of bad observations, model robustness becomes an issue. Synthetic minority over-sampling techniques are used to improve model robustness (Chang et al., 2016: 6083). Manual pruning (subjective) was not performed due to finalized trees not being cumbersomely large, hence unpruned trees were taken verbatim as final models. Cumbersomely large, unpruned trees have undesired outcomes of becoming overtly complex and prone to over fitting, losing robustness in the face of dynamic underlying populations. When resultant decision trees are inconsistent, bootstrapping techniques are used (Chang et al., 2016: 6083) to improve stability and consistency of resultant decision trees, however the decision trees produced in this study were consistent and stable, hence no requirement in applying bagging bootstrapping techniques.

The outcome variable modelled in this dissertation is defined as follows, loans in default on their monthly repayments (greater than or equal to two missed repayments or written-off/ moved off the balance sheet) at five months after loan inception is defined as 'Bad'. All others in the population are classified as 'Good'. The bad rate as defined in Equation 16 is the key outcome metric modelled in this dissertation. Tables 6 and 7 (Appendix B) represent the number of loans disbursed and corresponding bad rate on those loans disbursed, over the five-year review period. The bad rate on this reloan book is within a 2.4% variance band across the five observation years, with a floor of 7.8% and ceiling of 10.2%. This indicates that risk appetite does change, where a lower bad rate indicates a tightening on risk appetite (risk aversion) and higher bad rates indicating a loosening in risk appetite (risk taking). The tightening and loosening of risk strategies tie into the desired profit optics targeted by the company for the observation years under review. Hitting desired profit optics from deployed risk strategies depend heavily on stable operations activity (collections) in meeting certain efficiencies and a certain degree of macro-economic stability (limited economic shocks and related effects on income and employment).

The general flow in bad rate appears seasonal, where bad rates have local peaks for accounts originating in January/ February, then again for accounts in June and again starting to trend upward for accounts from September to December. These local trends don't hold for all years. Seasonal local peaks in bad rates are linked to an increased demand for credit in those months, where credit is generally used to fund certain lifestyle and economic requirements. Increase in festive spending results in an increased demand for credit and the company responds to that increased demand by accepting the higher risk accounts. This is all budgeted for though, and factored into the desired profit calculations.

$$\text{Bad rate} = \frac{(\sum_{i=1}^{n} B_i)}{(\sum_{i=1}^{n} B_i + \sum_{i=1}^{m} G_i)}$$

$B_i \; is \; the \; i_{th}$ bad observation out of n total bad observations
$G_i \; is \; the \; i_{th} \; good$ observation out of m total good observations
Total Population $= \; n + m$

Equation 16 – Bad Rate

Only *approved disbursed* reloans were modelled, with loans rejected ignored. Reject inference will be ignored in both the construction of the logistic regression models and classification tree models. In practice, rejected applications with no resulting performance indicators should be included in business decisioning models because those rejected applications form part of the total population of applications running through the models, however the models created in this dissertation are not optimised for business use, but intended to answer the two underlying research questions, (Anderson, 2019: 349).

The data is split into three distinct time periods each related to three distinct model builds. Model 1 is labelled Iteration 1, Model 2 is labelled Iteration 2 and Model 3 labelled Iteration 3. There are nuances to the data construction with time overlap, depicted in Figure 3 (below). The data development data set grows per iteration by including new data but with the same January 2014 start point. Development data sets are apportioned in ratio 4:1 toward training and validation, 80% of development data used train models and remaining 20% used to validate trained models, in both modelling techniques. The performance data sets are mutually exclusive independent

data sets to the development data sets per iteration. They are used to test the predictive ability of the built models. Each observed performance application has an actual outcome. If models are predictive they will accurately predict the performance of these performance observations. For each iteration, the performance data are observed applications over the twelve months, post the development data.

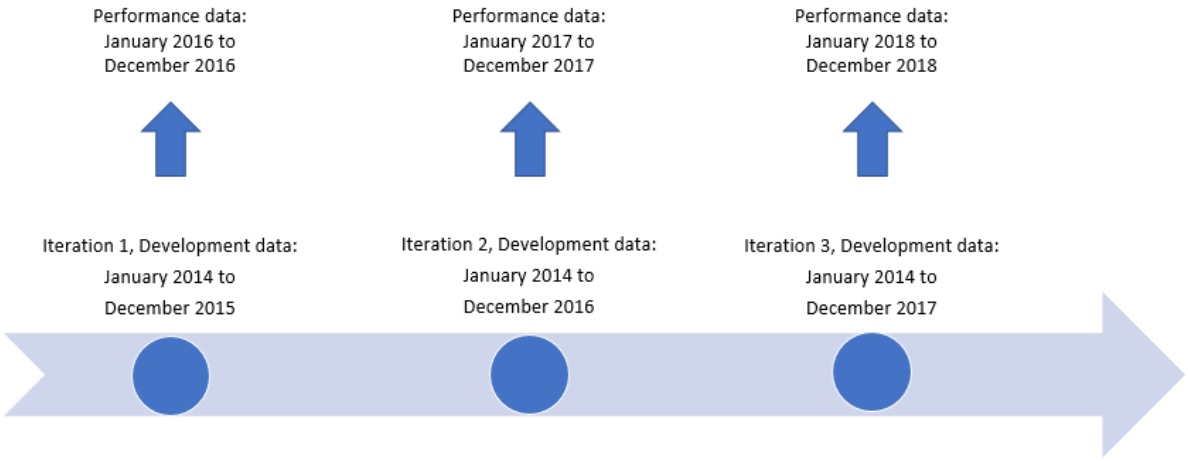

Figure 3 - Data dissection per model build

Setting up the data apportioning assists in answering the key analytical questions to this dissertation, which technique proves to be more predictive when modelling a short-term outcome and which technique proves more robust as development data sets are increased in size. Six models in total were built, three per modelling technique relating to each data iteration. At each iteration, identical data sets were used to build both the logistic regression and classification tree models. In this way, there could be no biases introduced, possibly unduly penalising either technique. Stein (2007: 77) highlights that when development data sets expand for credit models, modelling standard outcome periods, predictiveness and power of models will improve when the ratio of bads to goods increases (i.e. new information becomes available) else

the models incorporating more development observations will simply just be prone to overfitting without any improvement on accuracy.

4.2 Predictor Feature Variables

The fundamental driver of future payment behaviour or credit default risk is revealed by the past payment behaviour of a client. With brand new loan applications to the institution, credit providers seek this insight by consulting and analysing the applicant's payment behaviour at other credit institutions. In the case of an applicant who is brand new to credit (no trace of credit history), demographic data can be used to infer the future payment behaviour. This dissertation looks at predictive credit default risk of existing clients with a payment history within the organization and at other institutions. Once a customer is granted a credit product, internal payment histories weigh in more heavily on the internal credit scoring decisioning process. In general, across the credit industry, reloan credit scorecards using a mix of internal payment behaviour history and third-party payment behaviour have high credit default predictive strength. Historic third-party payment behaviour is encompassed in credit bureau information. Predictor variables are represented at the time of credit application. Post a due diligence process involving parametric testing and subjective industry norms, the universe of possible feature predictor variables was scoped down to thirteen variables, with those selected representing a good blended mix of the variables typically that would go into a reloan decisioning process. Hand & Henley (1997: 528) highlight expert judgement, stepwise statistical procedures and information values as appropriate criteria used in selecting feature predictor variables from a universe of predictor feature variables. A client's level of indebtedness is not included because the purpose of this model build is to specifically evaluate willingness to pay and not ability to pay. The thirteen scoped predictor variables are listed below (at time of application):

1. Client's bureau behavioural score (Bureau Score). As calculated at the Credit Bureau. The bureau score encapsulates the applicant's payment history payment across all the client's credit commitments at various other credit institutions.
2. Client classification. The original channel the applicant had been sourced from, outside the listed group (pure external) or sourced from the sister company (sister company lead).

3. Number of debit order disputes by client on past loans (Dispute Counter). Historic counter of the number of times the applicant has reversed a scheduled debit order at the credit provider when obliged to make a contractual payment.
4. Recency at time of application (Recency). The number of elapsed calendar months since the applicant made a contractual payment.
5. Sister company's internal behavioural score of the client (Sister Behaviour Score). Credit behavioural score on the applicant with a repayment credit history at the sister credit provider.
6. Worst contractual delinquency (Worst Contract Status). Returning the applicants highest ever count of arrears payments recorded for the 12 months preceding the application date.
7. Number of months client has been on book (Month on Book). The cumulative count of the number of months since the applicant had been initially granted credit (coming onto book).
8. Number of months since the client last missed payment (Months Since Last Missed Pmt). The cumulative calendar month counter since the applicant has been making contractual payments since missing their last contractual payment.
9. Number of times the client reloaned in the last 12 months. The cumulative counter the applicant had reloaned (recursively settled existing loans with new loans) for the 12 months preceding the time of application.
10. Concurrently having a short-term facility with a long-term loan at the point of applying or a repeat long-term loan (ST Concurrent)
11. Disbursement by source channel (Source Channel). Details on the procurement channel in concluding the sale of the reloan.
12. Client's age (Age). Applicant's age at the time of application.
13. Client's bank (Bank). Applicant's primary bank account for debit orders

The strongest variable among the thirteen model variables is the client's bureau behavioural score or bureau score, in short hand. Figure 4 (below) displays the monotonic relationship between the bureau score and bad rate, where the bad rate is linearly decreasing in trend accompanied with the bureau score increasing. The higher the bureau score the lower the bad rate and vice versa for a lower bureau score, hence the displayed relationship conforms to

expectations. The bad rate appears flat around the tails of the score distribution, indicating the score losing predictive power around the tails, which unfavourably reduces the bureau score's predictive power in these score value regions. Table 2 (below) shows the bureau score to have a statistically significant correlation with the outcome variable with a p-value of 0.0000 when performing the simple one-factor parametric test for correlation.

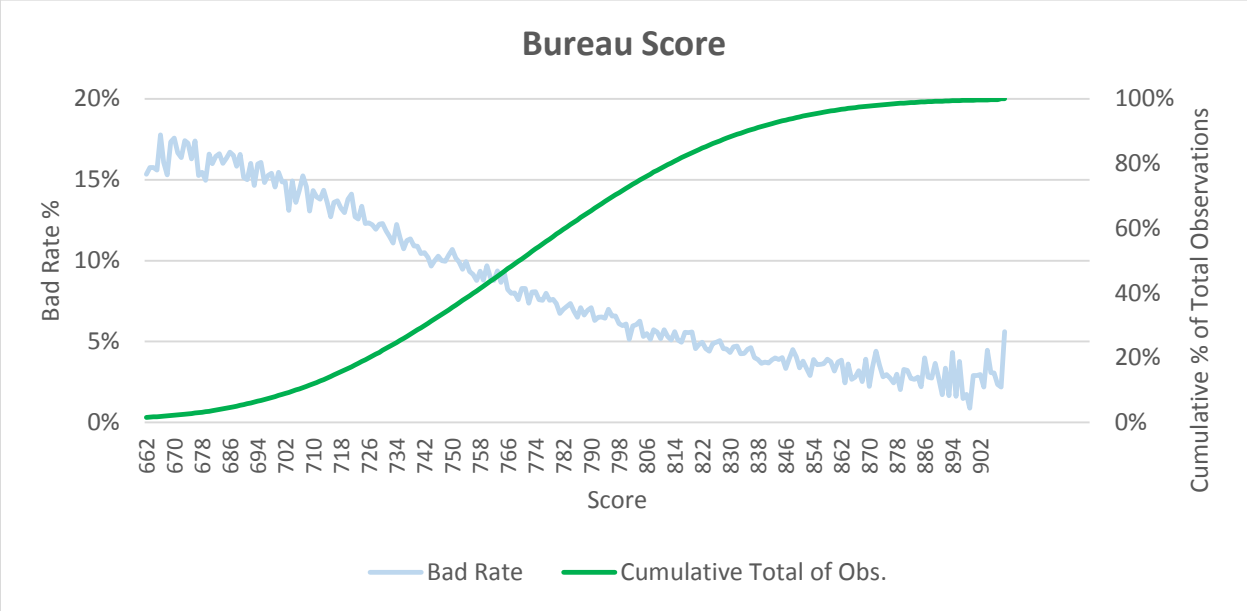

Figure 4 - Bureau Score

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Bureau Score	-0.0004	0.0000	-19.2764	0.0000

Table 2 - Bureau Score Correlation Statistics relative to the outcome variable

4.3 Model Construction

The next section deals with the model construction and the technical aspects associated with each technique under each data iteration. The model build follows the technical model building processes highlighted in Chapter 3. One model will be built per data iteration for both techniques, resulting in six models built. The most efficient way to convey the model construction results per technique per iteration, will be to decompose the model build process into three distinct partitions. The first topic will cover variable strength across the built models, secondly model fit

will be explored, thirdly views of the actual models. The analysis of the resultant models evaluating performance in context to the research questions will be covered in depth in Chapter 5.

Table 8 (Appendix B) displays the ranked power (entropy) of the variables under Classification Tree (CT) across the three data iterations. Bureau score is found to be the most powerful discriminant variable across all data iterations. One weakness in Visual Studio is that it does not have the option to display the entropy measure ranking variable power. Table 9 (Appendix B) displays the ranked power, IV, of the variables under the Logistic Regression(LR) technique with, bureau score being the most powerful variable across all three iterations with IV of 0.23. This finding is in line with the CT findings and the basic non-parametric linear regression tests. Month on Book is the next strongest variable ranking second and third under LR and CT respectively. Interestingly, recency features as preferable variable under CT and less so under LR and vice versa for sister behaviour score. Source channel appears weak to both techniques. Another pattern emerges, LR is consistent in its power ranking across the iterations whereas power ranking jostles under CT, especially from ranked variable 6 onward, with variables even dropping out of model consideration (Sister Behaviour Score under iteration 2). The jostling between variables under CT indicates that variables are matched in power to each other under these iteration. The CT algorithm is allowed to choose it's desired variables and tree design. Table 9 (Appendix B) includes variables with a significant IV (IV>=0.03). Analysis on the fine and course classes produced by the algorithm shows the classes to be monotonic and having significant volumes appropriate for modelling.

The variables are then processed through the respective algorithms per data iteration creating the resultant models. All three resultant models built under CT have a strong fit of the development data, with all having a normalisation score of 94%. Figure 8 (Appendix A) and Table 10 (Appendix B) depicts how the CT models (CTM's) diverge from the ideal population fit, with FPR of 8.56% (CTM1), 8.68% (CTM2) and 8.69% (CTM3). The trend is emerging that under CT the FPR starts to increase as the development data sets increase in volume as the first hint that the classification tree method has started to deteriorate (be it at a marginal 0.03%) with the inclusion of more development data. Table 11 (Appendix B) summarises the three models built under LR,

with some trends emerging. Firstly, the Model AIC increases under LR as the development data set size increases, as opposed to CT model strength measured by the normalisation score remaining flat at 94% as development data set size increasing. Secondly, the significant variables making up the three LR models remains consistent, apart from Months Since Last Missed PMT not being significant in LR1 and Client Classification entering the model for LR3. The coefficients produced for all LR models are in-line with business intuition. Table 12 (Appendix B) shows none of the variables have VIF greater than 5, meaning there is no evidence of correlation among variables in the resultant output models. Figures 9 – 11 (Appendix A) illustrates the K-S statistics across all three LR model's averages 0.26, maximized in the close neighbourhood around the 60th cumulative percentile observation depending on model (consistent for both the development and validation data sets). This reinforces the consistency observed in the LR resultant models. The K-S statistic of 0.26 indicates the model can separate between goods and bads, though still far from the ultimate K-S statistic of 1. The three models average with an AUC of 67.2% and 67.4% respectively between training and test data sets, which means the final model can effectively minimize the FPR generated. The six resultant models fit the development data appropriately and is fit for purpose to use to test the research questions.

With the six models significant and deemed fit for purpose, the actual models are presented. Tables 13 to 15 (Appendix B) represent the three CTM's. For CT, the views are neatly compressed visualizations of the model's raw tree's, with none of the key elements lost. Crucial elements such as cumulative population distributions and bad rates are included, making comparisons easier. CTM1 has 44 legs, CTM2 58 legs and CTM3 83 legs, clearly as more development data is included the number of legs increases with the models becoming more complex. Adding more data over a longer time horizon has worked negatively against to the final CTM's with the increased complexity leading to overfitting, with the technique boosting discrimination within localized segments of the development population. A consequence of high complexity in classification trees models makes future population stability by node an issue (talking to robustness in population stability). Interestingly adding more development data has not increased the model normalisation score (measure of fit) and the number of variables used in the final CTM's remaining relatively static.

Tables 16 to 21 (Appendix B) summarises the final LR calibrated scorecards. Each of the selected variables have categories who are assigned a numeric score. New observations are scored through these models by having a unique category assigned to each variable and the associated score of that variable's category. The individual categorical scores per observation are summed to give its aggregate score. With the inclusion of more development data, the number of included variables expands on the logistic regression model. Like the CTM's the parsimonious principle regarding variables should always be followed, avoiding resultant models being overly complex.

Table 22 (Appendix B) lines up all six models by variable included in each model. Core variables alike to both techniques are Bureau Score, Month on Book, Recency and Worst Contract Status which corresponds sensibly to the data analysis presented in prior chapters. Bureau score is the most powerful discriminant variable. Non-core variables or boost variables are defined to be used for specific legs of the model, applied on specific subsets of the larger population. Use of these variables significantly increases the explanatory power and fit within that localized regional population subset. This ability to use specific variables is the inherent boosting power of the classification tree algorithm. In contrast, all the variables listed in the logistic regression are core variables, each observation in the entire population is scored according to its variable category. This makes logistic regression a broad model with a certain set of variables attempting to explain behaviour across the entire population.

5. Analysis of Results

With the final models now complete, the dissertation moves onto the analysis of the resultant models' performance, with reference to the dissertation's two research questions. Firstly, the weighted sum of squared estimation error (SSE) metric for the two characteristics, population distribution stability and bad rate stability is introduced. Once the SSE measure is introduced, the population distribution stability SSE (PDSSEE) and bad rate stability (BRSSSE) are analysed with respect to the six built models, which leads into the deeper analytical findings related to the two research questions of this dissertation.

5.1 Weighted Sums of Squared Estimation Error (SSE)

Key to answering the dissertation research questions is the introduction of SSE measure. Mathematically, the measure is defined as follows:

$$Weighted\ Sums\ of\ Sqaured\ Estimation\ Error\ (SSE) = \sum_{i=1}^{n} (\widehat{y_i} - y_i)^2 / n_i$$

Equation 10 - Weighted Sums of Squared Estimation Error

Where $y_i = Development\ data\ set\ observation$

Where $\widehat{y_i} = Perfomance\ data\ set\ observation$

Where $n_i = Development\ data\ set\ population\ observation$

The SSE measure encompasses inherent strength in a model by quantifying the model's ability to stand up to expectation when applied to an independent, mutually exclusive, out of time performance data, simulating the predictiveness if the models were used to execute the credit decisioning. Figure 3 displays the development and performance time ranges covering the six models. Development and performance data sets per iteration were identically presented to each modelling technique, in effect giving both techniques the exact same playing field with no biases. Two SSE measures have been constructed, first, measurement of the shift in population distribution between development and performance data sets (PDSSEE) and the second,

measurement of the shift in expected bad rate per segment between development and performance data sets (BRSSSE). Both measures are equally important in credit scoring models, when population distributions and, or expected bad rates begin to shift from those observed in the model development data compared to that observed in performance data, these shifts can have negative effects on the business profitability. In effect rendering a scorecard less precise in its ability to accurately predict credit risk.

Figures 12 to 23 (Appendix A) depict how well the CT and LR fits on the respective performance datasets, with respect to population distribution and bad rate expectation. For the ideal model fit (100% fit and predictability) onto the performance data, the point differences between development and performance at each leg (CT) or score band (LR) will need to be 0%. The development models have population distributions and bad rate per leg (CT) or score (LR). If the model is working well, that population distribution and bad rate expectation will hold true for that leg or score band, within a certain tolerance on an independent performance data set. Realistically a 100% fit will never materialize, hence a model which minimizes these point differences at each model leg or score will result in an overall good fit. Table 23 (Appendix B) summarises the PDSSEE and BRSSSE across the six built models or two models per technique for three iterations.

5.2 Analytical Review of the Research Questions

Which technique proves to be more predictive, when modelling a short outcome period event? The primary tool to answering this question lies in the analysis of the BRSSSE between the two techniques spanning over the six models. A lower BRSSSE indicates a model is doing well to minimise the weighted sums of squares error deviance on the out of time independent and mutually exclusive performance data set when compared to its development model, meaning the model has good predictability. Table 24 (Appendix B) summarises the BRSSSE for each model. The power of the BRSSSE measure lies in its comparability strength, when lining up both techniques side by side for comparison. This is further boosted by the way in which the development and performance data was constructed to be identically available to both techniques. These individual models may not have the appropriate power as an acceptable

model, fit for purpose to making actual business credit risk decisioning, however that is separate topic not explored in this dissertation.

Both techniques have produced excellent predictability for iteration 1, with low BRSSSE statistics, when we consider the BRSSSE of 0.051% for LR1 and BRSSSE of 0.081% from CTM1. In the comparison between techniques, LR came in first position relative to the classification trees models, with consistently lower BRSSSE's overall iterations. The LR technique has a better ability to coming closer to the expected model bad rates than CT. In fact, both techniques suffered on the iteration 2, with the BRSSSE worsening, though improving when moving to iteration 3. In terms of better predictiveness on out of time independent performance data, logistic regression trumps over the classification tree technique, when modelling a short outcome period event. Having multiple iterations further assists solidify that LR was consistently more predictive than CT.

Which technique proves to be more robust and stable in population and bad rate prediction, as development data sets expand over time, when modelling a short outcome period event? Table 24 (Appendix B) is an expansion of Table 23 (Appendix B), assisting in depicting which technique proves more robust and stable, as development data sets expand over time. As mentioned in the last section, adding more data from iteration 1 to iteration 2 worsens both the BRSSSE's on LR and CT's, although improves BRSSSE's when moving from iteration 2 to iteration 3.

The average change in BRSSSE's between iterations is -0.009% for LR in comparison to that for 0.083% for CT. The better technique can keep the average change in for BRSSSE as close to 0% as possible (analysing Equation 10), which LR does better at minimising between the two techniques. As more data is added to model development, the LR technique proves more robust for BRSSSE stability.

The PDSSEE assists in analysing how robust or sensitive each technique is to population distribution movements, particularly when models are applied to independent and mutually exclusive performance data sets and when development populations expand. The analytical results are a stark contrast between techniques. As more data is included in LR model development, the PDSSEE becomes less deviant. LR invites more data for modelling purposes. On

the other hand, CT PDSSEE's increases as more data becomes available for modelling purposes. Table 24 (Appendix B) illustrates this point. With more data in the modelling, each CTM increased its number of legs (CTM1 with 45 legs, CTM2 with 59 legs and CTM3 with 85 legs). This makes intuitive sense, however, the increasing PDSSEE indicates that although the increased number of legs is a result of more development data, this resulted in the model almost being overfitted to the data and subsequent testing on the out of time performance data worsening the PDSSEE indicating that indeed the increased number of legs did lead to over fitment. On population stability, LR proves more robust to population distribution changes when more data is added on for development purposes all whilst when modelling a short outcome period.

6. Conclusion and Future Research

Literature has shown that on average, logistic regression proves to be a stronger credit risk modelling technique with respect to predictiveness and robustness over the classification tree algorithm, particularly when standard default outcome periods are modelled (refer to Chapter 2 for listing of comparative studies). Classification trees however have been shown to be predictive when modelling a shorter outcome period (Chang et al., 2016: 6805). This dissertation aims to prove if the classification tree technique could match or better the logistic regression technique in terms of predictiveness and robustness founded on credit risk data pertaining to a well-established South African micro lender over the period January 2016 to December 2018, when modelling a shorter credit default outcome periods.

The results were conclusive, with logistic regression able to be display better predictiveness and stability robustness to changes in population distribution, as development data sets expand, over the classification tree technique. Both techniques demonstrated strong predictive power, however the logistic regression models display more stable predictive strength over the classification tree models when fitted on independent, mutually exclusive out of time performance data. The optimal predictive bad rate fit for the classification tree technique was achieved only when one year of development data was used in model development (for this dissertation, the minimum development period was one year worth of loans data). Models built with one year development data under logistic regression however still had a better predictive fit over the classification tree technique. This finding corresponds with Stein (2007: 77) who highlights that when development data sets expand for credit models, modelling standard outcome periods, predictiveness and power of models will improve when the ratio of bads to goods increases on the newly incorporated development data else these newly formed models will simply just be prone to overfitting without any improvement on accuracy. The ratio of bads to goods remains relatively static when the development data sets expand (Iteration 1 to Iteration 3) in this dissertation, resulting the over fit results observed. The logistic regression technique responds better to having increased data added for modelling on both predictive fit and population stability. In fact, classification trees population stability degraded with increased data, highlighting the proneness to overfitting.

The trap into which classification trees fall is that of overfitting a model on development data, especially when development data sets expand. The issue is that the number of legs and specificity of the fitted legs on the development data does not hold true when time moves on. Significant branch splits in the tree do not hold true over subsequent snapshots of the underlying population making the branch splits less stable (with respect to population stability and bad rate prediction fit) relative to the logistic regression technique. The logistic regression technique, although possibly criticized for being too general a fit to the entire development population, copes better and is more robust on both predictive fit and population stability as time moves on.

Future research related to this study is to answer the same questions posed, with smaller development data sets time spans (for example, three month, six month, nine month development data sets) and comparing against similar out of time mutually exclusive performance period data. The primary point of doing this test would be to see if classification trees perform better relative to the logistic regression technique on stable loans data over shorter development time spans, accompanied with the shorter credit default outcome period.

In business, the reality is that such credit decisioning tools, such as credit scoring models are required to go through strict conformance requirements (hierarchies of approval through various committees). A simpler tool such as a logistic regression scorecard which demonstrates to be more stable and robust on predictive fit and population stability with a proven track record and being successfully used in the credit industry, will trump the classification tree model. Though, all this holds true in a consistent and stable industry with relation to the business specific risk appetite. Classification trees could have a different application in a shorter more fluid type of modelling scenario, such as modelling fraud outcomes, where fraudster strategies change often and models need to be dynamic to these nuanced changes in population. Different variables enter the equation where the static logistic regression will not pick up these nuances. This is a separate topic for future research, performing the same analysis as in this study with different

underlying data (for example, volatile fraud data and not stable credit risk data) to see how the techniques stack up against each other.

7. References

Altman, E. & Saunders, A. 1998. Credit risk measurement: Developments over the last 20 years. *Journal of Banking & Finance.* 21(11-12): 1721-1742.

https://www-sciencedirect-com.ezproxy.uct.ac.za/science/article/pii/S0378426697000368.

https://doi.org/10.1016/S0378-4266(97)00036-8.

Anderson, B. 2019. Using Bayesian networks to perform reject inference. *Expert Systems with Applications, Science Direct*. 137: 349-356.
https://doi.org/10.1016/j.eswa.2019.07.011.

Baesons, T., Van Gestel, S., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J. 2003. Benchmarking state-of-the-art classification algorithms for Credit Scoring. *The Journal of the Operational Research Society.* 54(6): 627-635.
DOI: https://www.jstor.org/stable/4101754?seq=1.

Blanco, A., Pino-Mejias R., Lara J., Rayo S. 2013. Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. *Expert Systems with Applications.* 40(1): 356-364.
www.elsevier.com/locate/eswa.
DOI: 10.1016/j.eswa.2012.07.051.

Breiman, L., Friedman, J., Olshen, R., Stone, C. 1984. *Classification and Regression Trees*. Pacific Grove, California. Wadsworth and Brooks Cole Advanced Books and Software.

Chang, Y-C., Chang, K-H., Chu, H-H., Tong, L-I. 2016. Establishing decision tree-based short-term default credit risk assessment models. *Communications in Statistics – Theory and Methods.* 45(23): 6803-6815.

https://doi.org/10.1080/03610926.2014.968730.

Charitou, A., Neophytou, E. & Charalambous, C. 2004. Predicting corporate failure: empirical evidence for the UK. *European Accounting Review.* 13 (3): 465-497.

https://doi.org/10.1080/0963818042000216811.

Churchill C. 2000. Banking on Customer Loyalty*. Journal of Microfinance.* 2(2): 1-21.

Crook, J., Elderman, D., Thomas, L. 2007. Recent developments in consumer credit risk assessment. *European Journal of Operational Research.* 183(3): 1447-1465.

Davis, R., Edelman, D. & Gammerman, A. 1992. Machine learning algorithms for credit-card applications. *IMA Journal of Mathematics Applied in Business & Industry.* 4(1): 43-52.

Dong, G. & Lai, K., Yen, J. 2012. Credit scorecard based on logistic regression with random coefficients. *Procedia Computer Science.* 1(1): 2463 – 2468.

Ertel, W. 2017. *Introduction to Artificial Intelligence (2nd ed.).* Weingarten, Germany. Springer International Publishing AG.

Glatting, G., Kletting, P., Reske, S., Hohl, K., Ring, C. 2007. Choosing the optimal fit function: Comparison of the Akaike information criterion and the F-test. *American Association of Physicists in Medicine.* 34 (11): 4285-4292.
DOI: 10.1118/1.2794176.

Galindo, J. & Tamayo, P. 2000. Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics. Kluwer Academic Publishers.* 15: 107-143.

Gurny, P. & Gurny, M. 2013. Comparison of credit scoring models on probability of default estimation for US Banks. *Prague Economic Papers.* 22(2): 163 – 181.

DOI: 10.18267/j.pep.446.

Hand, D. & Henley, W. 1997. Statistical classification methods in consumer credit scoring: A Review. *Journal of the Royal Statistical Society*. 160(3): 523-541.

Horng, I-H., Tsung-Pei, L. & Tian-Shyug, L. 2010. Data Mining in Building Behavioral Scoring Models. *2010 International Conference on Computational Intelligence and Software Engineering.* 1(1): 1-4.

Huang, C., Chen, M. & Wang, C. 2007. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications.* 33(4): 847-856.
www.elsevier.com.
https://doi.org/10.1016/j.eswa.2006.07.007.

Hui, L., Sun, J. & Wu, J. 2010. Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. *Expert Systems with Applications, Elsevier.* 37(8): 5895-5904.

Ibtissem B. 2014. A credit scoring model for microfinance bank based on Fuzzy Classifier optimized by a differential evolution algorithm. *The UIP Journal of Financial Risk Management*. 11(2): 7-24.

Khandani, A., Kim A. & Lo, A. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*. 34: 2767-2787.
www.elsevier.com/locate/jbf.
DOI: 10.1016/j.jbankfin.2010.06.001.

Kocenda, E. & Vojtek, M. 2009. Default predictors and credit scoring models for Retail banking. *CESifo Working Paper No.2862, Category 12: Empirical and Theoretical Methods.*
https://www.cesifo.org/DocDL/cesifo1_wp2862.pdf.

Leontyev, S., Legare J-F., Borger, M., Buth, K., Funkat, A., Gerhard J., Mohr, F. 2016. Creation of a scorecard to predict in-hospital death in patients undergoing operation for acute Type A dissection. *The Annals of thoracic surgery, Elsevier BV*. 101(5): 1700-1706.
DOI: 10.1016/j.athoracsur.2015.10.038.

Makowski, P. 1985. Credit scoring branches out. *Credit World.* 74(2): 30-37.

Mester, L. 1997. What's the point of scoring?. *Business Review (Federal Reserve Bank of Philadelphia. Academic Search Premier.* 1-29.

https://oheg.org/ty_s_tybeh_ti.pdf.

Microsoft. 2020. *Microsoft Decision Trees Algorithm Technical Reference, January 2020.* Available: https://docs.microsoft.com/en-us/analysis-services/data-mining/microsoft-decision-trees-algorithm-technical-reference?view=asallproducts-allversions.

National Credit Act, No. 34 of 2005. 2005. *Government gazette*. 202(37882). 15 March 2015. Government notice no. 10242. Cape Town: Government Printer.

https://www.creditombud.org.za/wp-content/uploads/2015/11/NATIONAL-CREDIT-REGULATIONS-INCLUDING-AFFORDALITY.pdf.

National Credit Act, No. 34 of 2005. 2005. *Government gazette*. 489(230). 15 March 2006. Government notice no. 286619. Cape Town: Government Printer.

https://www.justice.gov.za/mc/vnbp/act2005-034.pdf.

O'Brien, R. 2007. A caution regarding rules of thumb for variance inflation factors. *Springer Science and Business Media LLC Quality & quantity.* 41(5): 673-690.

DOI: 10.1007/s11135-006-9018-6.

Peng, C., Lee, K., Ingersoll, G. 2002. An introduction to Logistic regression analysis and reporting. *The Journal of Educational Research, Taylor & Francis Ltd*. 96(1): 3-14.

Ruonan, L. 2013. The application and assessment of consumer credit scoring models in measuring consumer loan issuing risk. M.Comm dissertation. Stony Brook University.

Sandro, S. 2018. *Introduction to deep learning, From logical calculus to artificial intelligence.* University of Zagreb, Zagreb, Croatia.

Shannon, C. & Holden, W. 1976. *Mathematische Grundlagen der Informationstheorie.* Oldenbourg Verlag, Germany.

Stein R. 2007. Benchmarking default prediction models: pitfalls and remedies in model validation. *Journal of Risk Model Validation*. 1(1): 77-113.

StatsSA. 2020. *Statistical Release. P0441. Gross Domestic Product. Fourth Quarter 2019*. Pretoria. Available: http://www.statssa.gov.za/publications/P0441/P04414thQuarter2019.pdf [2020, September 03].

Thomas L. 2000. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*. 16(1): 149-172.

Tam, K. & Kiang, M. 1992. Managerial applications of neural networks: The case of bank failure predictions. *Management Science*. 38(7): 926-947.
DOI: https://www.jstor.org/stable/2632376?seq=1.

TransUnion. 2020. *TransUnion Industry Insights Report. Quarterly Overview of Consumer Credit Trends Released by TransUnion Financial Services. Fourth Quarter 2019*. Johannesburg. South Africa.

Available: https://www.transunion.co.za/resources/transunion-za/doc/campaign/IIR/INT-AF-19-IIR-Q4-2019/TU-IIR-Report-Q4-2019.pdf [2020, September 03].

Ong, C., Haung, J. & Tzeng, G. 2005. Building credit scoring models using genetic programming. *Expert Systems with Applications.* 29: 41-47.

https://www.journals.elsevier.com/expert-systems-with-applications.

DOI: 10.1016/j.eswa.2005.01.003.

Peach, C. 2016. Logistic Regression [Lecture notes]. Stanford University. California.

Available: https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/pdfs/40%20LogisticRegression.pdf.

Accessed: 27thMay2020.

Date of publication: 20thMay2016.

Vojtek, M. & Kocenda, E. 2006. Credit Scoring Methods. *Finance a uver – Czech Journal of Economics and Finance.* 56 (3-4): 152-167.

West, D. 2000. Neural network credit scoring models. *Computers and Operations Research*. 27: 1131-1152.

Zhang, D., Agterberg, F., Cheng, Q., Zuo, R. 2013. A comparison of modified fuzzy weights of evidence, fuzzy weights of Evidence and logistic regression for mapping mineral prospectivity. *Math Geosci*. 46: 869 – 885.

DOI: 10.1007/s11004-013-9496-8.

Appendix A - Figures and Charts

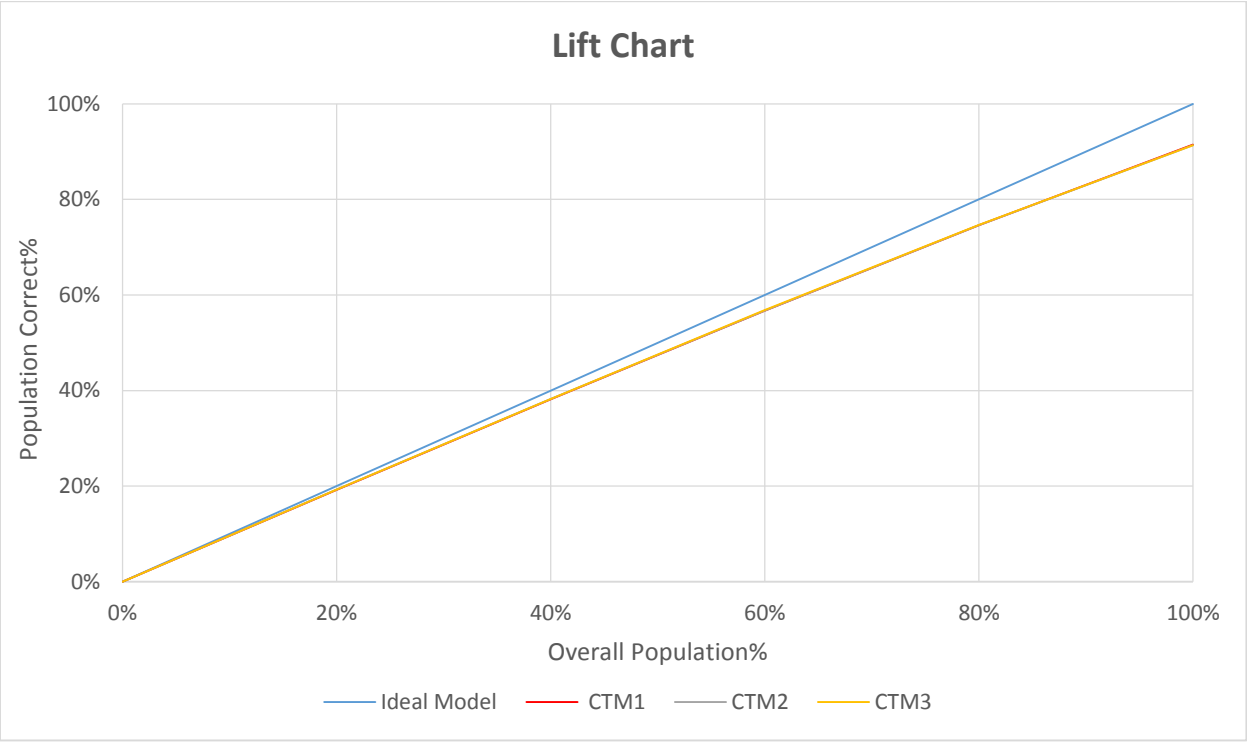

Figure 5 - Lift Chart

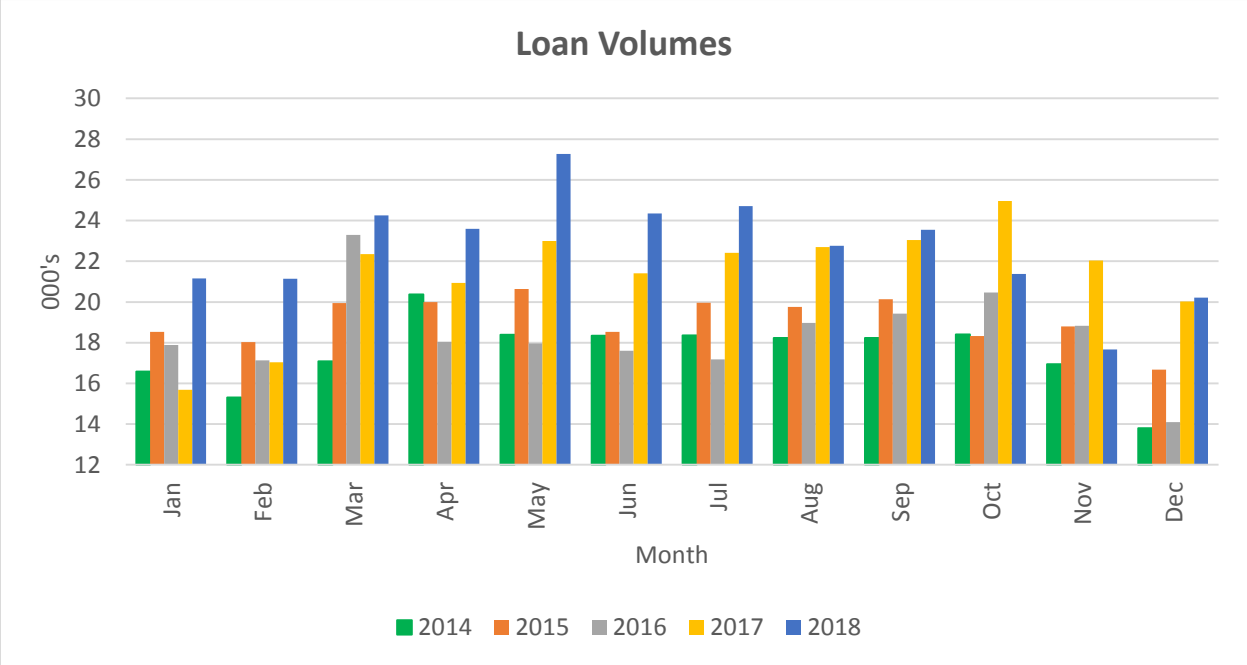

Figure 6 - Loans volumes disbursed by month

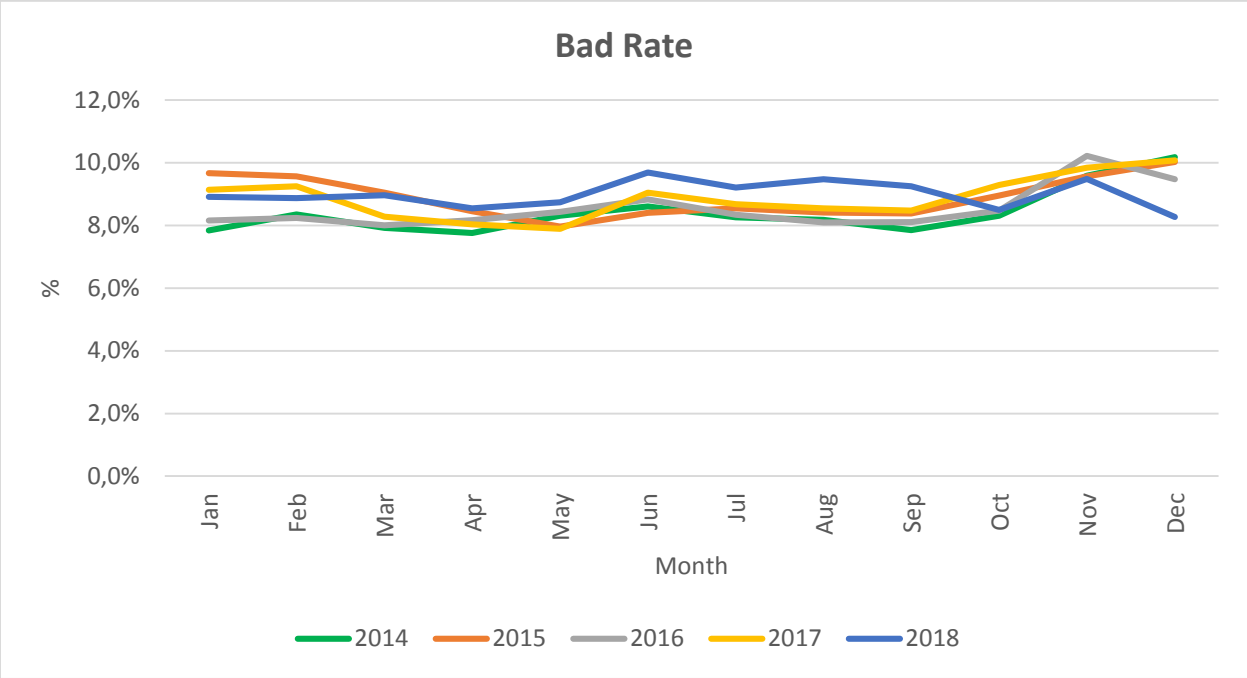

Figure 7 - Bad rate per disbursed month

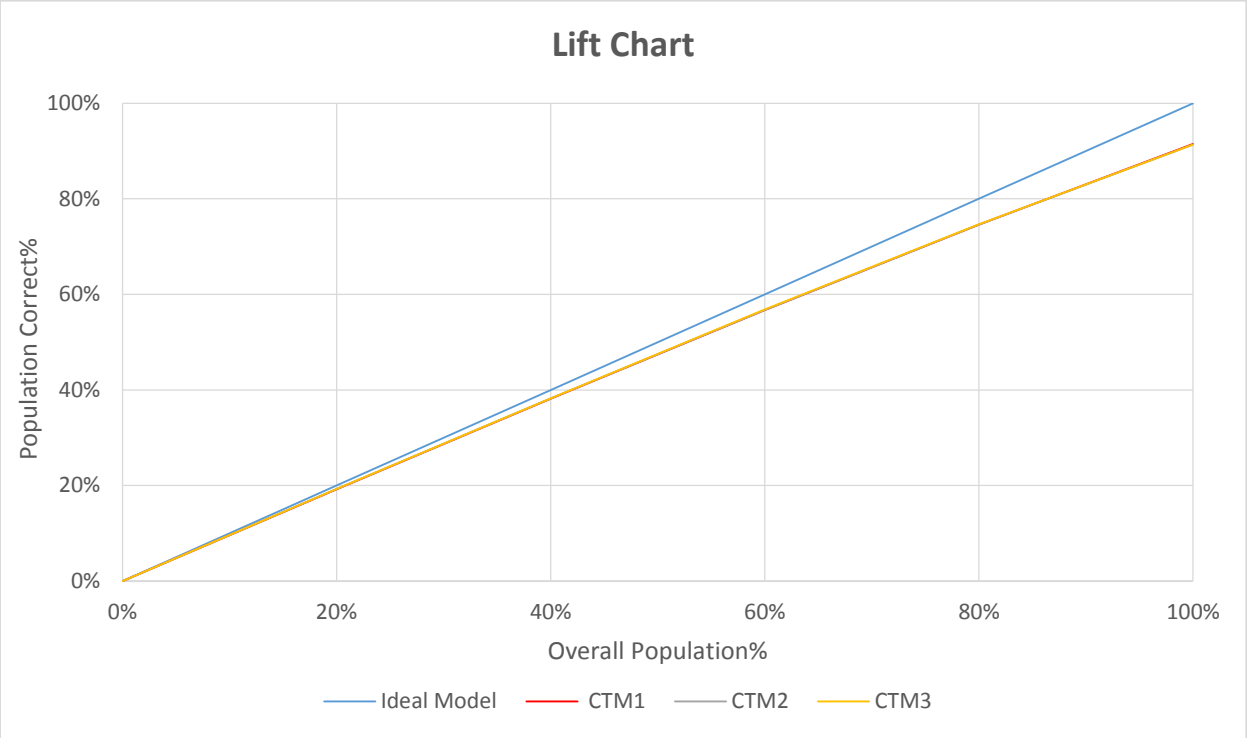

Figure 8 - CTM Lift Chart

Logistic Regression Iteration 1 AUC and ROC Curve

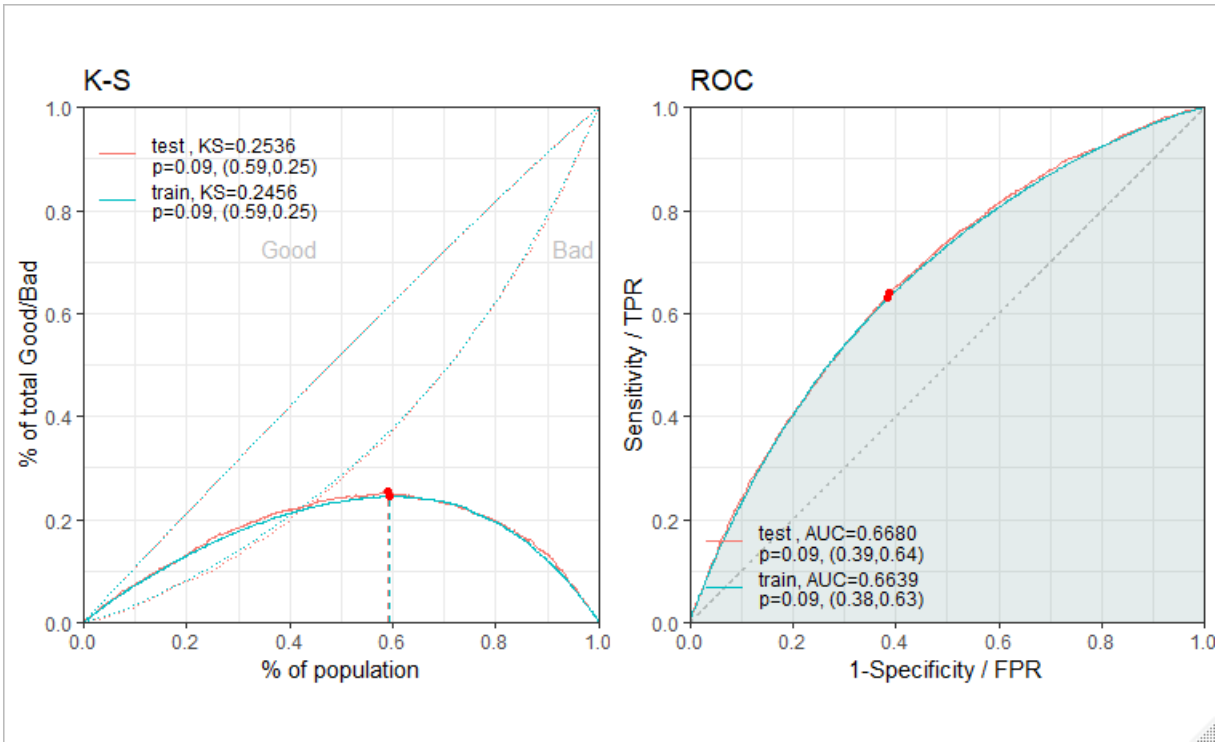

Figure 9 - LR1 ROC Curve and AUC

Logistic Regression Iteration 2 AUC and ROC Curve

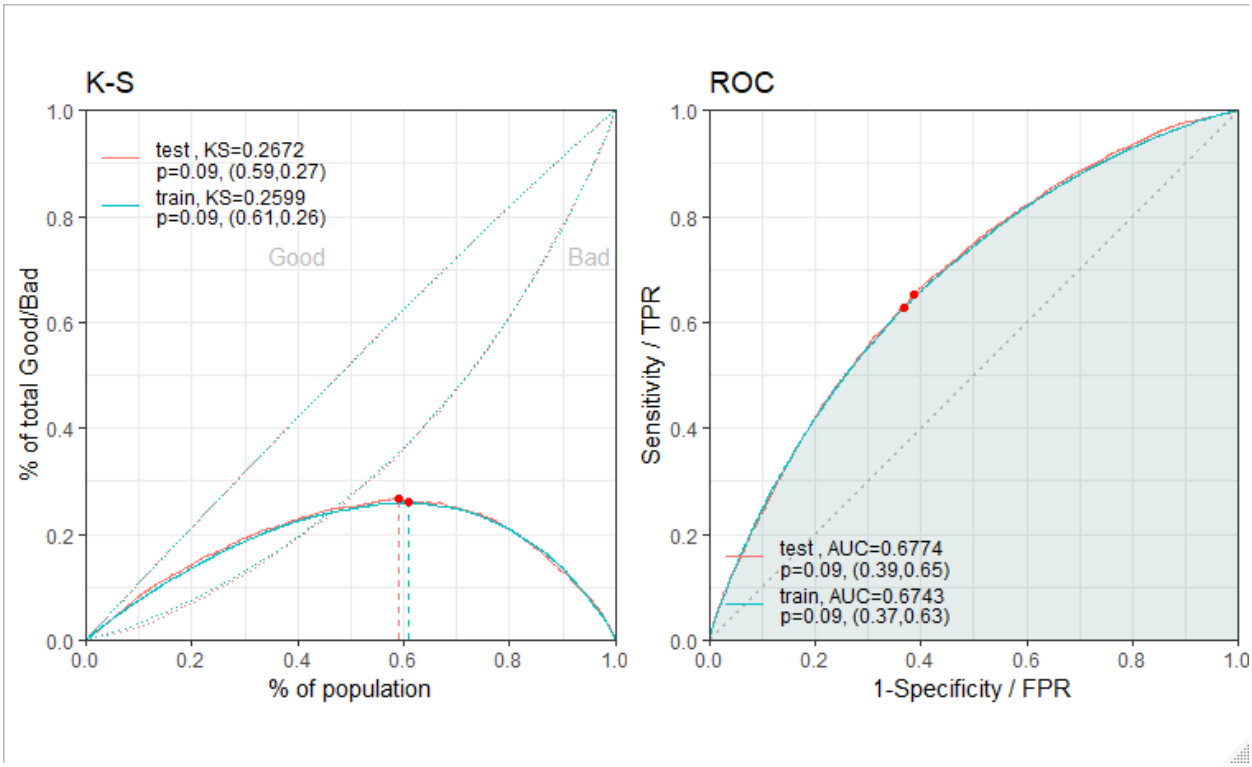

Figure 10 - LR2 ROC Curve and AUC

Logistic Regression Iteration 3 AUC and ROC Curve

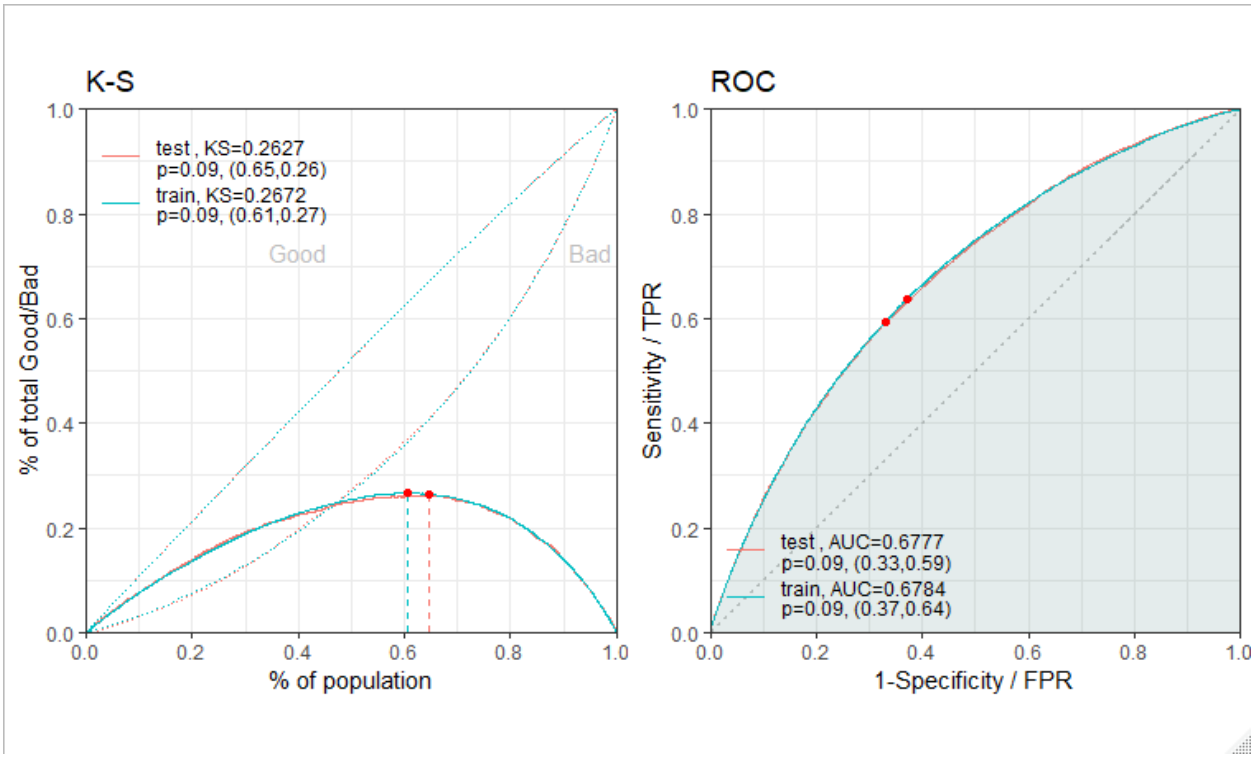

Figure 11 - LR3 ROC Curve and AUC

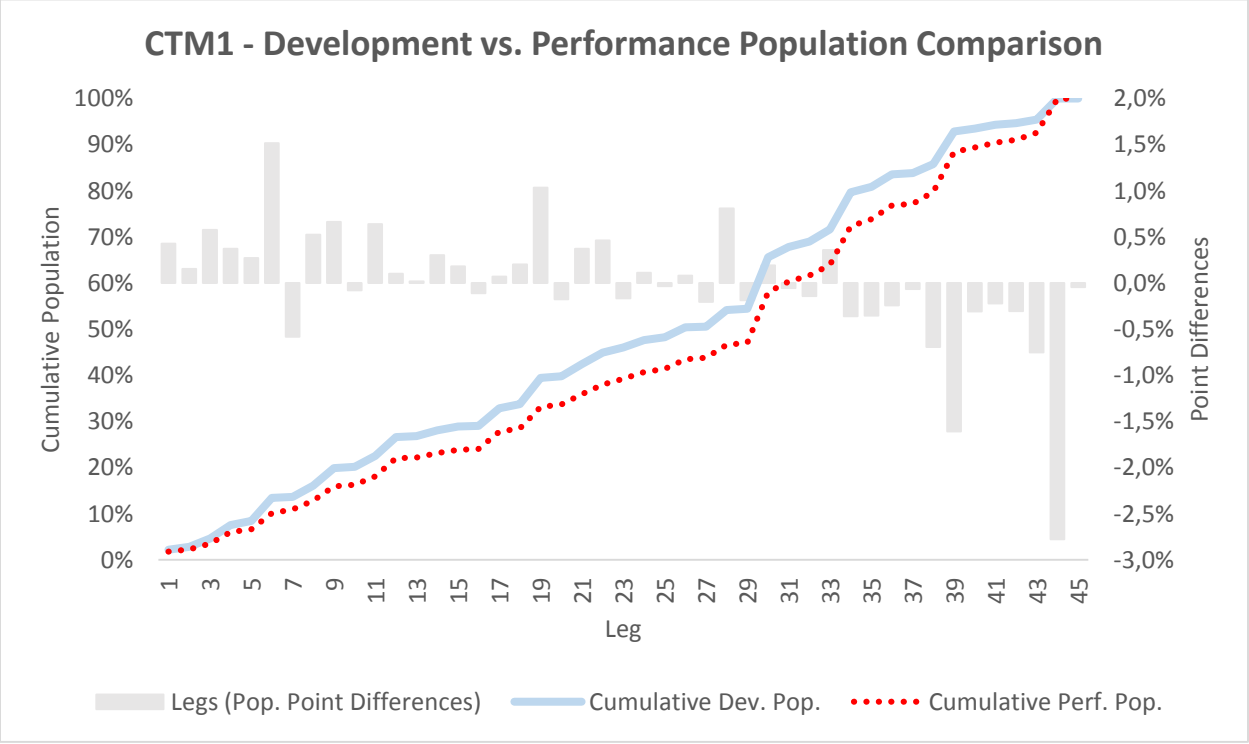

Figure 12 - CTM1 - Development vs. Performance population comparison

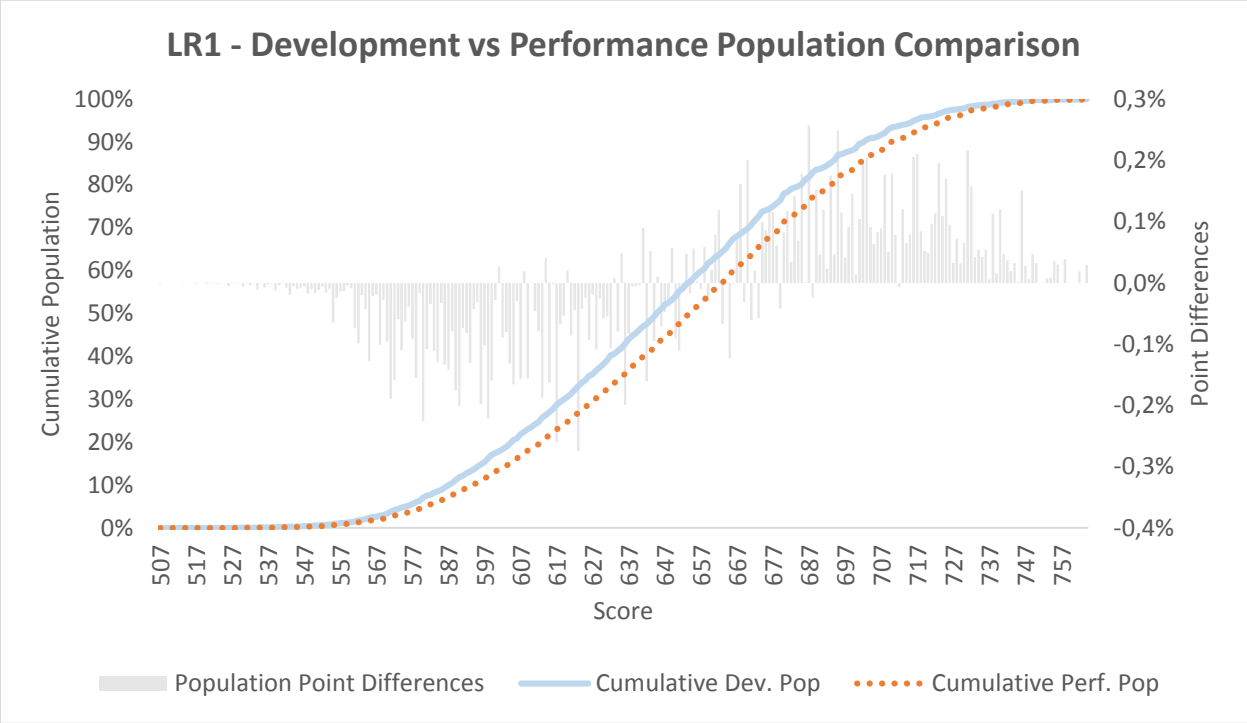

Figure 13 - LR1 - Development. vs. Performance population comparison

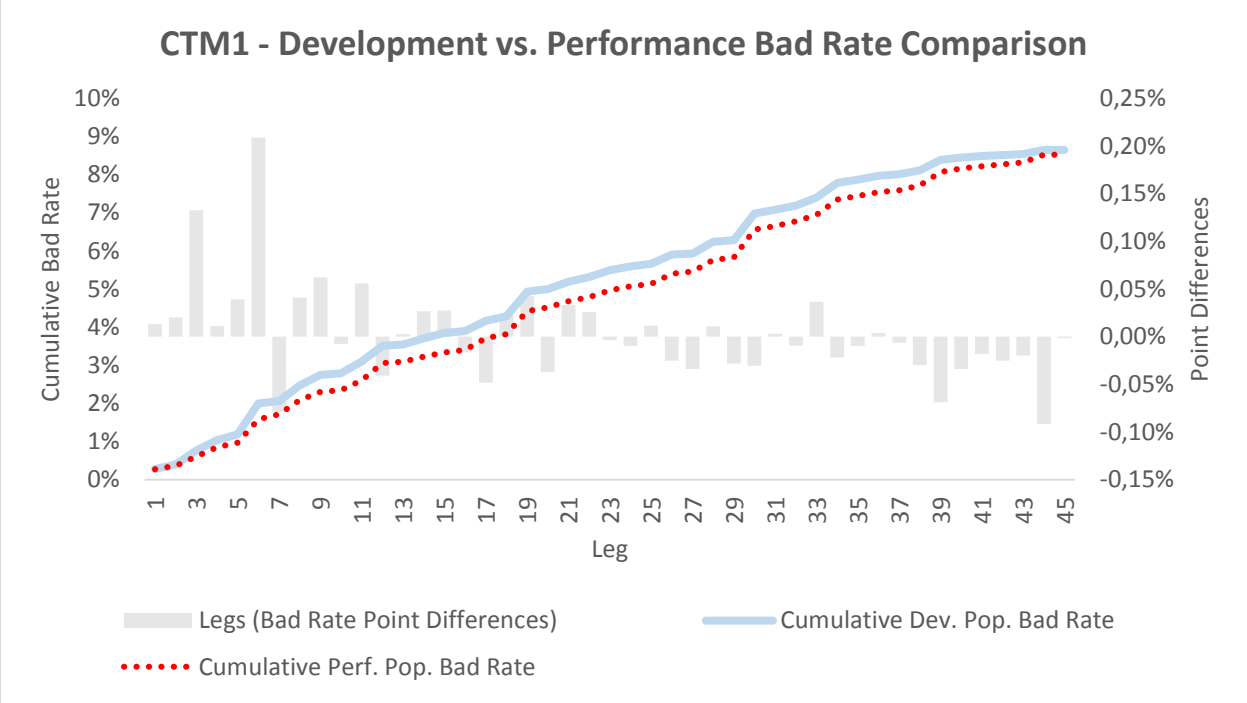

Figure 14 - CTM1 - Development. vs. Performance bad rate comparison

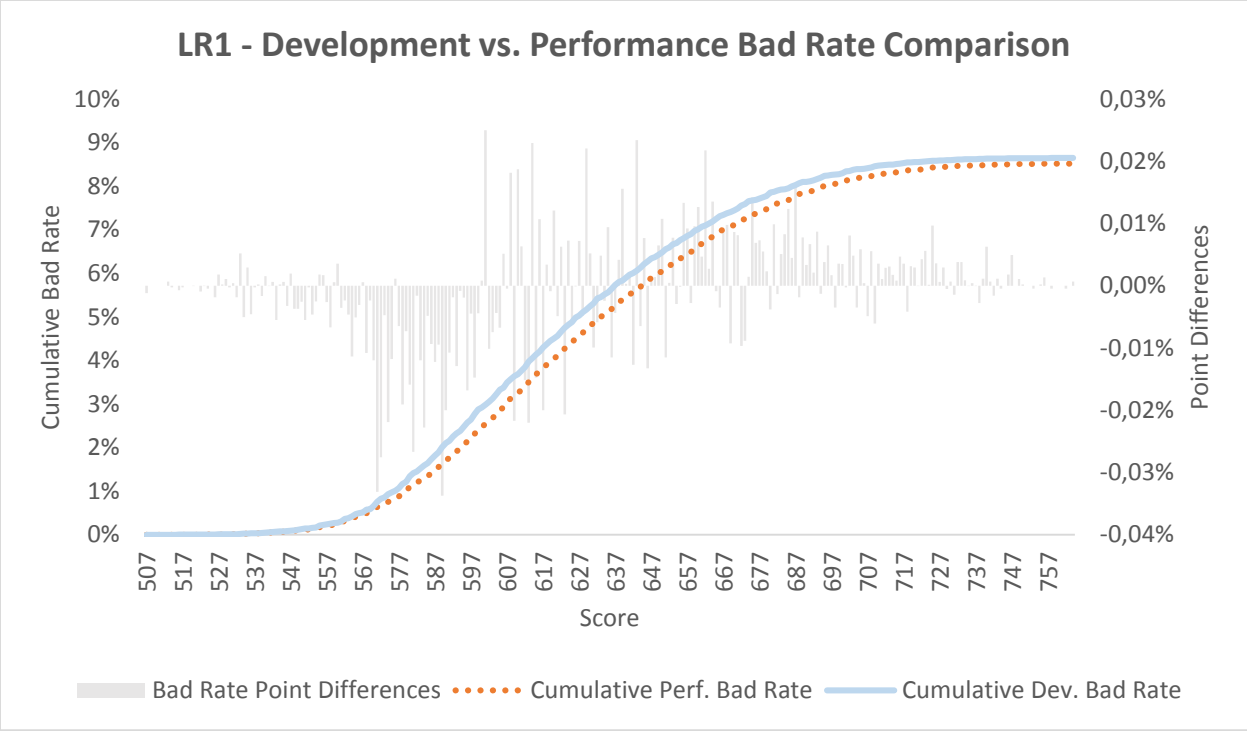

Figure 15 - LR1 - Development. vs. Performance bad rate comparison

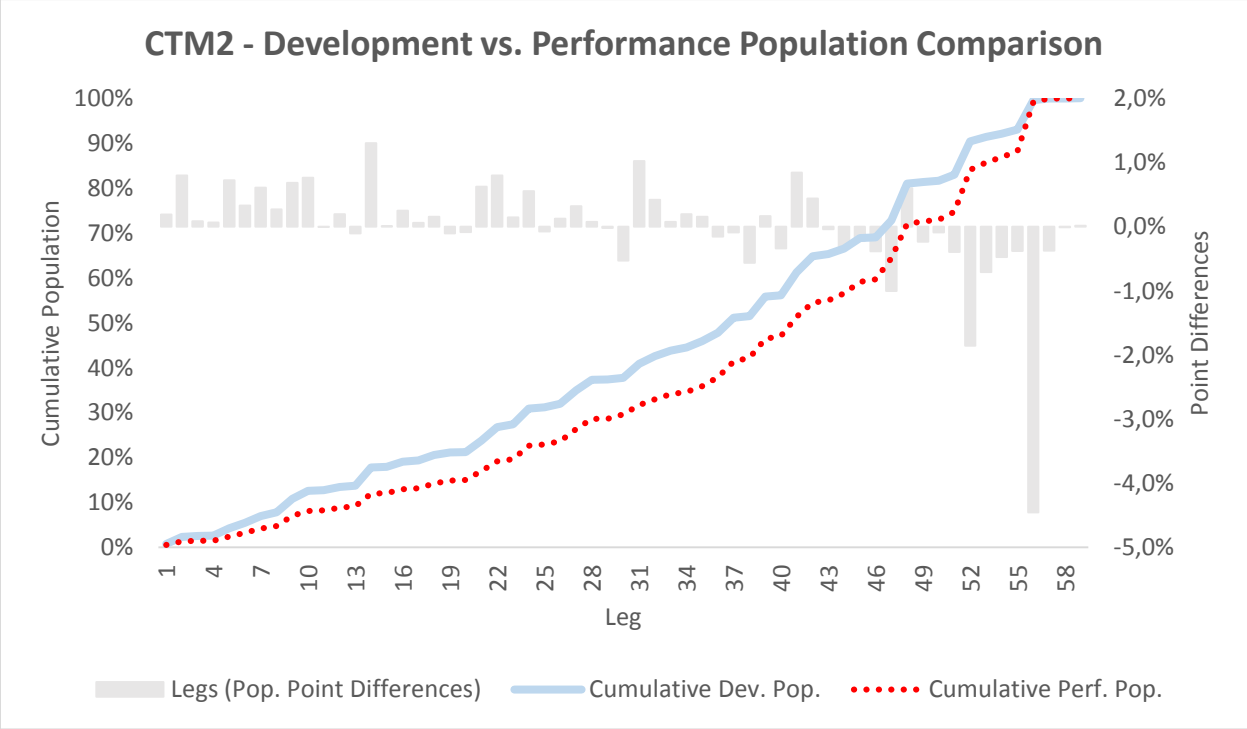

Figure 16 - CTM2 - Development vs. Performance population comparison

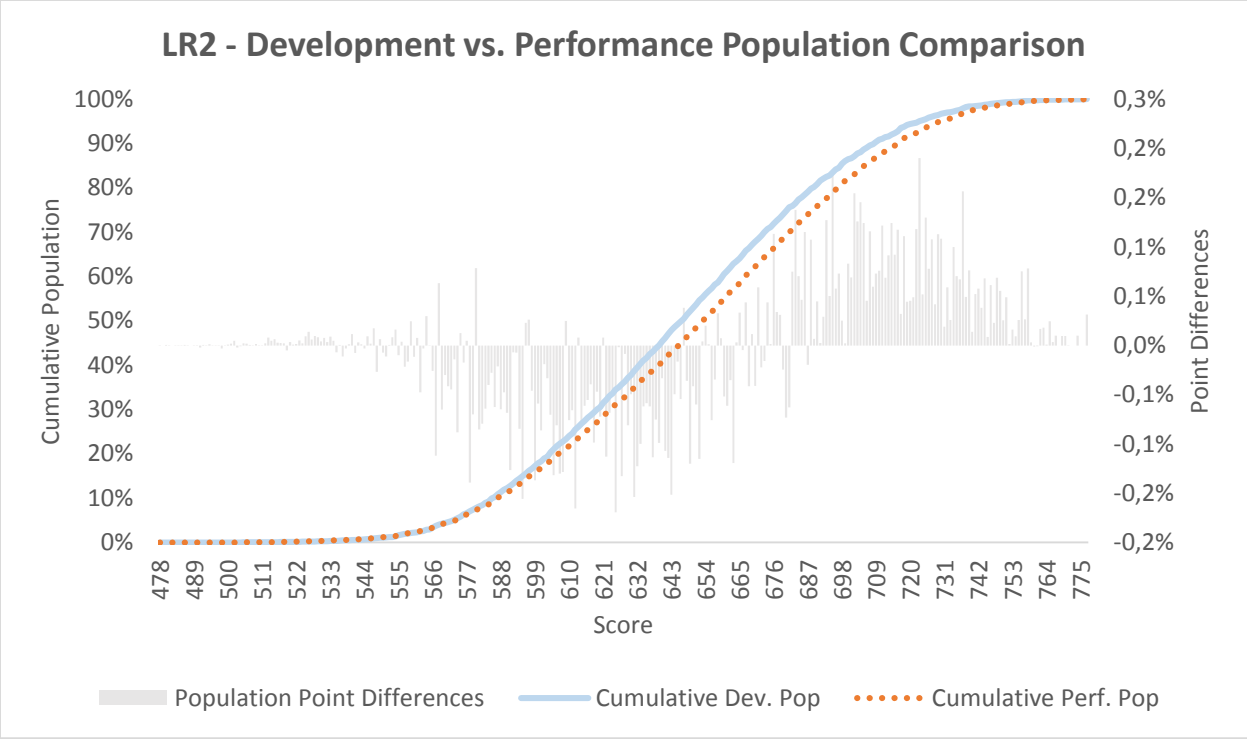

Figure 17 - LR2 - Development vs. Performance population comparison

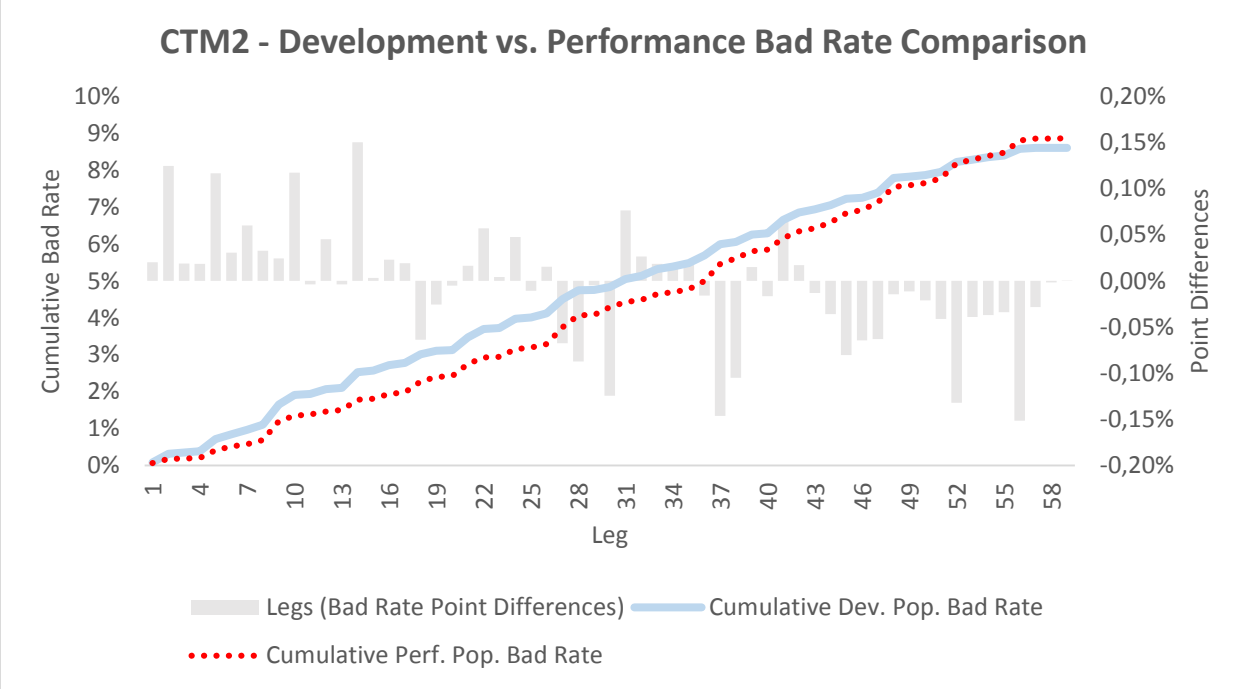

Figure 18 - CTM2 - Development vs. Performance bad rate comparison

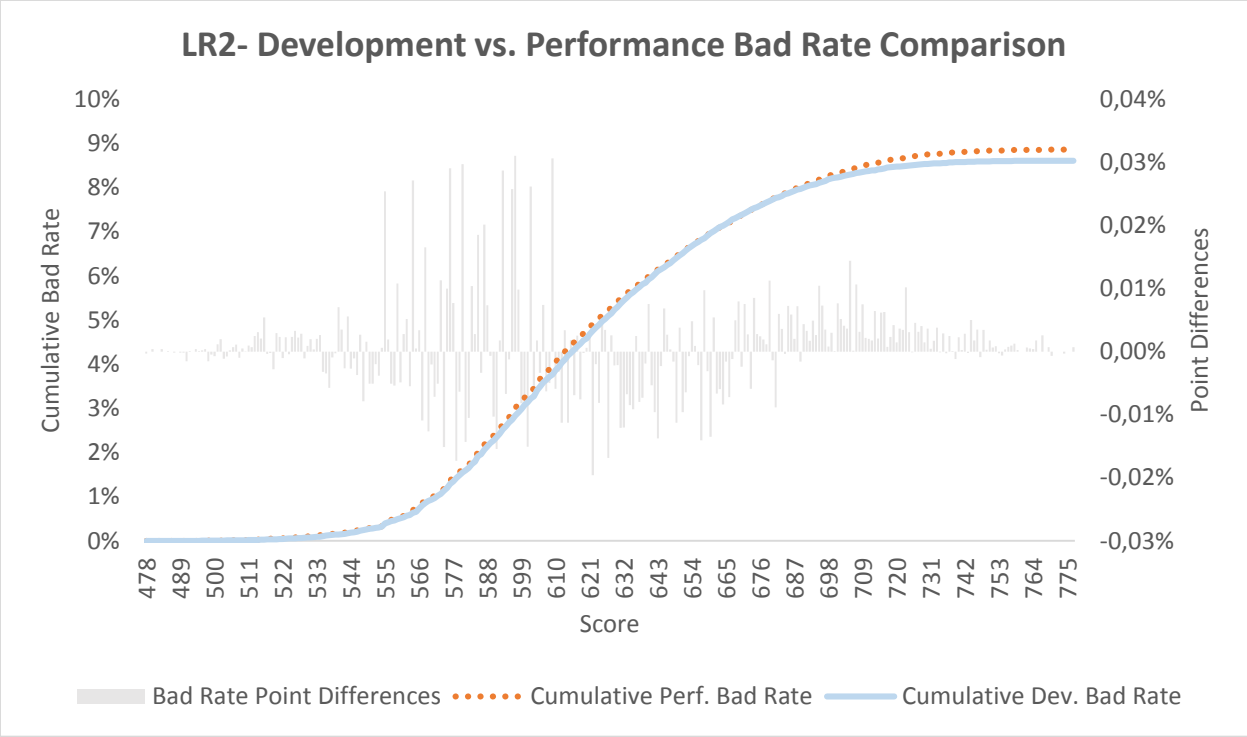

Figure 19 - LR2 - Development vs. Performance bad rate comparison

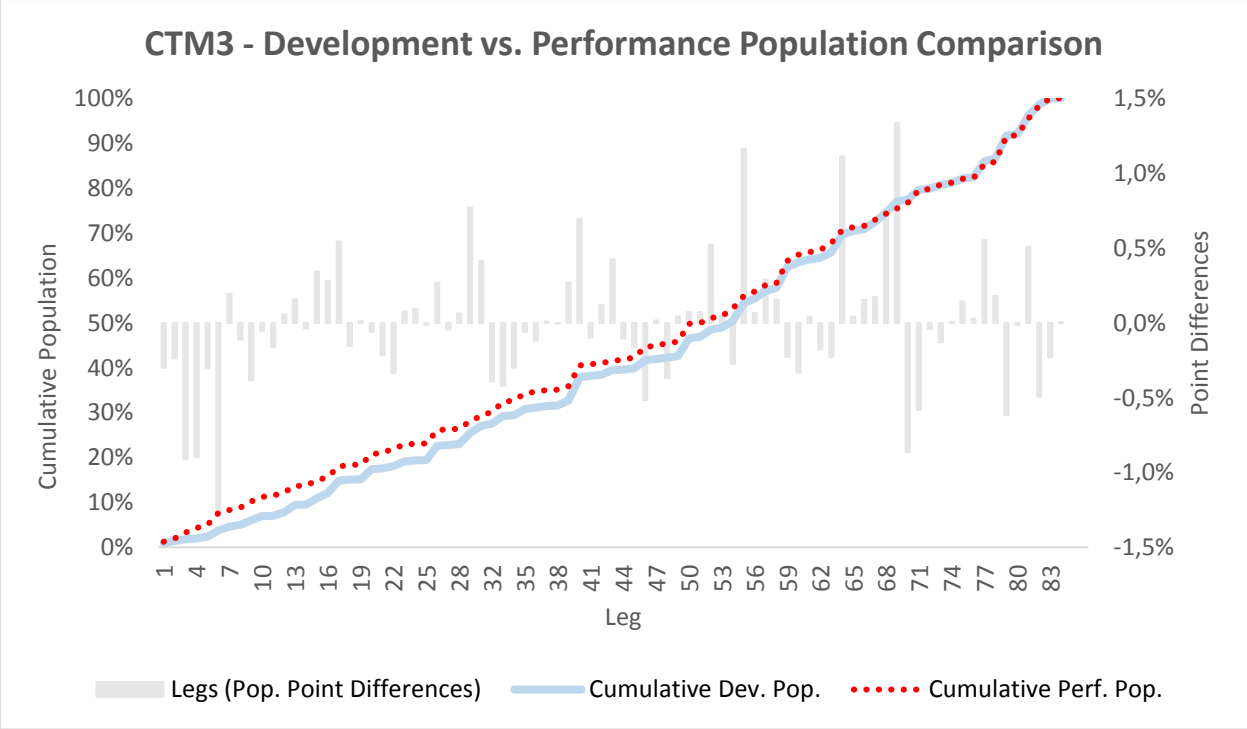

Figure 20 - CTM3 - Development vs. Performance population comparison

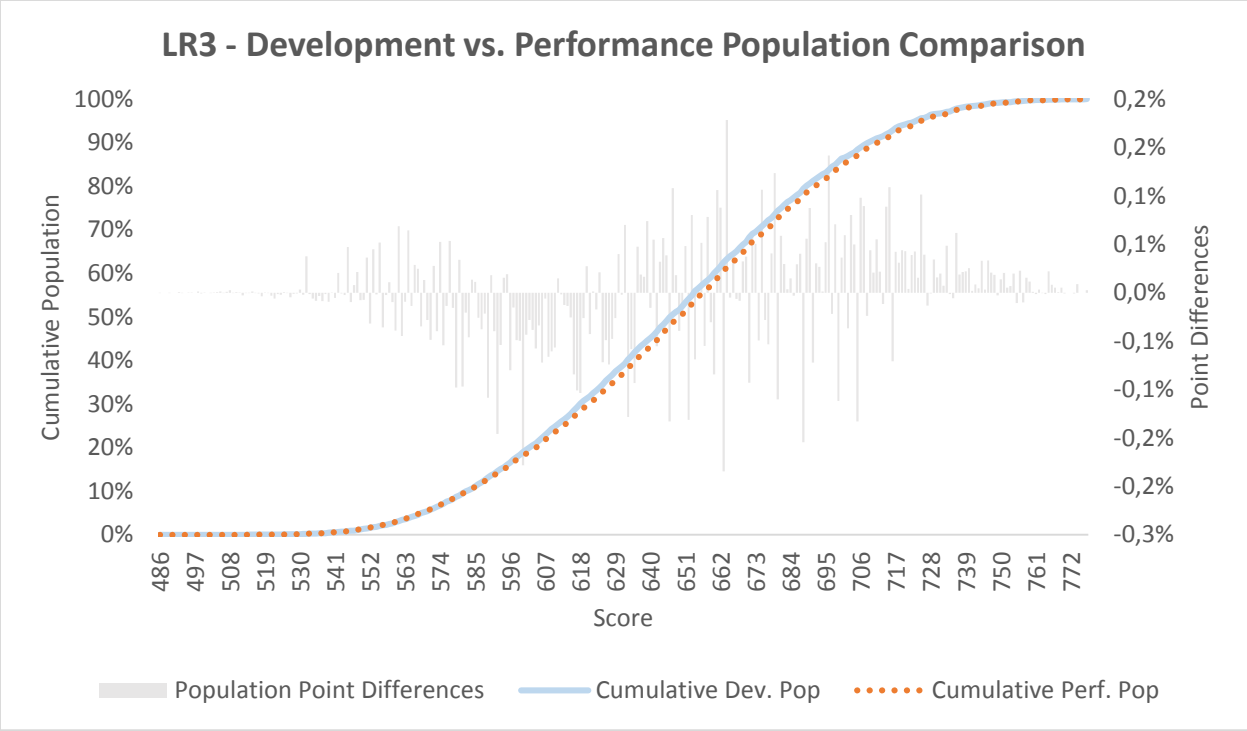

Figure 21 - LR3 - Development vs. Performance population comparison

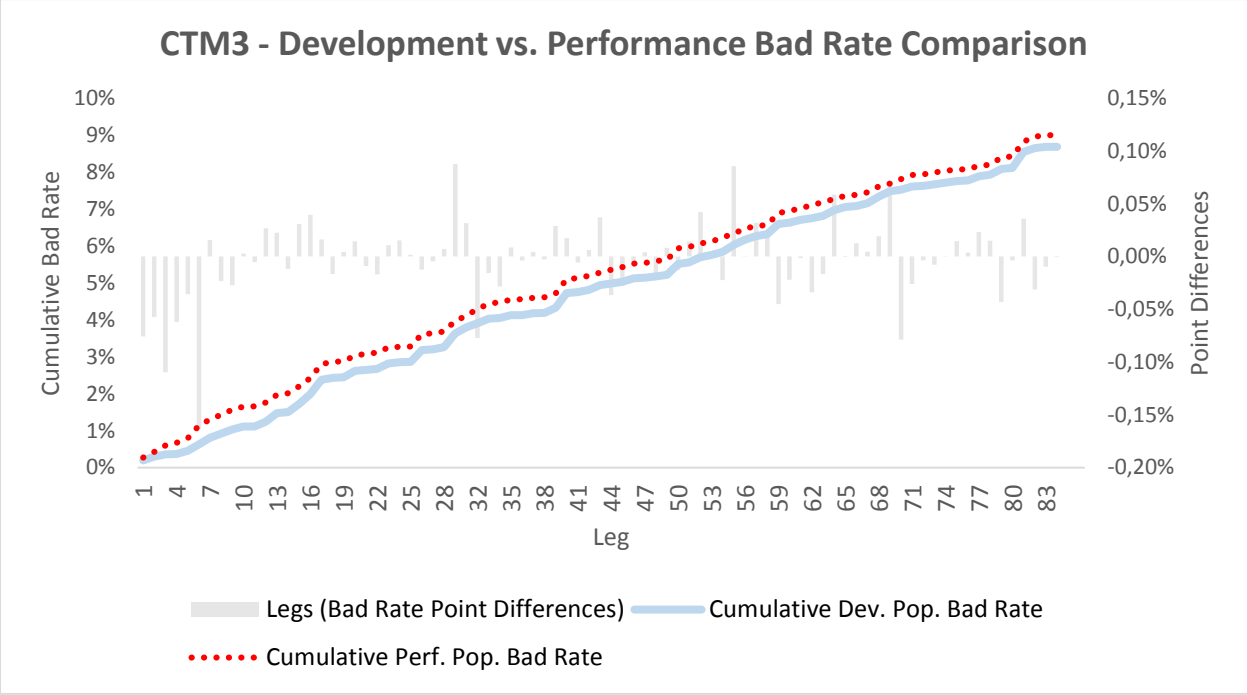

Figure 22 - CTM3 - Development vs. Performance bad rate comparison

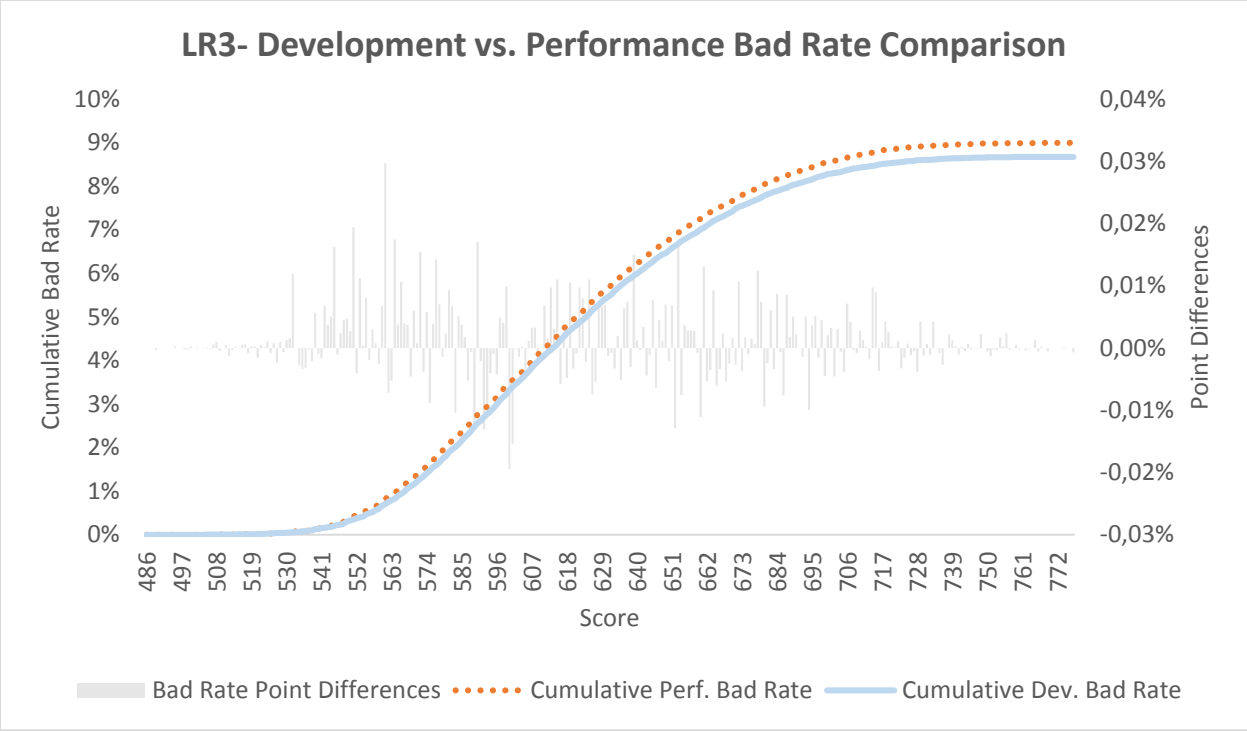

Figure 23 - LR3 - Development vs. Performance bad rate comparison

Appendix B - Tables

Overall population%	Fitted model Population correct%	Ideal model population correct %
0%	0%	0%
20%	19.31%	20%
40%	38.26%	40%
60%	56.83%	60%
80%	74.69%	80%
100%	91.31%	100%

Table 3 - Lift chart example

Predicted	Goods (actual)	Bad (actual)	Total
Goods	250,915	23,843	274,758
FPR			8.68%

Table 4 - False Positive Rate

Information Value	**Predictive Power of Variable**
<0.02	No prediction ability
0.02 to 0.1	Weak predictor
0.1 to 0.3	Medium predictor
0.3 to 0.5	Strong predictor
>0.5	Very strong predictor

Table 5 - Information Value decisioning criteria

	Loans Volumes Disbursed				
	2014	2015	2016	2017	2018
January	16,584	18,529	17,892	15,690	21,157
February	15,307	18,021	17,137	17,033	21,135
March	17,085	19,951	23,292	22,355	24,249
April	20,372	19,996	18,048	20,931	23,594
May	18,391	20,641	17,966	23,000	27,273
June	18,336	18,528	17,611	21,401	24,341
July	18,361	19,962	17,175	22,418	24,711
August	18,236	19,764	18,965	22,693	22,762
September	18,239	20,135	19,427	23,049	23,548
October	18,402	18,324	20,460	24,958	21,382
November	16,940	18,795	18,825	22,038	17,672
December	13,796	16,678	14,101	20,023	20,211

Table 6 - Total loan volumes

	Bad Rate				
	2014	2015	2016	2017	2018
January	7.8%	9.7%	8.2%	9.1%	8.9%
February	8.3%	9.6%	8.2%	9.3%	8.9%
March	7.9%	9.0%	8.0%	8.3%	9.0%
April	7.8%	8.5%	8.2%	8.0%	8.5%
May	8.3%	8.0%	8.4%	7.9%	8.7%
June	8.6%	8.4%	8.8%	9.0%	9.7%
July	8.3%	8.5%	8.3%	8.7%	9.2%
August	8.2%	8.4%	8.1%	8.5%	9.5%
September	7.9%	8.4%	8.1%	8.5%	9.3%
October	8.3%	9.0%	8.5%	9.3%	8.5%
November	9.6%	9.6%	10.2%	9.8%	9.5%
December	10.2%	10.0%	9.5%	10.1%	8.3%

Table 7 - Bad Rate corresponding to disbursed month

Rank	CTM1	CTM2	CTM3
1	Bureau Score	Bureau Score	Bureau Score
2	Recency	Recency	Month on Book
3	Month on Book	Month on Book	Recency
4	Worst Contract Status	Worst Contract Status	Worst Contract Status
5	Client Classification	Client Classification	Client Classification
6	Age	Dispute Counter	Dispute Counter
7	Dispute Counter	ST Concurrent	Age
8	Months Since Last Missed PMT	Bank	Months Since Last Missed PMT
9	ST Concurrent	Age	ST Concurrent
10	Bank	Sister Behaviour Score	Source Channel
11	Source Channel		Bank

Table 8 - Variables Ranked by Power - CTM

Variable	Rank-LR1	IV- LR1	Rank-LR2	IV- LR2	Rank-LR3	IV- LR3
Bureau Score	1	0.23	1	0.23	1	0.23
Month on Book	2	0.11	2	0.11	2	0.16
Age	3	0.08	3	0.08	3	0.1
Sister Behaviour Score	4	0.07	4	0.07	4	0.08
Worst Contract Status	5	0.05	5	0.05	5	0.05
Recency	6	0.04	6	0.04	6	0.04
Bank	7	0.03	7	0.03	7	0.03
Months Since Last Missed PMT	8	0.03	8	0.03	8	0.03

Table 9 - Initial variable summary - LR

Ideal Model	CTM1	CTM2	CTM3
0%	0%	0%	0%
20%	19.19%	19.26%	19.31%
40%	38.14%	38.23%	38.26%
60%	56.71%	56.74%	56.83%
80%	74.60%	74.59%	74.69%
100%	91.44%	91.32%	91.31%

Table 10 - CTM Model Fit

Variable	LR1 – P-Value	LR2 – P-Value	LR3 – P-Value
Intercept	0.00%	0.00%	0.00%
Sister Behaviour Score	0.00%	0.00%	0.00%
Bureau Score	0.00%	0.00%	0.00%
Recency	0.00%	0.00%	0.00%
Worst Contract Status	0.00%	0.00%	0.00%
Month on book	0.00%	0.00%	0.00%
Months Since Last Missed PMT*	27.50%	0.00%	0.02%
Age	0.00%	0.00%	0.00%
Bank	0.00%	0.00%	0.00%
Client Classification			0.00%
Model AIC	222,960	330,603	459,786

Table 11 - LR output

Variable	VIF-LR1	VIF-LR2	VIF-LR3
Client Classification			1.05
Bank	1.03	1.03	1.03
Sister Behaviour Score	1.04	1.06	1.08
Recency	1.22	1.14	1.21
Age	1.14	1.14	1.13
Bureau Score	1.21	1.20	1.20
Month on Book	1.35	1.33	1.38
Months Since Last Missed PMT		2.51	2.60
Worst Contract Status	1.09	2.60	2.70

Table 12 - LR VIF statistics

Leg	BureauScore	Recency	Month on Book	Worst Contract Status	Age	Dispute Counter	Months since last missed pmt	ST Concurrent	Bank	Source Channel	Cumulative Dev. Pop.	Cumulative Perf. Pop.	Legs (Pop. Point Differences)	Cumulative Dev. Pop. Bad Rate	Cumulative Perf. Pop. Bad Rate	Legs (Bad Rate Point Differences)
1	<687		>=9	<1							2.2%	1.8%	0.4%	0.28%	0.27%	0.01%
2	<687		>=9	>=1							2.8%	2.2%	0.2%	0.39%	0.36%	0.02%
3	<687		<9								4.7%	3.5%	0.6%	0.77%	0.60%	0.13%
4	>=687 & <712		>=18	<1							7.5%	6.0%	0.4%	1.03%	0.85%	0.01%
5	>=687 & <712		>=18	>=1							8.5%	6.7%	0.3%	1.18%	0.97%	0.04%
6	>=687 & <712		<18							<>Mobi Internet	13.4%	10.1%	1.5%	2.00%	1.58%	0.21%
7	>=687 & <712		<18							Mobi Internet	13.7%	10.9%	-0.6%	2.06%	1.71%	-0.08%
8	>=712 & <737		<4	<1							16.1%	12.8%	0.5%	2.47%	2.08%	0.04%
9	>=712 & <737		>=27	<1		<1					19.9%	16.0%	0.7%	2.75%	2.30%	0.06%
10	>=712 & <737		>=27	<1		>=1					20.2%	16.3%	-0.1%	2.78%	2.34%	-0.01%
11	>=712 & <737		>=4 & <9	<1							22.6%	18.1%	0.6%	3.10%	2.60%	0.06%
12	>=712 & <737		>=9 & <27	<1							26.7%	22.1%	0.1%	3.51%	3.05%	-0.04%
13	>=712 & <737		<9	>=1							26.8%	22.2%	0.0%	3.55%	3.09%	0.00%
14	>=712 & <737		>=27	>=1							28.0%	23.1%	0.3%	3.70%	3.22%	0.03%
15	>=712 & <737		>=9 & <27	>=1		<1					28.9%	23.8%	0.2%	3.85%	3.34%	0.03%
16	>=712 & <737		>=9 & <27	>=1		>=1					29.1%	24.1%	-0.1%	3.89%	3.40%	-0.02%
17	>=737 & <762	<0.76	>=18	<1							32.9%	27.8%	0.1%	4.16%	3.71%	-0.05%
18	>=737 & <762	<0.76	>=18	>=1							33.7%	28.5%	0.2%	4.28%	3.81%	0.02%
19	>=737 & <762	<0.76	<18							<>pureexternal	39.5%	33.2%	1.0%	4.92%	4.41%	0.04%
20	>=737 & <762	<0.76	<18							pureexternal	39.8%	33.7%	-0.2%	4.99%	4.52%	-0.04%
21	>=737 & <762	>=0.76 & <0.95	<27								42.4%	35.9%	0.4%	5.18%	4.68%	0.03%
22	>=737 & <762	>=0.76 & <0.95	>=27								45.0%	38.0%	0.5%	5.32%	4.78%	0.03%
23	>=737 & <762	>=0.95 & <1.71									46.1%	39.3%	-0.2%	5.49%	4.96%	0.00%
24	>=737 & <762	>=1.71		<1							47.6%	40.7%	0.1%	5.59%	5.07%	-0.01%
25	>=737 & <762	>=1.71		>=1							48.2%	41.4%	0.0%	5.66%	5.13%	0.01%
26	>=762 & <787			>=1		<2					50.4%	43.4%	0.1%	5.90%	5.39%	-0.03%
27	>=762 & <787			>=1		>=2					50.5%	43.8%	-0.2%	5.93%	5.46%	-0.03%
28	>=762 & <787		<9	<1						<>pureexternal	54.2%	46.6%	0.8%	6.24%	5.76%	0.01%
29	>=762 & <787		<9	<1						pureexternal	54.4%	47.1%	-0.2%	6.28%	5.83%	-0.03%
30	>=762 & <787		>=9	<1	<53						65.6%	58.0%	0.2%	6.98%	6.55%	-0.03%
31	>=762 & <787		>=9	<1	>=53						67.8%	60.3%	-0.1%	7.08%	6.65%	0.00%
32	>=787 & <812			1							69.0%	61.7%	-0.1%	7.19%	6.77%	-0.01%
33	>=787 & <812		<9	<1							71.6%	63.9%	0.4%	7.40%	6.94%	0.04%
34	>=787 & <812		>=9	<1	<53				<>Bank1		79.7%	72.3%	-0.4%	7.78%	7.35%	-0.02%
35	>=787 & <812		>=9	<1	<53				Bank1		80.9%	73.9%	-0.4%	7.86%	7.44%	-0.01%
36	>=787 & <812		>=9	<1	>=53						83.6%	76.8%	-0.2%	7.96%	7.54%	0.00%
37	>=787 & <812			>=2							83.9%	77.2%	-0.1%	8.01%	7.59%	-0.01%
38	>=812 & <837			<1				<>No			85.8%	79.8%	-0.7%	8.11%	7.72%	-0.03%
39	>=812 & <837			<1				No			92.9%	88.5%	-1.6%	8.39%	8.06%	-0.07%
40	>=812 & <837			>=1							93.5%	89.4%	-0.3%	8.45%	8.16%	-0.03%
41	>=837		<9								94.3%	90.5%	-0.2%	8.49%	8.22%	-0.02%
42	>=837		>=9				<1000				94.7%	91.1%	-0.3%	8.51%	8.27%	-0.03%
43	>=837		>=9				>=1000	<>No			95.4%	92.6%	-0.8%	8.54%	8.32%	-0.02%
44	>=837		>=9				>=1000	No			100.0%	99.9%	-2.8%	8.65%	8.52%	-0.09%
45	Null										100.0%	100.0%	0.0%	8.65%	8.52%	0.00%

Table 13 - CTM1

Leg	BureauScore	Recency	Month on Book	Worst Contract Status	Client Classification	Age	Dispute Counter	ST Concurrent	Bank	Sister Behaviour Score	Cumulative Dev. Pop.	Cumulative Perf. Pop.	Legs (Pop. Point Differences)	Cumulative Dev. Pop. Bad Rate	Cumulative Perf. Pop. Bad Rate	Legs (Bad Rate Point Differences)
1	<687		>=27	<1							0.8%	0.6%	0.2%	0.08%	0.06%	0.02%
2	<687		>=9 & <27	<2							2.3%	1.3%	0.8%	0.32%	0.17%	0.12%
3	<687		>=27	>=1							2.5%	1.5%	0.1%	0.35%	0.19%	0.02%
4	<687		>=9 & <27	>=2							2.6%	1.5%	0.1%	0.38%	0.20%	0.02%
5	<687		<9								4.3%	2.4%	0.7%	0.71%	0.41%	0.12%
6	>=687 & <712	<0.76	>=18	<1							5.5%	3.3%	0.3%	0.85%	0.52%	0.03%
7	>=687 & <712	>=0.76	>=18	<1							7.0%	4.2%	0.6%	0.97%	0.58%	0.06%
8	>=687 & <712		>=18	>=1							7.9%	4.8%	0.3%	1.11%	0.69%	0.03%
9	>=687 & <712		<9								10.8%	7.1%	0.7%	1.65%	1.20%	0.02%
10	>=687 & <712		>=9 & <18				<1				12.6%	8.1%	0.8%	1.91%	1.35%	0.12%
11	>=687 & <712		>=9 & <18				>=1				12.7%	8.2%	0.0%	1.94%	1.38%	0.00%
12	>=712 & <737		>=9 & <27	1							13.5%	8.8%	0.2%	2.06%	1.46%	0.05%
13	>=712 & <737		>=27	<1			>=1				13.8%	9.2%	-0.1%	2.10%	1.50%	0.00%
14	>=712 & <737		>=9 & <27	<1							17.8%	11.9%	1.3%	2.53%	1.78%	0.15%
15	>=712 & <737		>=2 & <9	>=1							18.0%	12.1%	0.0%	2.57%	1.81%	0.00%
16	>=712 & <737		>=27	>=1							19.1%	13.0%	0.2%	2.72%	1.94%	0.02%
17	>=712 & <737		>=9 & <27	>=2							19.4%	13.2%	0.1%	2.78%	1.98%	0.02%
18	>=712 & <737		<2								20.6%	14.2%	0.2%	3.01%	2.28%	-0.06%
19	>=712 & <737		>=2 & <9	<1	<>pureexternal			Yes			21.1%	14.9%	-0.1%	3.10%	2.40%	-0.03%
20	>=712 & <737		>=2 & <9	<1	pureexternal			Yes			21.2%	15.0%	-0.1%	3.13%	2.43%	-0.01%
21	>=712 & <737		>=2 & <9	<1				<>Yes			23.8%	17.0%	0.6%	3.47%	2.75%	0.02%
22	>=712 & <737		>=27	<1		<48	<1				26.8%	19.2%	0.8%	3.70%	2.92%	0.06%
23	>=712 & <737		>=27	<1		>=48	<1				27.4%	19.7%	0.1%	3.73%	2.95%	0.00%
24	>=737 & <762	<0.76	>=18	<1			<1				30.9%	22.6%	0.6%	3.98%	3.15%	0.05%
25	>=737 & <762	<0.76	>=18	<1			>=1				31.2%	23.0%	-0.1%	4.01%	3.19%	-0.01%
26	>=737 & <762	<0.76	>=18	>=1							32.0%	23.7%	0.1%	4.12%	3.29%	0.02%
27	>=737 & <762	<0.76	<18		<>pureexternal		<2			<622	35.0%	26.3%	0.3%	4.50%	3.74%	-0.07%
28	>=737 & <762	<0.76	<18		<>pureexternal		<2			>=622	37.3%	28.6%	0.1%	4.74%	4.07%	-0.09%
29	>=737 & <762	<0.76	<18		<>pureexternal		>=2				37.4%	28.7%	0.0%	4.75%	4.08%	0.00%
30	>=737 & <762	<0.76	<18		pureexternal						37.8%	29.6%	-0.5%	4.83%	4.29%	-0.12%
31	>=737 & <762	>=0.76 & <0.95	<36								40.9%	31.7%	1.0%	5.05%	4.43%	0.08%
32	>=737 & <762	>=0.76 & <0.95	>=36								42.6%	33.0%	0.4%	5.14%	4.49%	0.03%
33	>=737 & <762	>=0.95 & <1.71									43.8%	34.1%	0.1%	5.31%	4.65%	0.02%
34	>=737 & <762	>=1.71	<27								44.5%	34.6%	0.2%	5.38%	4.70%	0.02%
35	>=737 & <762	>=1.71	>=27								45.9%	35.9%	0.2%	5.48%	4.78%	0.02%
36	>=762 & <787			1							47.8%	37.9%	-0.2%	5.69%	5.00%	-0.02%
37	>=762 & <787		<9	<1	<>pureexternal						51.2%	41.4%	-0.1%	6.00%	5.45%	-0.15%
38	>=762 & <787		<9	<1	pureexternal						51.5%	42.3%	-0.6%	6.05%	5.61%	-0.11%
39	>=762 & <787		>=45	<1			<1				55.9%	46.5%	0.2%	6.25%	5.80%	0.02%
40	>=762 & <787		>=45	<1			>=1				56.2%	47.1%	-0.3%	6.28%	5.85%	-0.02%
41	>=762 & <787		>=9 & <45	<1		<43					61.3%	51.4%	0.8%	6.65%	6.15%	0.06%
42	>=762 & <787		>=9 & <45	<1		>=43					64.8%	54.5%	0.4%	6.85%	6.34%	0.02%
43	>=762 & <787			>=2							65.3%	55.0%	0.0%	6.93%	6.43%	-0.01%
44	>=787 & <812			1							66.5%	56.6%	-0.3%	7.05%	6.58%	-0.04%
45	>=787 & <812		<9	<1	<>pureexternal						68.9%	59.1%	-0.2%	7.22%	6.83%	-0.08%
46	>=787 & <812		<9	<1	pureexternal						69.0%	59.7%	-0.4%	7.25%	6.92%	-0.06%
47	>=787 & <812		>=54	<1							72.8%	64.4%	-1.0%	7.39%	7.13%	-0.06%
48	>=787 & <812		>=9 & <54	<1			<1				81.0%	72.0%	0.6%	7.79%	7.54%	-0.01%
49	>=787 & <812		>=9 & <54	<1			>=1				81.3%	72.6%	-0.2%	7.82%	7.59%	-0.01%
50	>=787 & <812			>=2							81.6%	73.0%	-0.1%	7.86%	7.65%	-0.02%
51	>=812 & <837		<9	<1							83.0%	74.8%	-0.4%	7.95%	7.78%	-0.04%
52	>=812 & <837		>=9	<1					<>Bank1		90.4%	84.0%	-1.9%	8.23%	8.19%	-0.13%
53	>=812 & <837		>=9	<1					Bank1		91.4%	85.7%	-0.7%	8.28%	8.28%	-0.04%
54	>=812 & < 837			>=1							92.1%	86.9%	-0.5%	8.35%	8.39%	-0.04%
55	>=837		<9	<1							93.0%	88.2%	-0.4%	8.40%	8.47%	-0.03%
56	>=837		>=9	<1							99.6%	99.2%	-4.5%	8.58%	8.80%	-0.15%
57	>=837			>=1		<68					99.9%	99.9%	-0.4%	8.60%	8.86%	-0.03%
58	>=837			>=1		>=68					100.0%	100.0%	0.0%	8.60%	8.86%	0.00%
59	Null										100.0%	100.0%	0.0%	8.61%	8.86%	0.00%

Table 14 - CTM2

	BureauScore	Month on Book	Recency	Worst Contract Status	Client Classification	Dispute Counter	Age	Months since last missed pmt	ST Concurrent	Source Channel	Bank	Cumulative Dev. Pop.	Cumulative Perf. Pop.	Legs (Pop. Point Differences)	Cumulative Dev. Pop. Bad Rate	Cumulative Perf. Pop. Bad Rate	Legs (Bad Rate Point Differences)
1	<687	<9									<>Bank1	1.0%	1.3%	-0.3%	0.19%	0.27%	-0.08%
2	<687	<9									Bank1	1.4%	2.0%	-0.2%	0.30%	0.43%	-0.06%
3	<687	>=36					<43					1.9%	3.3%	-0.9%	0.35%	0.60%	-0.11%
4	<687	>=36					>=43					2.0%	4.4%	-0.9%	0.36%	0.67%	-0.06%
5	<687	>=9 & <36						<1000				2.4%	5.1%	-0.3%	0.45%	0.79%	-0.04%
6	<687	>=9 & <36						>=1000				3.7%	7.7%	-1.3%	0.63%	1.13%	-0.16%
7	>=687 & <712	<3							No			4.6%	8.3%	0.2%	0.81%	1.30%	0.02%
8	>=687 & <712	<9							<>No			5.1%	8.9%	-0.1%	0.92%	1.43%	-0.02%
9	>=687 & <712	>=18	<0.57	<1								6.0%	10.3%	-0.4%	1.03%	1.57%	-0.03%
10	>=687 & <712	>=18	>=0.57	<1			<48					6.9%	11.2%	-0.1%	1.11%	1.64%	0.00%
11	>=687 & <712	>=18	>=0.57	<1			>=48					7.1%	11.5%	-0.2%	1.11%	1.65%	-0.01%
12	>=687 & <712	>=18		>=1								7.9%	12.3%	0.1%	1.24%	1.76%	0.03%
13	>=687 & <712	>=9 & <18				<1						9.5%	13.7%	0.2%	1.47%	1.97%	0.02%
14	>=687 & <712	>=9 & <18				>=1						9.6%	13.9%	0.0%	1.50%	2.01%	-0.01%
15	>=687 & <712	>=9 & <9							No			10.9%	14.9%	0.3%	1.73%	2.20%	0.03%
16	>=712 & < 737	<2										12.1%	15.8%	0.3%	1.98%	2.42%	0.04%
17	>=712 & < 737	>=2 & <9		<1	<>pureexternal							14.9%	18.0%	0.5%	2.38%	2.79%	0.02%
18	>=712 & < 737	>=2 & <9		<1	pureexternal							15.1%	18.4%	-0.2%	2.43%	2.86%	-0.02%
19	>=712 & < 737	>=2 & <9		>=1								15.2%	18.4%	0.0%	2.44%	2.87%	0.00%
20	>=712 & < 737	>=27		<1		<1	<48					17.4%	20.7%	-0.1%	2.62%	3.03%	0.01%
21	>=712 & < 737	>=27		<1		>=1	<48					17.6%	21.1%	-0.2%	2.65%	3.07%	-0.01%
22	>=712 & < 737	>=27		<1			>=48					18.1%	22.0%	-0.3%	2.67%	3.11%	-0.02%
23	>=712 & < 737	>=27		>=1								19.2%	23.0%	0.1%	2.81%	3.25%	0.01%
24	>=712 & < 737	>=9 & <27		1		<1						19.4%	23.1%	0.1%	2.85%	3.27%	0.02%
25	>=712 & < 737	>=9 & <27		1		>=1						19.5%	23.2%	0.0%	2.86%	3.28%	0.00%
26	>=712 & < 737	>=9 & <27		<1		<1						22.7%	26.1%	0.3%	3.18%	3.61%	-0.01%
27	>=712 & < 737	>=9 & <27		<1		>=1						22.8%	26.3%	0.0%	3.20%	3.63%	0.00%
28	>=712 & < 737	>=9 & <27		>=2								23.0%	26.4%	0.1%	3.25%	3.68%	0.01%
30	>=737 & <762	<9	<0.38 & >=0.95		<>pureexternal					Mobi Internet		25.5%	28.2%	0.8%	3.63%	3.97%	0.09%
31	>=737 & <762	<9	>=0.38 & <0.95		<>pureexternal							27.0%	29.3%	0.4%	3.80%	4.10%	0.03%
32	>=737 & <762	<9			pureexternal							27.6%	30.2%	-0.4%	3.91%	4.29%	-0.08%
33	>=737 & <762	>=36	<0.76	<1		<1	<58					29.3%	32.3%	-0.4%	4.03%	4.43%	-0.02%
34	>=737 & <762	>=36	<0.76	<1		>=1	<58					29.4%	32.8%	-0.3%	4.05%	4.48%	-0.03%
35	>=737 & <762	>=36	<0.76	<1			<58					30.9%	34.3%	-0.1%	4.12%	4.54%	0.01%
36	>=737 & <762	>=36		<1			>=58					31.2%	34.7%	-0.1%	4.13%	4.55%	0.00%
37	>=737 & <762	>=36		>=1			<48					31.5%	35.0%	0.0%	4.17%	4.60%	0.00%
38	>=737 & <762	>=36		>=1			>=48					31.6%	35.1%	0.0%	4.18%	4.61%	0.00%
39	>=737 & <762	>=9 & <36		1								32.7%	35.9%	0.3%	4.32%	4.71%	0.03%
40	>=737 & <762	>=9 & <36		<1		<1						38.0%	40.5%	0.7%	4.72%	5.10%	0.02%
41	>=737 & <762	>=9 & <36		<1		>1						38.2%	40.8%	-0.1%	4.75%	5.13%	-0.01%
42	>=737 & <762	>=9 & <36		>=2								38.5%	41.0%	0.1%	4.81%	5.18%	0.01%
43	>=762 & <787	<2	<0.76		<>pureexternal							39.4%	41.6%	0.4%	4.94%	5.28%	0.04%
44	>=762 & <787	<2	<0.76		pureexternal							39.6%	41.9%	-0.1%	4.98%	5.36%	-0.04%
45	>=762 & <787	>=2 & <9	<0.76		pureexternal							39.9%	42.3%	-0.2%	5.02%	5.42%	-0.02%
46	>=762 & <787	>=45	<0.76	<1		<1						41.8%	44.7%	-0.5%	5.12%	5.52%	-0.01%
47	>=762 & <787	>=45	<0.76	>=1		<1						42.0%	44.9%	0.0%	5.14%	5.54%	0.00%
48	>=762 & <787	>=45	<0.76			>1						42.3%	45.5%	-0.4%	5.17%	5.59%	-0.02%
49	>=762 & <787	>=9 & <45	<0.76	1								42.6%	45.8%	0.0%	5.21%	5.62%	0.01%
50	>=762 & <787	>=9 & <45	<0.76	<1								46.6%	49.8%	0.1%	5.51%	5.93%	-0.01%
51	>=762 & <787	>=9 & <45	<0.76	>=2								46.8%	49.9%	0.1%	5.55%	5.96%	0.01%
52	>=762 & <787		<0.76	>=2 & <9	<>pureexternal						<>Bank1	48.5%	51.1%	0.5%	5.69%	6.06%	0.04%
53	>=762 & <787		<0.76	>=2 & <9	<>pureexternal						Bank1	49.0%	51.5%	0.0%	5.75%	6.12%	0.00%
54	>=762 & <787		>=0.76 & <0.95						<>No			50.4%	53.2%	-0.3%	5.84%	6.23%	-0.02%
55	>=762 & <787		>=0.76 & <0.95						No			54.4%	56.0%	1.2%	6.03%	6.33%	0.09%
56	>=762 & <787		>=0.95 & <1.71									55.4%	57.0%	0.1%	6.16%	6.47%	0.00%
57	>=762 & <787		>=1.71	<1								57.2%	58.4%	0.3%	6.26%	6.53%	0.03%
58	>=762 & <787		>=1.71	>=1								57.8%	58.9%	0.2%	6.32%	6.57%	0.02%
59	>=787 & <812	>=9	<0.76	<1	<>pureexternal		<53					62.5%	63.8%	-0.2%	6.6%	6.9%	0.0%
60	>=787 & <812	>=9	<0.76	<1	<>pureexternal		>=53					63.5%	65.2%	-0.3%	6.6%	6.9%	0.0%
61	>=787 & <812		<0.76	>=1	<>pureexternal							64.1%	65.7%	0.0%	6.7%	7.0%	0.0%
62	>=787 & <812		<0.76		pureexternal							64.5%	66.3%	-0.2%	6.8%	7.1%	0.0%
63	>=787 & <812		>=0.76 & <0.95						<>No			65.7%	67.7%	-0.2%	6.8%	7.2%	0.0%
64	>=787 & <812		>=0.76 & <0.95						No			69.7%	70.6%	1.1%	7.0%	7.3%	0.1%
65	>=787 & <812		>=0.95 & <1.71									70.5%	71.3%	0.0%	7.0%	7.4%	0.0%
66	>=787 & <812		>=1.71	<27								70.9%	71.6%	0.2%	7.1%	7.4%	0.0%
67	>=787 & <812		>=1.71	>=27								72.4%	73.0%	0.2%	7.2%	7.4%	0.0%
68	>=787 & 812	<9	<0.76	<1	<>pureexternal							74.5%	74.3%	0.8%	7.3%	7.6%	0.0%
69	>=812 & <837	<36	<0.57							<>Mobi Internet		77.0%	75.5%	1.3%	7.5%	7.7%	0.1%
70	>=812 & <837	<36	<0.57							Mobi Internet		77.5%	76.8%	-0.9%	7.5%	7.8%	-0.1%
71	>=812 & <837	>=36	<0.57	<1								79.7%	79.6%	-0.6%	7.6%	7.9%	0.0%
72	>=812 & <837	>=36	<0.57	>=1								79.9%	79.8%	0.0%	7.6%	7.9%	0.0%
73	>=812 & <837		>=0.57 & <0.836									80.7%	80.8%	-0.1%	7.7%	8.0%	0.0%
74	>=812 & <837		>=0.95 & <1.71									81.1%	81.2%	0.0%	7.7%	8.0%	0.0%
75	>=812 & <837		>=1.71	<1								82.2%	82.1%	0.1%	7.8%	8.1%	0.0%
76	>=812 & <837		>=1.71	>=1								82.4%	82.3%	0.0%	7.8%	8.1%	0.0%
77	>=812 & <837		>0.836 & <0.95									86.0%	85.3%	0.6%	7.9%	8.2%	0.0%
78	>=837 & <862	<9		<1								86.7%	85.8%	0.2%	7.9%	8.2%	0.0%
79	>=837 & <862	>=9		<1								91.7%	91.4%	-0.6%	8.1%	8.4%	0.0%
80	>=837 & <862			>=1								92.0%	91.8%	0.0%	8.1%	8.4%	0.0%
81	>=862		<0.57 or >=0.95				<53					96.2%	95.5%	0.5%	8.5%	8.8%	0.0%
82	>=862		<0.57 or >=0.95				>=53					98.6%	98.4%	-0.5%	8.6%	9.0%	0.0%
83	>=862		>=0.57 & <0.95									100.0%	100.0%	-0.2%	8.7%	9.0%	0.0%
84	Null											100.0%	100.0%	0.0%	8.7%	9.0%	0.0%

Table 15 - CTM3

	Bureau Score	Sister Behaviour Score	Age	Month on Book	Recency	Worst Contract Status	Bank
Cat1	>=662 & <=721	<10	<31	<6	<0.1	<1	Bank1
Cat2	>=722 & <=755	>=10 & <630	>=31 & <41	>=6 & <14	>=0.1 & <0.8	>=1	Bank2
Cat3	>=756 & <=796	>=630 & <640	>=41 & <56	>=14 & <28	>=0.8 & <0.9	N/A	Bank3&Bank4
Cat4	>=797 & <=836	>=640 & <650	>=56 & <61	>=28	>=0.9 & <1.8	N/A	Bank5&Bank6
Cat5	>=837	>=650	>=61	N/A	>=1.8	N/A	

Table 16 - LR1 - Variables grouped by category

	Bureau Score	Sister Behaviour Score	Age	Month on Book	Recency	Worst Contract Status	Bank	Dispute Counter	Client Classification	Months Since Last Missed Pmt
Cat1	>=662 & <=725	<10	<34	<6	<0.5	<1	Bank1	<1	Existing Cust.	<500
Cat2	>=726 & <=765	>=10 & <630	>=34 & <41	>=6 & <14	>=0.5 & <0.8	>=1	Bank2	>=1	Pure External & Sister comp. lead	>=500
Cat3	>=766 & <=798	>=630 & <650	>=41 & <54	>=14 & <28	>=0.8 & <0.9	N/A	Bank3&Bank4	N/A	N/A	N/A
Cat4	>=799 & <=837	>=650 & <660	>=54 & <61	>=28	>=0.9 & <1.8	N/A	Bank5&Bank6	N/A	N/A	N/A
Cat5	>=838	>=660	>=61	N/A	>=1.8	N/A	N/A	N/A	N/A	N/A

Table 17 - LR2 - Variables grouped by category

	Bureau Score	Sister Behaviour Score	Age	Month on Book	Recency	Worst Contract Status	Bank	Client Classification	Months Since Last Missed Pmt
Cat1	>=662 & <=725	<10	<31	<4	<0.1	<1	Bank1	Existing Cust.	<500
Cat2	>=726 & <=765	>=10 & <630	>=31 & <40	>=4 & <14	>=0.1 & <0.6	>=1	Bank2	Pure External & Sister comp. lead	>=500
Cat3	>=766 & <=798	>=630 & <650	>=40 & <54	>=14 & <36	>=0.6 & <0.8	N/A	Bank3&Bank4	N/A	N/A
Cat4	>=799 & <=837	>=650 & <660	>=54 & <62	>=36	>=0.8 & <0.9	N/A	Bank5&Bank6	N/A	N/A
Cat5	>=838	>=660	>=62	N/A	>=0.9 & <1.8	N/A	N/A	N/A	N/A

Table 18 - LR3 - Variables grouped by category

	Bureau Score	Sister Behaviour Score	Age	Month on Book	Recency	Worst Contract Status	Bank
Cat1	63	92	80	66	90	98	81
Cat2	81	83	87	79	84	50	87
Cat3	100	92	95	93	90	N/A	91
Cat4	121	98	104	108	98	N/A	97
Cat5	143	107	113	N/A	96	N/A	N/A
Catch All	143	107	113	108	96	63	97

Table 19 - LR1 - Score per variable category

	Bureau Score	Sister Behaviour Score	Age	Month on Book	Recency	Worst Contract Status	Bank	Dispute Counter	Client Classification	Months Since Last Missed Pmt
Cat1	63	65	52	38	58	70	52	66	65	61
Cat2	81	54	61	51	60	29	59	33	46	64
Cat3	100	65	69	64	64	N/A	64	N/A	N/A	N/A
Cat4	121	76	78	82	72	N/A	71	N/A	N/A	N/A
Cat5	143	84	90	N/A	69	N/A	N/A	N/A	N/A	N/A
Catch All	143	84	90	82	69	29	71	33	46	64

Table 20 - LR2 - Score per variable category

	Bureau Score	Sister Behaviour Score	Age	Month on Book	Recency	Worst Contract Status	Bank	Client Classification	Months Since Last Missed Pmt
Cat1	43	73	55	41	69	77	58	73	67
Cat2	62	61	65	57	60	31	65	54	71
Cat3	79	72	76	74	70	N/A	71	N/A	N/A
Cat4	96	81	87	91	72	N/A	79	N/A	N/A
Cat5	115	91	99	N/A	81	N/A	N/A	N/A	N/A
Catch All	N/A	N/A	N/A	N/A	78	N/A	N/A	N/A	N/A

Table 21 - LR3 - Score per variable category

Variable	**LR1**	**LR2**	**LR3**	**CTM1**	**CTM2**	**CTM3**
Bureau Score	✓	✓	✓	✓	✓	✓
Month on Book	✓	✓	✓	✓	✓	✓
Age	✓	✓	✓	✓✓	✓✓	✓✓
Sister Behaviour Score	✓	✓	✓		✓✓	
Worst Contract Status	✓	✓	✓	✓	✓	✓
Recency	✓	✓	✓	✓	✓	✓
Bank	✓	✓	✓	✓✓	✓✓	✓✓
Months Since Last Missed Payment		✓	✓	✓✓		✓✓
Client classification		✓	✓		✓✓	✓✓
Source Channel				✓✓		✓✓
Dispute Counter		✓		✓✓	✓✓	✓✓
Reloan Counter						
ST Concurrent				✓✓	✓✓	✓✓
Total Number Variables in Model	**7**	**10**	**9**	**10**	**10**	**11**

Table 22 - Consolidated View of Included Variables in Each Final Model

Legend	
Symbol	Description
✓	Core Variable
✓✓	Non Core Variable used in the final model
	Un-used in the final model

	Logistic Regression			**Classification Tree**			**SSE(CT) – SSE(LR)**		
SSE	**LR1**	**LR2**	**LR3**	**CTM1**	**CTM2**	**CTM3**	**IT 1**	**IT 2**	**IT 3**
BRSSSE	0.051%	0.076%	0.042%	0.081%	0.185%	0.164%	0.030%	0.110%	0.122%
PDSSEE	5.391%	3.367%	1.513%	7.966%	12.153%	17.925%	2.575%	8.786%	16.412%

Table 23 - Final models' SSE comparison

	Logistic Regression			Classification Tree			SSE(CT) – SSE(LR)		
SSE	**LR1**	**LR2**	**LR3**	**CTM1**	**CTM2**	**CTM3**	**IT 1**	**IT 2**	**IT 3**
BRSSSE	0.051%	0.076%	0.042%	0.081%	0.185%	0.164%	0.030%	0.110%	0.122%
PDSSEE	5.391%	3.367%	1.513%	7.966%	12.153%	17.925%	2.575%	8.786%	16.412%
SSE Diff	**IT1-IT1**	**IT2-IT1**	**IT3-IT2**	**IT1-IT1**	**IT2-IT1**	**IT3-IT2**	**IT1-IT1**	**IT2-IT1**	**IT3-IT2**
BRSSSE	0.000%	0.024%	-0.03%	0.000%	0.104%	-0.021%	0.000%	0.080%	0.012%
PDSSEE	0.000%	-2.024%	-1.85%	0.000%	4.188%	5.772%	0.000%	6.211%	7.626%

Table 24 - SSE difference over time for bad rate prediction and population fit