Credit application scoring for consumers without credit history

LY Mathebula

orcid.org/0000-0002-4101-7605

Mini-dissertation accepted in partial fulfilment of the requirements for the degree *Master of Science in Computer Science* at the North-West University

Supervisor: Dr I Takaidza

Co-supervisor: Mr MB Seitshiro

Graduation: July 2019

Student number: 20929919

Acknowledgements

- Heavenly salutation to Most High, thanking you God for carrying me through all the challenges and for giving me the ability to complete this work.
- Dr Isaac Takaidza, my promoter, a special thank you for the academic support, your patience, guidance and all the hard work you have put into this mini dissertation.
- To my co-promoter, Mr Modisane Seitshiro, thank you for your academic knowledge, support, guidance and encouragement when needed.
- To my dearest family; my mother (Hildah Mathebula), brothers (Monty, Happy and Jabu) and my lovey sister, Nompumelelo, for your words of encouragement and your presence in my life.
- To all my friends, Baby-Joyce, Ndivhuo and Kgomotso, your support really means a lot to me. Thank you for being there.
- To everybody who supported me, thank you very much for carrying me with you.

"Praise be to the Lord my Rock, who trains my hands for war, my fingers for battle."

(Psalm 144:1)

Abstract

Credit scoring is a tool that is used to either qualify or disqualify credit applicants by quantifying the risk factors relevant to classify them to high risk or low risk. Due to a demand in credit inclusion, financial institutions, especially banks must come up with a way of screening and scoring applicants. In most cases, applicants are required to have credit history, or risk being denied credit because these institutions cannot charge high interest rates, mainly because they are obliged legally not to do so on the repayment of the loans due to a lack of the applicant's credit history. In this study, the concept and application of credit scoring is explained. The steps necessary to develop a credit scoring model are outlined with the focus on data that do not have any credit history. Literature is reviewed discussing the background information regarding the performance of the logistic regression model and other statistical models in classifying consumers. Datasets, statistical models, methodology and variables were reviewed and used to assist in building the scorecard. Secondary data collected from the General Household Survey (GHS) is used to classify credit applicants into two groups of high risk and low risk. Binary logistic regression is used to identify the variables that best predict these two groups. The forward selection technique is used in determining variables that are significant. The developed model is tested for prediction accuracy and thereafter, this is followed by key findings and recommendations. In conclusion, the developed model is found to be fitting the data well.

Key words: credit scoring, binary logistic regression, credit history

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION AND BACKGROUND OF THE STUDY

1.1 Introduction

1.1.1 Defining credit scoring

When a financial institution offers a loan, which should be paid back within a specified time, this is known as credit (De la Rey, 2007). Lenders provide money with the anticipation that borrowers will be able to repay their money. However, lenders are required by law to perform an affordability check on their potential customers before approving a loan. This is done to guarantee that the borrower will be able to afford to repay the loan and thus protecting the rights of the consumer. Furthermore, the lender will also determine the probability that the consumer will default on the repayment of the loan and on any possible fraud that might take place during the authorisation of the loan. The lender will then rate its applicants according to their ability to repay the loan as either 'low risk' or 'high risk'. A scale mostly used to rate the applicants is known as a scorecard (De la Rey, 2007). A scorecard assigns a numerical value to applicants which is then compared to a specified cut-off score to separate high risk applicants from the low risk applicants. Applicants with values above the cut-off score are classified as being low risk whilst applicants with values below the cut-off score are classified as being high risk. In addition, other authors define a credit scorecard as the application of numerical data measures have been transformed using statistical models (Abdou & Pointon, 2011). The same authors also defined credit scoring as the use of statistical models to determine the likelihood that a potential borrower is going to default on a loan. Credit scoring models are widely used to evaluate business, real estate, and consumer loans. Furthermore, credit scoring is defined as a set of decision models and their underlying techniques that support lenders with decision making in terms of who is eligible to get credit, how much should they get and what strategies should be implemented in order to improve the profitability of the borrowers (Abdou & Pointon, 2011).

There are different types of scorecards, each with its own use and some lenders use a combination of them. In sub-section 1.1.2 these different scorecards are mentioned and briefly discussed. According to the Federal Deposit Insurance Corporation (FDIC), scorecards that are poorly developed lead to inaccurate scoring and potential big losses to lenders (Cavendish, 2009).

1.1.2 Credit scoring techniques

There are three different techniques or methods used to build scorecards, namely judgmental, statistical and hybrid (Caire, 2004). Judgmental scorecards are built based on a company's credit policy and management risk preferences then ranking applicants according to risk, statistical scorecards are built based on historical data of applicants (examples include decision trees,

artificial neural networks, logistic regression, etc.) and hybrid scorecards are built by combining statistically derived models and judgmental weighted variables (Caire, 2004).

The study will focus on the statistical method of building scorecards. The advantages and disadvantages of statistical credit scoring when compared to subjective scoring as explained by Schreiner (2004) are provided in Table 1-1:

Table 1-1: Advantages and disadvantages of statistical credit scoring when compared to subjective scoring

Advantages	Disadvantages
• The risk is measured − the probability that the loan applicant will default is known; • There is stability – people who portray the same characteristics as measured by the scorecard are treated the same; • It is explicit – the process used to predict risk is clear; • Statistical credit scoring accounts for a wide range of risk factors – applicants must meet financial ratios and policies as described by the lender thus risk can be evaluated and managed. • The scorecard is evaluated and tested before use – the scorecard is tested on current data to determine its effectiveness whereby the predicted risk is compared to the observed risk; • Trade-offs can be revealed – it can assist management with decision making; • Associations between risk and the characteristics of the applicant, the lender and the loan can be identified – characteristics of the repayment of the loan by the applicant can be revealed as well as the corresponding	• Data are required on many loans; • Each loan requires more data – this is useful for applicants with little or no credit history; • Data of high quality is required – data must be cleaned; • It requires a consultant – the building of a scorecard requires an experienced person; • It depends on the integration with the management information systems; • It seems to fix what is not broken – it adds an additional step to traditional ways of evaluation; • Applicants can only be rejected – rejected applicants cannot be approved or modified; • Quantified characteristics are assumed to be linked with risk - it is assumed that risk is associated with gender, age, place of resident, past arrear, etc.; • The history is the determinant of the future – the behaviour of the past is used to predict the future; • Probabilities not certainties – the score outcome is a percentage;

Advantages	Disadvantages
characteristics associated with that applicant; • Changes are not necessary in the current evaluation process before the credit committee – the current data-base used is usually in its current form; • Collection time is reduced – the number of loans and the value paid to high-risk applicants is reduced thus leading to less time spent in collections; • The effect of profits by the scorecards can be revealed as well as the estimation of the first-round; • It performs much better than the automatic grade – scoring can reveal the origin of historic relationship between past arrears and future risk, risk of new applicants can be predicted and, finally, scoring is based on both historical performance and other characteristics of the applicant.	• It is subject to abuse – management might not fully use the information revealed by the scorecard, exceptions can be made and data can be cooked; • Discriminatory predictors might be used – certain groups of applicants can be related to risk.

1.1.3 The process of credit scoring

The application of credit scoring stretches from consumer credit (credit card, personal loan, auto loan, home loan, etc.) to business credit (small business loan) starting from the pre-application stage to credit application stage and finally the credit performance stage (Liu, 2001). The process is depicted in Figure 1-1.

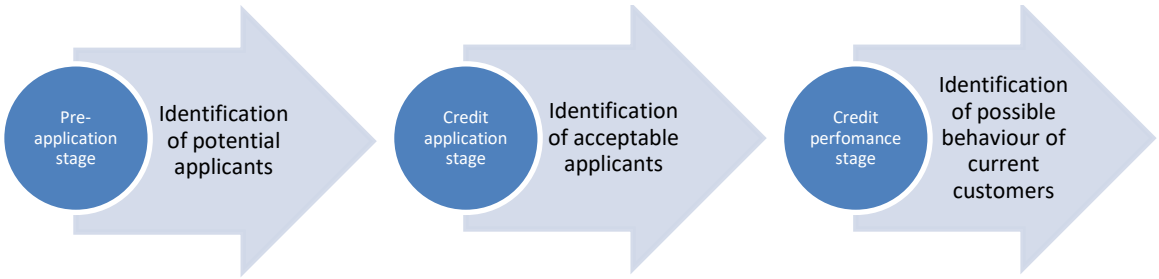

Figure 1-1: The process of credit scoring (Adapted from Liu (2001)

The main purpose of credit scoring is to provide a brief description of the measure of a consumer's creditworthiness (Abdou *et al*., 2016). The objectives of credit scoring are categorised into four aspects as follows (Cavendish, 2009):

Marketing aspect - The cost of acquiring new customers can be reduced when the marketing and management team are assisted to identify credit-worthy customers who have a possibility of responding to their products' promotions (Cavendish, 2009). Moreover, by knowing how many customers are likely to switch to other brands that offer the same products, the marketing team can introduce effective strategies to keep their profitable customers. In addition, more focus is paid on the most profitable account to evaluate the revenue possibility as well as forecasting risk. In this way, customer churn is reduced and more valuable customers are retained. Customer churn is defined as the movement of customers from one provider to another (Hung *et al.*, 2006). The type of scoring models used in this case are the response scoring model, revenue scoring model and the retention or attrition model.

Application aspect - During the application process of the loan, the consumer can be categorised as either low risk or high risk. Low risk consumers are granted the loan and the amount of credit to give the consumer is properly stated. The prediction of the repayment of the loan by the consumer is then determined by calculating the probability of the consumer defaulting on the loan. The type of a credit score card used is known as the applicant scoring (Avery *et a*l., 1996).

Performance aspect - Just like the application scoring, the behaviour of the repayment of the loan is forecast to give attention to consumers that might need assistance, consequently, reducing the probability of default. The type of credit scorecard used is the behavioural scoring, as stated by the Federal Deposit Insurance Corporation (FDIC) (Cavendish, 2009).

Bad debt management aspect - When consumers default, financial institutions must find means to get all or part of the loan owed to them. Proper collection methods must be followed to such consumers to reduce administration cost and maximise the collection of the amount owed. The

type of credit scorecards used are collection scoring model, payment projection model, bankruptcy scoring model and recovery scoring model as mentioned by FDIC (Cavendish, 2009).

The study will focus on the first two stages of the credit scoring processes. Potential applicants will be identified and thereafter scored to determine whether they will be granted credit or not. The credit scorecard to be built is based on the application aspect of the applicant.

1.2 Background to study

Consumers with no credit history are denied credit or cannot secure a loan through regulated financial institutions such as banks, investment companies, insurance companies, mortage companies etc. because these institutions cannot charge high interest rates mainly because they are obliged legally not to do so on the repayment of the loans. Such households are then compelled to rely heavily on their pay day lending, title loans, rent to own, pawn broking and loans with very high interest rates from unregulated financial institutions (Caskey, 2002). These unregulated financial institutions take advantage of the poor's reliance on them and thus enforce negative externalities on the rest of community (Caskey, 2002). Since credit scoring is the mostly used tool to screen consumers and predict loan default, reliance is on credit bureau scores or credit history of consumers. Some lenders solely rely on credit scores to approve loans (Mester, 1997). Due to the lack of credit history, credit scores become unfavourable to consumers with no credit history and they are likely to be denied access to credit.

1.3 Problem statement

Consumers without credit history have unfavourable credit scores or have a high probability to be denied access to credit by regulated financial institutions and might be compelled to take up credit with unregulated financial institutions such as loan sharks who charge very high interest rates.

1.4 Objectives

This mini dissertation comprises of primary, secondary, theoretical and empirical objectives.

i. The primary objective is to apply a logistic regression model to score credit applicants with no credit record.
ii. The secondary objective is to suggest a criteria based on several measures of predictive accuracy of the logistic regression model.
iii. The theoretical objective is to understand the field of credit scoring in retail banks.
iv. Another theoretical objective is to be able to produce the criteria for scoring consumers with no credit record.

v. The empirical objective is the building of the statistical models using data mining methods to classify consumers with no credit record into low risk and high risk.

1.5 Mini dissertation outline

In Chapter 2 a review of the literature on logistic regression applicable to credit scoring is compared with other statistical methods.

Chapter 3 provides the methodology used to develop the binary logistic regression model. The results of the model developed are presented and interpreted in Chapter 4.

Finally, Chapter 5 provides conclusions derived from the findings and recommendations offered with suggestions for future research.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

In this section, the association between credit scoring model building and previous research is considered. Section 2.2 highlights how logistic regression has been used to build a credit scorecard model and how it performs in comparison to other methods. Section 2.3 concludes the chapter.

2.2 Classification models

Logistics regression (LR) and discriminant analysis (DA) techniques are the widely used statistical techniques for building credit scoring models (Abdou and Pointon, 2011). Fisher (1936) was the first person to propose the DA for discrimination and classification purposes. The earliest use of multiple DA to credit scoring was applied by Durand (1941) to examine the application of car loans. Altman (1968) used the DA model to predict corporate bankruptcy. In their research the operational scoring model was based on five financial ratios taken from eight variables from corporate financial statements. The Z-score was produced as a linear combination of the financial ratios and used to assess the financial position of a company. In 1977, Eisenbeis (1978) discovered statistical drawbacks associated with the application of the DA model. The drawbacks included the usage of linear functions instead of quadratic functions, groups definition, a prior probabilities inappropriateness and classification error prediction.

An attempt was made by Desai *et al.* (1996) to investigate the predictive power of feedforward neural networks in comparison to linear discriminant analysis (LDA) and LR. They collected data from three credit unions L, M and N in the South eastern United States for the period 1988 to 1991. Credit union L is mostly teachers (962 observations), credit union M is mostly telephone company employees (918 observations) and credit union N (853 observations) represents a more diverse state-wide sample. The results revealed that all models correctly classified the loans, however, the neural network (NN) model came out as the best classifier with the LDA being last.

To extract the main features for small business credit scoring, Bensic *et al.* (2005) conducted a study on data collected from a Croatian savings and loan association specialising in financing small-sized and medium-sized enterprises of 160 applicants. The accuracy of logistic regression, neural network, classification and regression trees (CART) was compared. Furthermore, backpropagation, radial basis function network, probabilistic and learning vector quantisation NN algorithms were tested by using the forward nonlinear variable selection strategy. The results revealed that the NN model is the best and CART gave based on the contingency coefficient, Kendall's tau- c and the Spearman rank R. The LR model produced the lowest Type II error while

the CART model produced the highest. Overall the NN model was best, wherein 10 variables were identified as important. These variables are of personal and business nature, microeconomic conditions and credit programme characteristics.

The estimation of a credit scoring model for agricultural loans in Thailand was conducted by Limsombunchai *et al*. (2005). Data were obtained from the Bank of Agriculture and Agricultural Cooperative (BAAC) of Thailand and comprises 14 383 good loans and 2 177 bad or default loans. To construct the credit model, the logit multi-layer feed-forward neural network (MLFN) and probabilistic neural network (PNN) were applied to predict the risk of default and creditworthiness of the borrower. The PNN and MLFN are special classes of the artificial neural network (ANN) and the logit model a special class of LR. The credit models were divided into duration (model I) and without duration (model II). Duration in this study refers to the period the borrower is with the bank (in years). The results revealed that the logistic credit scoring model correctly predicted the risk. Insignificant variables are assets, education, return on assets and borrowing from others, while significant variables include age, collateral, leverage ratio, capital turnover and duration at a significant level of 5 percent to predict the probability of default.

Bellotti and Crook (2009a) tested the general economic conditions that are measured by macroeconomic variables hypotheses affected by probability of default (PD). The LR model is compared to the survival analysis (SA) method as credit scoring methods for prediction. The data used were obtained from a UK bank of 100 000 credit accounts opened from 1997 to mid-2005. The training dataset consist of accounts opened from 1997 to 2001 and the testing dataset consisted of accounts opened from 2002 to 2005. The results revealed that both models were able to predict the probability of default accurately.

Abdou *et al.* (2008) investigated the ability of PNN and MLFN, DA, probit analysis (PA) and LR, in evaluating credit risk in Egyptian banks. The data used are from one of the commercial banks in Egypt and comprise 581 personal loans with 433 good loans and 148 bad loans. The results revealed that the LR model has the highest classification rate when compared to the DA model. The DA scoring model has the lowest classification rate.

Data of payment history of members from a recreational club that consists of 977 (35%) defaulters and 1 788 (65%) non-defaulters were used by Yap *et al.* (2011) to improve the assessment of creditworthiness using credit scoring models through data mining. The data were divided into 70 percent for training and 30 percent for validation. The LR, credit scorecard and decision tree models were compared. The results revealed that both the credit scorecard and LR model are quite comparable and outperform the decision tree model when looking at the receiver operating characteristic (ROC) charts. The ROC curve is used to determine the fit of a model by showing

the ability of a model to classify two groups; one that experience an outcome of interest and one that does not (Bolton, 2010). Furthermore, the LR model has the highest sensitivity and the lowest Type II error (a defaulter misclassified as non-defaulter) when compared to the other two models. The model with the highest Type II error and the lowest sensitivity is the decision tree model.

An empirical study of instance sampling in predicting consumer repayment behaviour was conducted by Crone and Finlay (2012). They used two datasets, one was supplied by Experian UK and contained details of credit applications made between April and June 2002, and the second dataset was a behavioural scoring dataset from a mail order catalogue retailer providing revolving credit. The multilayer perceptron (MLP) and CART models were compared with DA and LR. Results obtained from the study show the support for efficient modelling techniques, such as LR obtain near an optimal performance using fewer observations, unlike other methods such as CART and NN. Oversampling has shown increased accuracy relative to under-sampling. The LDA and CART methods demonstrated a greater sensitivity to sample distribution; whereas, for LR it appeared to be less sensitive.

Mansouri and Dastoori (2013) attempted to identify and employ financial ratios affecting the creditworthiness of banking customers by using DA and LR to determine the reliability of different credit scoring models. Data were supplied by one of the branches of Iranian commercial banks, which contained credit files of customers of which a total of 54 creditable (solvent) and 46 non-creditable (insolvent) firms were identified. The results revealed that the LR and DA models scored high classification rates of borrower companies but LR scored the highest. Furthermore, both models showed to be very powerful methods to predict and classify banking customers. However, the LR model outperformed the DA model as demonstrated by the area under the ROC curve.

Abdou *et al.* (2016) conducted a study to identify and investigate the currently used approaches to assessing consumer credit in the Cameroonian banking sector. The authors built appropriate predictive scoring models for creditworthiness then compared performances with the traditional system. Thereafter, they identified significant variables that should be considered in making decisions. The data provided by the largest Cameroonian banks in 2011, comprised 505 good and 94 bad credit cases. Furthermore, 479 observations were used for training and 120 observations were used for validation. There were 23 independent variables (describing the consumer's demographic and financial information) and a dependent variable (loan status). The LR, CART and Cascade Correlation Neural Network (CCNN), a special case on NN, models were then compared. The results revealed that the classification rates of the models range between 88.68 percent and 92.32 percent on average, CCNN obtaining the highest rate whilst the LR model obtained the lowest rate. In terms of the Type II errors, the CCNN model obtained the

lowest whilst the LR model obtained the highest on average. All models were of good quality as confirmed by results shown by the appropriate use criteria (AUC) and Gini coefficients. The significant variables that were identified in the study in all models are previous employments, borrower's account functioning, guarantees, other loans and monthly expenses. A study was conducted by Abid *et al*. (2016) to uncover the issue of allocating credit to bad borrowers. LR and DA models were developed to differentiate between bad and good borrowers. A commercial Tunisian bank data that were collected from 2010 to 2012 with 603 loans granted to new and existing customers was used and consisted of four selected and ordered variables. Only 341 of the consumers were not creditworthy and the remaining 262 creditworthy. The results of the study reveal that the LR model outperformed the DA model in terms of forecasting payment defaults with a 99 percent correct classification rate. The DA model had a correct classification rate of 68.49 percent.

The LR models is an example of predictive models. Predictive models are faced with certain challenges (Potts & Patetta, 2000). Firstly, predictive models use data that are not initially collected for purposes not related to statistical analysis. The data are known to be massive, not clean, erroneous, contain missing values, contain outliers and very dynamic, which makes it difficult to prepare for modelling. It is advisable to only acquire the part of data that is relevant for analysis. Secondly, input variables can contain mixed scales of measures such as intervals, binary, nominal, ordinal and counts, which can be complicated for some models. Thirdly, computational performance is affected more by the number of input variables than the number of cases. The higher the number of input variables, the more challenging it becomes to explore and model the relationship amongst the variables. This is known as the curse of dimensionality. To solve this problem, the number of input variables should be reduced; this phenomenon is known as dimension reduction (Huber, 1997). Furthermore, redundant and irrelevant dimensions should be ignored without carelessly ignoring those that are important. Principal component analysis is one of the methods used for dimension reduction. Fourthly, predictive modelling encounters nonlinearity and interaction challenges. The challenges arise when a dimension affects the target in a complicated way or an input variable depending on another input variable. The curse of dimensionality makes this difficult to unravel. Logistics regression is one model that does not have input variables depending on each other. Lastly, model selection − a model that over-fits data is too sensitive and will not generalise well to new data. An under-fitted model fits the true features in the data.

2.3 Summary

Although variables incorporated in the credit scoring building model differ for various credit models, almost all use credit history or credit record of applicants. This can be seen from the

literature discussed above. The research aims to build a credit scoring model for people without credit history or credit record. It would be interesting to discover which variables are significant in scoring these applicants. Based on the studies described above, the LR model proved to be capable to classify customers. However, its classification capabilities seemed to differ from one study to the next. In this study the logistic regression that will be applied to classify applicants with no credit history will then be evaluated to determine its classification capabilities. In the next chapter the research methodology will be reviewed.

CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY

3.1 Introduction

The previous chapter highlighted the literature review on logistic regression compared to other statistical methods previously used in building a credit scorecard. In this chapter, we address the methodology and data that will be used to develop a credit scoring model for consumers that do not have credit history in South Africa. This includes the data source used, the characteristics of the data and the reconstruction of the datasets.

3.2 Binary logistic regression

This section provides some basic information regarding binary logistic regression modelling.

3.2.1 Introduction

According to Stoltzfus (2011), the logistic regression model is an efficient and powerful way to analyse the effect of a group of independent variables by quantifying each independent variable's unique contribution. Furthermore, by using the components of linear regression that reflect in the logit scale, Stoltzfus (2011) mentions that the logistic regression model can identify the strongest linear combination of variables with the strongest probability of detecting the observed income. The difference between the logistic regression model and the linear regression model is that the logistic regression model has a dichotomous outcome (Abdou & Pointon, 2011).

In order to develop a logistic regression model, background information regarding the methodology of this model is required. This background information is provided in the remainder of this section.

3.2.2 Binary logistic regression model equation

Let Y_i denote a Bernoulli random variable that indicates that an applicant is high risk or low risk at a given period of time taking on one and zero with probabilities of π and $1-\pi$ respectively (Agresti, 2018). Let $X_{i,k}$ represent the k^{th} attribute of individual i and $\beta_0, \beta_1, \dots, \beta_K$ denote unknown parameters.

The conditional probability that an individual is high risk is given by $P\left(Y_i = 1 \mid X_{i,k}\right) = \pi\left(X_{i,k}\right)$, while the conditional probability that an individual is low risk is given by $P\left(Y_i = 0 \mid X_{i,k}\right) = 1 - \pi\left(X_{i,k}\right)$. The odds that an individual is high risk can be calculated as $\frac{\pi\left(X_{i,k}\right)}{1-\pi\left(X_{i,k}\right)}$. Also, the probability that an applicant is high risk is defined by the equation:

$$\pi_i = \pi\left(X_i\right) = \frac{exp\left(\sum_{k=0}^{K}\beta_k X_{i,k}\right)}{1+exp\left(\sum_{k=0}^{K}\beta_k X_{i,k}\right)}, \quad (1)$$

where $i = 1, \ldots, n$ and $k = 0, \ldots, K$.

Rearranging equation (1) gives:

$$\frac{\pi_i}{1-\pi_i} = e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\ldots+\beta_K X_{i,K}}. \quad (2)$$

To create a linear model, the logit transformation is applied to the probability given by the logistic regression model. The logit transformation is the log of the odds, that is, the ratio of the probability of the outcome to the probability of no outcome:

$$G\left(X,\beta\right) = logit\left(\pi_i\right) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_K X_{i,K}. \quad (3)$$

The logit transformation is linear in its parameters, it is unbounded and the estimated probabilities are between zero and one (Jupiter, 2013).

The maximum likelihood methodology is used to estimate the binomial logistic regression. The density function is expressed as follows:

$$f_i\left(Y_i\right) = \left(\pi_i\right)^{Y_i}\left[1-\pi_i\right]^{1-Y_i}. \quad (4)$$

Since it is assumed that observations are independent, the likelihood function will be the product of all individual likelihoods:

$$l\left(\beta\right) = \prod_{i=1}^{n}\left(\pi_i\right)^{Y_i}\left(1-\pi_i\right)^{1-Y_i}. \quad (5)$$

Therefore, the equation (6) below shows the log-likelihood function to be maximised:

$$L\left(\beta\right) = \ln\left(l\left(\beta\right)\right) = \sum_{i=1}^{n}[Y_i\left[\ln\pi_i\right] + \left(1-Y_i\right)\ln\left[1-\pi_i\right]]. \quad (6)$$

3.2.3 Basic assumptions of a binary logistic regression model

The basic assumptions of logistic regression are as follows (Kiveu, 2015) :

- The response variable, Y_i, follows a Bernoulli distribution with parameter $\pi(X_{i,k})$
- There is no need for independent variables to be interval, nor normally distributed
- Little or no multicollinearity
- Mutually exclusive and exhaustive categorical variables
- The sample should have more than 30 observations
- The error terms are independent (Fourie, 2015).

3.2.4 Variable selection techniques

As soon as the data needed to develop the binary logistic model have been collected, there are several techniques that could be applied in order for the dependent variable to be best predicted by determining independent variables. These techniques help in explaining the most possible variance in the dependent variable with the least number of independent variables. Common variable selection techniques include backward elimination (also known as backward selection), forward addition (also known as forward selection), stepwise selection (also known as stepwise) and best subset (also known as all-possible-subset) (Fourie, 2015).

The chosen subset selection method for this research is forward selection method. The method begins with a model that is empty (SAS/STAT, 2010). An adjusted chi-square statistic is computed for each independent variable that is not in the model (SAS/STAT, 2010). At each step an independent variable with the largest adjusted chi-square statistic is added in the model if it is significant at a specified significant level (SAS/STAT, 2010). A significant variable is a variable that has a lower p −value than the specified significance level value, in this study a significance level of $p = 0.05$ is used. The variable remains in the model after it has been entered. The process carries on until none of the remaining variables enters the model because they do not meet the specified level for entry (SAS/STAT, 2010).

3.2.5 Model fit statistics

In this section, model fitting criteria that are considered in the study are described. They are used to compare different models for the same data by comparing the differences between the fitted values and observed values of the data (Sibanda & Pretorius, 2012).

3.2.5.1 -2Log likelihood

The -2Log likelihood equation is:

$$-2ln\,L = -2\sum_i \frac{w_i}{\sigma^2} f_j \left[Y_i \ln\left(\hat{\pi}_i\right) + \left(1-Y_i\right)\ln\left(1-\hat{\pi}_i\right)\right], \tag{7}$$

where σ^2 denotes the dispersion parameter which has a default value of 1, w_i denote the weight values, Y_i is the number of events and $\hat{\pi}_i$ the estimated event probability of the i^{th} observation.

3.2.5.2 AIC – Akaike information criterion

The AIC equation is:

$$\begin{aligned} \text{AIC} &= -2\text{lnL}\left(\sum_{i=1}^{n}[Y_i\left[\ln \pi_i\right] + \left(1-Y_i\right)\ln\left[1-\pi_i\right]]\right) + 2\text{q} \\ &= -2\text{l}\left(\prod_{i=1}^{n}\left(\pi_i\right)^{Y_i}\left(1-\pi_i\right)^{1-Y_i}\right) + 2\text{q} \end{aligned} \tag{8}$$

where *q* is the number of parameters in the model.

3.2.5.3 SC – Schwarz Bayesian information criterion

The SC equation is:

$$\begin{aligned} SC &= -2\text{lnL}\left(\sum_{i=1}^{n}[Y_i\left[\ln \pi_i\right] + \left(1-Y_i\right)\ln\left[1-\pi_i\right]]\right) + \text{q}\log\left(\text{n}\right) \\ &= -2\text{l}\left(\prod_{i=1}^{n}\left(\pi_i\right)^{Y_i}\left(1-\pi_i\right)^{1-Y_i}\right) + \text{qln}\left(\text{n}\right) \end{aligned} \tag{9}$$

where n is the number of trials in the model. The AIC and SC should have values that are as minimum as possible to represent a good fit. These two criterions penalise the model for including too many variables.

3.2.6 Statistical inference for parameters

Three statistical tests for inferences are discussed in this section. These test the significance of coefficients (β_j) from zero by estimating the parameters in the mathematical model (Sibanda & Pretorius, 2012). The hypothesis tested is given as follows:

- $H_0 : \beta_k = 0,\ k = 0,1,2,\ldots,K$,
- $H_A : \beta_k \neq 0,\ k = 0,1,2,\ldots,K$.

The decision rule is to reject H_0 at $\alpha = 0.05$ level of significance, if the p-value is less than the level of significance.

3.2.6.1 Likelihood ratio test

The likelihood ratio test tests whether there is a difference between the reduced and full model as shown below.

Let the full model (0) be

$$\pi_i = \frac{e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\ldots+\beta_K X_{i,K}}}{1+e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\ldots+\beta_K X_{i,K}}} \tag{10}$$

and the reduced model (1) be denoted by

$$\pi_i = \frac{e^{\beta_q X_{iq}+\ldots+\beta_K X_{i,K}}}{1+e^{\beta_q X_{iq}+\ldots+\beta_K X_{i,K}}}. \tag{11}$$

Let L_0 and L_1 denote maximised likelihood for full and reduced model respectively. Let l_0 and l_1 denote the maximised log likelihood for full and reduced model respectively. The likelihood ratio test statistic is given by

$$G^2(X) = -2ln\left(\frac{l_1}{l_0}\right) = -2\left[\ln l_1 - \ln l_0\right]. \tag{12}$$

$G^2(X)$ has a chi-square distribution with $p-q$ degrees of freedom under H_0.

3.2.6.2 Wald test

The Wald test statistics is given by

$$Z^{*2} = \frac{b_k^2}{s^2(b_k)}, \tag{13}$$

Z^{*2} is calculated as the squared value of the coefficient $\left(b_k^2\right)$ divided by the variance of the coefficient $\left(s^2(b_k)\right)$.

3.2.6.3 Score test

The score test is based on the distribution of the derivatives of the log likelihood. Suppose L is the likelihood function that depends on a univariate parameter β and the data are denoted by x, then the score is $U(\beta)=\frac{\partial \log L(\beta \mid x)}{\partial \beta}$ and the observed Fisher information is $I(\beta)=\frac{\partial^2 \log L(\beta \mid x)}{\partial \beta^2}$.

The score test statistics is given by:

$$S(\theta)=\frac{U(\beta_0)^2}{I(\beta_0)}, \tag{14}$$

taking on a chi-square distribution (χ_1^2) when H_0 is true.

3.2.7 Statistical inference for goodness-of-fit

The Pearson and deviance goodness-of-fit tests determine how well the selected model fit the observed data by comparing the overall differences between observed values and fitted values (Sibanda & Pretorius, 2012). The hypothesis tested are given as follows:

$$H_0: E(Y_i)=\frac{e^{\boldsymbol{X}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{X}_i'\boldsymbol{\beta}}},$$

$$H_A: E(Y_{ij})\neq\frac{e^{X_j'\beta}}{1+e^{X_j'\beta}}.$$

The decision rule is to reject H_0 at $\alpha=0.05$ level of significance, if the p-value is less than the level of significance.

3.2.7.1 Pearson's chi-squared test

The test statistic is given by

$$\chi_P^2=\sum_{i=1}^{m}\sum_{j=1}^{k+1}\frac{\left(r_{ij}-n_i\hat{\pi}_{ij}\right)^2}{n_i\hat{\pi}_{ij}}, \tag{15}$$

where m represents the number of subpopulation profiles, $k+1$ is the number of levels of responses, r_{ij} is the sum of the product of the frequencies and the weights associated with jth level responses in the ith profile, $n_i=\sum_{j=1}^{k+1}r_{ij}$ and $\hat{\pi}_{ij}$ is the fitted probability for the jth level at the

ith profile. The degrees of freedom of this test statistic is $mk - p$, where p is the number of parameters estimated. This formulation is taken from SAS/STAT (2010).

3.2.7.2 Deviance goodness-of-fit test

Both the deviance goodness-of-fit test whether a selected model fits the observed data, however, the statistics used are different. The test statistic is given by

$$\chi_D^2 = 2\sum_{i=1}^{m}\sum_{j=1}^{k+1} r_{ij}\frac{r_{ij}}{n_i\hat{\pi}_{ij}} \tag{16}$$

where m represents the number of subpopulation profiles, $k+1$ is the number of levels of responses, r_{ij} is the sum of the product of the frequencies and the weights associated with jth level responses in the ith profile, $n_i = \sum_{j=1}^{k+1} r_{ij}$ (n_i is the value of trials at the ith profile) and $\hat{\pi}_{ij}$ is the fitted probability for the jth level at the ith profile. The degrees of freedom of this test statistic is $mk - p$, where p is the number of parameters estimated. This formulation is taken from (SAS/STAT, 2010).

3.2.7.3 Hosmer-Lemeshow goodness of fit test

The Hosmer-Lemeshow goodness-of-fit test is for datasets that are not replicated or with few replicates. Furthermore, there should be one or more continuous predictors and the response should be binary. Therefore, the grouping of observations based on the values of the estimated probabilities is proposed by Hosmer and Lemeshow (2000) to show that through simulation there is a statistic that is chi-squared distributed when there is no replication of the dataset (SAS/STAT, 2010). The observations are sorted in an ascending order of their estimated response level probability specified in the response variable. The observations are then divided into approximately 10 groups according to $M = \lfloor 0.1 \times N + 0.5 \rfloor$, where N and M denote the total number of observations and the number of observations for each class respectively.

Suppose there are n_1 observations in the first block and n_2 observations in the second block where the first block of observations is placed in the first class. If $n_1 < M$ and $n_1 + [0.5 \times n_2] \leq M$ then subjects in the second block can be added to the first class or else in the second class.

In general, suppose subjects of the $(j-1)$th block have been placed in the kth class and c is the total number of observations in the kth class. If $c < M$ and $c + \lfloor 0.5 \times n_j \rfloor \leq M$ then observations

in the jth block containing n_j observations can be added to the kth class or the next class. In addition, the last two classes are collapsed to form only one class if the number of observations in the last class does not exceed $\left[0.5 \times N\right]$. This formulation is adapted from (SAS/STAT, 2010).

The test statistic of Hosmer and Lemeshow is given by

$$X_{HL}^2 = \sum_{i=1}^{g} \frac{\left(O_i - N_i \bar{\pi}_i\right)^2}{N_i \bar{\pi}_i}, \tag{17}$$

where g is the number of classes, $\bar{\pi}_i$ is the average estimated predicted probability of an event outcome for the ith class, O_i is the total frequency of event outcome in the ith class and N_i is the total frequency of observations in the ith class.

The decision rule is to reject H_0 at $\alpha = 0.05$ level of significance, if the p -value is less than the level of significance. A lack of fit of the model is indicated by large values of $X_{HL}{}^2$. Also the model is inadequate if the p-value is less than the specified 0.05 significance level. Thus, the null hypothesis that the fitted model is adequate is rejected.

3.2.8 Classification table

The classification table groups the input binary response observations according to whether the predicted event or non-event probabilities are above or below a specified cut-off value *z* in the range zero to one. If the predicted event probability is greater than or equal to *z,* then an observation is predicted as an event. If the predicted event probability is less than *z* then an observation is predicted as a non-event. The prediction of event and non-event is only applicable for a binary response data where event is regarded as ordered value one while the response with ordered value zero is the non-event (SAS/STAT, 2010).

Sensitivity and specificity are measures used to determine the accuracy of the classification. Sensitivity is the ability that an event is correctly predicted and is calculated as the proportion of event responses that were predicted to be events. Specificity is the ability that a non-event is correctly predicted and is calculated as the proportion of non-event responses that were predicted to be non-events. Included in the classification table are the false positive percentages, false negative percentages and percentages of correct classification. The percentages of predicted event responses that were observed as non-events is known as false positive or Type I error. The percentages of predicted non-event responses that were observed as events is known as false negative or Type II error (SAS/STAT, 2010).

Sensitivity, also known as recall rate or true positive rate, is the proportion of event responses that were correctly predicted as positive and is calculated as

$$\text{TPR} = \frac{\text{True Positive}}{\text{True Positive} + \text{ False Negative}} \tag{18}$$

Specificity also known as true negative rate is the proportion of non-event responses that were correctly predicted negative and is calculated as

$$\text{TNR} = \frac{\text{True Negative}}{\text{True Negative} + \text{ False Positive}} \tag{19}$$

Precision also known as positive predicted value is calculated as

$$\text{PPV} = \frac{\text{True Positive}}{\text{True Positive} + \text{ False Positive}} \tag{20}$$

Accuracy is calculated as

$$\text{ACC} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{ False Positive} + \text{True Negative} + \text{False Negative}} \tag{21}$$

The F score is calculated as

$$\begin{aligned} F &= \frac{2\text{True Positive}}{2\text{True Positive} + \text{ False Positive} + \text{False Negative}} \\ &= 2\frac{\text{PPV*TPR}}{\text{PPV} + \text{TPR}} \end{aligned} \tag{22}$$

A confusion matrix is then constructed to show the rate of correctly and incorrectly predicted observations for both event and non-event classes. Table 3-1 shows the example of the confusion matrix (Powers, 2011).

Table 3-3: Confusion matrix

	Observed	
Predicted	TP = True Positives (Sensitivity)	FP = False Positives
	FN = False Negative	TN = True Negatives (Specificity)

3.2.9 Area under the ROC curve

The area under the ROC curve is used to measure the model's predictive power. In simple terms, it measures the ability of the model to discriminate between observations that experience the

outcome of interest versus those which do not. It is estimated by a concordance index denoted as C in the "Association of Predicted Probabilities and Observed Response" output table and ranges between zero and one. The higher the C statistic value the more predictive power the model has and the model is a good classifier, that is, the closer the value it is to 1. The C statistic is the area under the ROC curve (SAS/STAT, 2010).

Suppose there is a sample of *m* observations, where B_1 denotes a class of m_1 observations that have been observed to have a certain event. Let B_2 denote a class of $m_2 = m - m_1$ observations that do not have a certain event. A logistic regression is fitted to the data when risk factors are identified. An estimated probability of the event of interest $\hat{\pi}_k$ is calculated for the *k*th observation.

Suppose the *m* observations undergo a test of predicting the event and this test is based on the estimated probability of the event. The event is associated with higher values of the estimated probability. Therefore, ROC curve is constructed by changing the cut-off value that determines the events predicted by the estimated probabilities. Let v denote the cut-off value, therefore:

$$POS(v) = \sum_{i \in B_1} T(\hat{\pi}_i \geq v) \tag{23}$$

$$NEG(v) = \sum_{i \in B_2} T(\hat{\pi}_i < v) \tag{24}$$

$$FALPOS(v) = \sum_{i \in B_2} T(\hat{\pi}_i \geq v) \tag{25}$$

$$FALNEG(v) = \sum_{i \in B_1} T(\hat{\pi}_i < v) \tag{26}$$

$$SENSITIVITY(v) = \frac{POS(v)}{m_1} \tag{27}$$

$$1\_SPECIFICITY(v) = \frac{FALPOS(v)}{m_2} \tag{28}$$

where $T(.)$ denotes the indicator function, $POS(v)$ is the number of event responses that were correctly predicted, $NEG(v)$ is the number of non-event responses that were correctly predicted, $FALPOS(v)$ is the number of event responses that were falsely predicted, $FALNEG(v)$ is the number of non-event responses that were falsely predicted, $SENSITIVITY(v)$ is the sensitivity

of the test, $1\_SPECIFICITY(v)$ one minus the specificity of the prediction of the event test and $\hat{\pi}_i$ is the estimated probability of the event of interest.

The ROC curve is a graphical plot which shows the performance of the binary response classification model when its discrimination cut-off value is changed. It, therefore, is produced by plotting the $SENSITIVITY$ against $1-SPECIFICITY$ at various cut-off values (SAS/STAT, 2010).

3.3 Data description

3.3.1 Data source

This research followed a secondary data analysis (SDA) methodology using data from the South African general household survey (GHS) conducted in 2016 and released in 2017 by Statistics South Africa (StatsSA, 2017). According to Vartanian (2010), secondary data refers to existing data collected by others. Secondary data are chosen over primary data because (Boslaugh, 2007, Vartanian, 2010):

- it is cheaper to obtain,
- the time needed to organise the data is less,
- it can be stored for a long time,
- it comes prepared for use to make organising, coding and analysis easy,
- a variety of questions can be addressed,
- it has the ability to capture policy effects when there is a shift in policy,
- it allows for advanced analysis techniques application for large datasets, and
- the data collection process involves experts.

With the benefits of secondary data being discussed, there are some drawbacks associated with the use of the data, such as the following (Vartanian, 2010):

- the user does not have enough control over the framing and wording of survey items,
- information about people who participated in the survey cannot be followed up,
- one does not have enough information regarding the data collection process, and
- more time is spent retrieving documents related to the data.

The 2017 GHS survey report provides sufficient details on the data collection process.

3.3.2 The study population

The GHS 2017 data were downloaded from the Statistics South Africa website, www.statssa.gov.za. The GHS is a South African survey data record on education, health and social development, housing, household access to services and facilities, food security and

agriculture. The kind of data used is a sample survey where the units of analysis are individuals and households. The household characteristics include dwelling type, home ownership, access to water and sanitation, access to services, transport, household assets, land ownership and agricultural production while the individuals' characteristics include demographic characteristics, relationship to household head, marital status, language, education, employment, income, health, fertility, disability, access to social services and mortality. The main topics covered in the survey are employment, unemployment, labour and employment and demography and population. The GHS had a national coverage and the lowest level of geographic aggregation is province. The coverage area of the survey includes residents of the household and residents in workers' hostels in all nine provinces. Student hostels, old age homes, hospitals, prisons and military barracks are not covered in the survey (StatsSA, 2017).

3.3.2.1 Dataset description

The number of variables and observations in the GHS data are 299 and 21 601, respectively. The larger the data, the more accurate the model should be (Abdou and Pointon, 2011). In terms of the number of observations used for building a credit scoring model, some studies opted for many observations (Banasik *et al.*, 2003, Hsieh, 2004, Bellotti & Crook, 2009b) whilst few observations were used in other studies (Fletcher, Lee & Chen, 2005, Šušteršič *et al*., 2009).

The selection of variables when building a credit scoring model depends on the type of data based on the nature of data used as well as economic and cultural variables that may have an effect on the model. The determinants of scoring a consumer are divided into four areas, namely financial indicators, demographic indicators, employment indicators and behavioural indicators (Thabiso, 2014). The financial indicators refer to the financial status or position of a consumer and they include total assets, gross income of household, monthly costs of household. Demographic indicators include age, gender, marital status, number of dependants, home status, and residential location. Employment indicators include type of employment, length of current employment and number of previous employments. Lastly, behavioural indicators include existing financial account, average balance on existing financial account, loans outstanding, loans defaulted or delinquent, number of payments per year and collateral. It should be noted that this study does not take into account the credit history since it considers consumers without credit history.

Many researchers (Orgler, 1970, Steenackers & Goovaerts, 1989, Banasik *et al*., 2003, Chen & Huang, 2003, Lee & Chen, 2005, Hand *et al*., 2005, Šušteršič *et al*., 2009) used characteristics such as gender, age, marital status, dependants, having a telephone, educational level, occupation, time at present address and credit card holder status to build their credit scorecards.

Characteristics such as time at present job, loan amount, loan duration, house owner, monthly income, bank accounts, having a car, mortgage, purpose of loan and guarantees were also used to build some credit scorecard models, (Orgler, 1970, Steenackers & Goovaerts, 1989, Greene, 1998, Šarlija *et al*., 2004, Ong *et al*., 2005, Lee & Chen, 2005). Some characteristics that are not frequently used are court judgement, worst account status, time in employments and time with bank (Banasik *et al*., 2003, Andreeva, 2006, Banasik & Crook, 2007, Bellotti & Crook, 2009b).

Creditors are advised not to discriminate against credit applicants because of their race, religion, colour, national origin, gender, marital status or age (Mester, 1997). However, for the sake of this study, the study wanted to examine whether the inclusion of such variables will have any significance in classifying consumers into high risk or low risk. Variables that are proposed for the purpose of the study are recorded in Table 3-2. They were selected based on the type of data that is available and what authors have chosen in their studies as discussed above. Table 3-1 highlights the features of the variables. Furthermore, variables with "Not Applicable" and/or "Unspecified" replies that are more than 50 percent are not proposed. The variable Q812Netincome is selected as the dependent variable while the rest of the variables as independent variables. The variable Q812Netincome has been altered into a binary variable called $Response\_Y$, taking on one or zero to indicate low risk or high risk. The value of zero denotes a household that receives an income of less than or equal to R 10 169,12 and is categorised as high risk, while the value of one represents a household that receives an income that is greater than R 10 169,12 and is categorised as low risk. The value of R 10 169,12 is selected to be the average benchmark of affordability in households based on the average net income household of the GHS data of 2017, as shown in Table 4-1.

Table 3-2: Proposed variables for building the credit scoring model

k	X_k	Label	Type / Format	Qualitative /Quantitative
1	$Q511Subs$	House subsidy received	discrete	Qualitative
2	$econact\_hh$	Economic active	discrete	Quantitative
3	$Geotype$	Geography Type	discrete	Qualitative
4	$head\_age$	Age of household head	continuous	Quantitative
5	$head\_popgrp$	Population group of household head	discrete	Qualitative
6	$head\_sex$	Sex of household head	discrete	Qualitative
7	$hholdsz$	Household size	discrete	Quantitative
8	$Metro\_code$	Metro	discrete	Qualitative
9	$Q510aRDP$	RDP or state subsidised dwelling	discrete	Qualitative
10	$Q51MainD$	Main dwelling	discrete	Qualitative
11	$Q56Owner$	Ownership of dwelling	discrete	Qualitative
12	$Q58Val$	Market value of the property	discrete	Qualitative
13	$Q61Phon$	Telephone	discrete	Qualitative

k	X_k	Label	Type / Format	Qualitative /Quantitative
14	$Q62aCell$	Cellular telephone	discrete	Qualitative
15	$Q62bNrCell$	Total number of cellular telephones in the household	discrete	Quantitative
16	$Q64Int1$	Internet connection in the household	discrete	Qualitative
17	$Q64Int7$	Any place via a mobile cellular phone	discrete	Qualitative
18	$Q64Int8$	Any place via any other mobile access services	discrete	Qualitative
19-	$Q812Netincome$	Net household income per month in Rand	continuous	Quantitative
20	$Q814Exp$	Household expenditure	discrete	Qualitative
21	$Q815TotVeh$	Total number of vehicles	discrete	Quantitative
22	$Q89bMain$	Income sources	discrete	Qualitative
23	$Q820Happy$	happier, the same or less happy with life	discrete	Qualitative
24	$Q821Comp$	Computer/Desktop/Laptop	discrete	Qualitative
25	$Q821DVD$	DVD Player/Blu-ray Player	discrete	Qualitative
26	$Q821EStove$	Electric Stove	discrete	Qualitative
27	$Q821Fridge$	Refrigerator or Combined Fridge Freezer	discrete	Qualitative
28	$Q821HomeTh$	Home Theatre System	discrete	Qualitative
29	$Q821MicroW$	Microwave Oven	discrete	Qualitative
30	$Q821TV$	TV Set	discrete	Qualitative
31	$Q821WashM$	Washing Machine	discrete	Qualitative

3.3.3 Data preparation

In this section, the data is evaluated by detecting the missing values, outliers and illogical values and balancing the data. Missing or unspecified values in the Q812Netincome variable are dropped, which reduces the number of records from 21 225 to 21 086. The $head\_age$ variable ranges between eight and 110. The legal age for borrowing is 18 years and older therefore records having the values of less than 18 in the $head\_age$ variable are dropped and this reduces the number of records from 21 086 to 21 030.

The overall data are such that the dependent variable $Response\_Y_i$ has 16 296 records that are categorised as zero and the remaining 4 734 records are categorised as one, taking into account a cut-off value of R 10 169,12 that was made to divide the data into two groups. This cut-off value is the average monthly net income of the survey sample as shown in Table 3-3. The ratio of zero to one is 77:23.

Categorical independent variables have been coded into dummy variables. See Table A-1 in Appendix A for the new variables that are proposed to build the logistic regression. While most independent variables used are qualitative and some are quantitative. The categorised variables were then quantified such that they represent a dummy variable. For example, the province

variable has nine categories which represent all the nine provinces in South Africa. If for example, Mpumalanga were to be chosen, one (1) would denote the event that a person lives in Mpumalanga and zero (0) would denote the event that a person does not lives in Mpumalanga. The same resemblance has been applied in all provinces. Provinces in the dataset are numbered from 1 to 9, where 1 is WC, 2 is EC, 3 is NC, 4 is FS, 5 is KZN, 6 is NW, 7 is GP, 8 is MP and 9 is L.

Table 3-3: Data balancing

Province	*Response_$Y = 0$*	*Response_$Y = 1$*	*Total*
WC	1432	659	2091
EC	2408	400	2808
NC	768	148	916
FS	985	324	1309
KZN	2718	640	3358
NW	1225	217	1442
GP	3258	1779	5037
MP	1366	396	1762
L	2136	171	2307
Total	16296	4734	21030

3.3.4 Model validation

For purposes of validating the credit scorecard model, authors have divided the sample into two parts (Limsombunchai *et al.*, 2005, Landajo *et al.*, 2007), namely training and testing. Other studies have divided their sample into three parts, namely training, validation and testing (Spear and Leis, 1997). Samples have been divided into a 50:50 ratio (Lenard *et al*., 1995), while in others the ratio is 70:30 for validation (Boritz & Kennedy, 1995, Lenard *et al*., 1995, Lee *et al*., 2002). In this study the data are divided into 50:50 ratio, namely training and testing after cleaning it.

The model validation process comprised calculating and comparing one or more quantities to determine the appropriateness of the model (Neter, 2004). The study will compare the values given by accuracy, precision, sensitivity, specificity and the F-measure to determine if there is a difference between the training and testing datasets for the purpose of validating the model.

3.4 Summary

This chapter discussed the data and methodology that will be used in the subsequent chapters to develop part of the credit scoring methodologies for South Africa's households. Section 3.2 documented general information regarding the data used and Section 3.3 contains detailed information regarding the methodology used and the validation of the model after it has been built.

In summary, the research followed a secondary data analysis methodology (SDA) using data from the GHS conducted in 2017. The GHS is a South African survey data record on education, health and social development, housing, household access to services and facilities, food security, and agriculture. This dataset as well as the methodology discussed will be applied in Chapter 4 to derive a logistic regression model that could be used as a credit scoring methodology within the South African household context. The resulting model will contain just quantitative information and, therefore, can only be viewed as the quantitative part of a credit scoring methodology. The model will then be fitted and validated to monitor its behaviour on a new dataset.

CHAPTER 4: RESULTS AND FINDINGS

4.1 Introduction

The previous chapter highlighted data and variables used in building a credit scorecard model. Furthermore, the methodology that will be used to develop a credit scoring models for consumers that do not have credit history was reviewed. In this chapter, variables are briefly explained, a comparison between the nine provinces is carried out and results thereof interpreted. The binary logistic regression model is developed and the results interpreted.

4.2 Variables description

4.2.1 Net household income per month

The Q812Netincome represents the net household income per month, in Rands, and it ranges between R 0 to R 900 000. The value of 999 999 denoted a response that is unspecified.

Table 4-1: Average monthly net income per province

Province	*Average of Monthly Net Income*
WC	11380,76
EC	7069,00
NC	8029,42
FS	11309,16
KZN	8328,09
NW	7571,12
GP	15599,63
MP	11196,93
L	4708,87
Average Grand Total	**10169,12**

The results in Table 4-1 reveal that Gauteng Province (GP) has the highest average monthly net income of R 15 599,63 followed by Western Cape (WC) with a value of R 11 380,76. The province with the least average monthly income is Limpopo with an average monthly net income of R 4 708,87. The average grand total of the monthly net income for the whole population is R 10 169,12.

4.2.2 Household population group

The head_popgrp variable represents the race of the household head and has four categories. The categories are Black or African, which is denoted by 1, Coloured which is denoted by 2, Indian or Asian, which is denoted by 3 and White, which is denoted by 4.

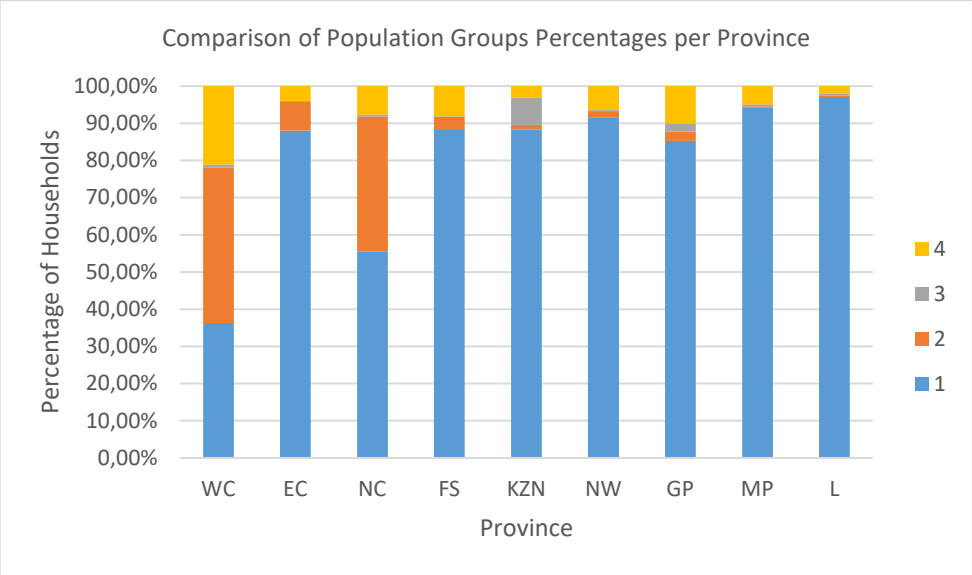

Figure 4-1: Comparison of population group percentages per province

The results in Figure 4-1 indicate that the dominating population group is Black in all provinces except in the Western Cape (WC) where the dominating population group is Coloured. The least dominating population group is Indian or Asian in all provinces except in KwaZulu Natal (KZN). The province with the highest percentage of the Black population group is Limpopo (L) with 97.05 percent while the province with the least percentage is the Western Cape (WC) with 36.25 percent.

4.2.3 Household head gender

The head_sex variable represents the gender of the head of the household, where 1 denotes male and 2 female.

Table 4-2: Household head gender per province

Head_Sex	*WC*	*EC*	*NC*	*FS*	*KZN*	*NW*	*GP*	*MP*	*L*
1	64,85%	49,43%	55,90%	55,61%	51,22%	61,30%	64,70%	58,46%	50,07%
2	35,15%	50,57%	44,10%	44,39%	48,78%	38,70%	35,30%	41,54%	49,93%

The results in Table 4-2 show that a high percentage of males are the heads of the households in all provinces, except in the Eastern Cape (EC) where the highest percentage of household heads are females. The percentages between the two genders in eight provinces differ with 0.3 percent. The province with the highest percentage of females as household heads, is the Eastern Cape (EC) with 50.57 percent, while Western Cape (WC) has the least with only 35.15 percent of females as household heads.

4.2.4 Household head age

The head_age variable represents the age of the household head. The age group has been categorised such that it starts at the minimum borrowing age of 18 grouped by ten until the last group.

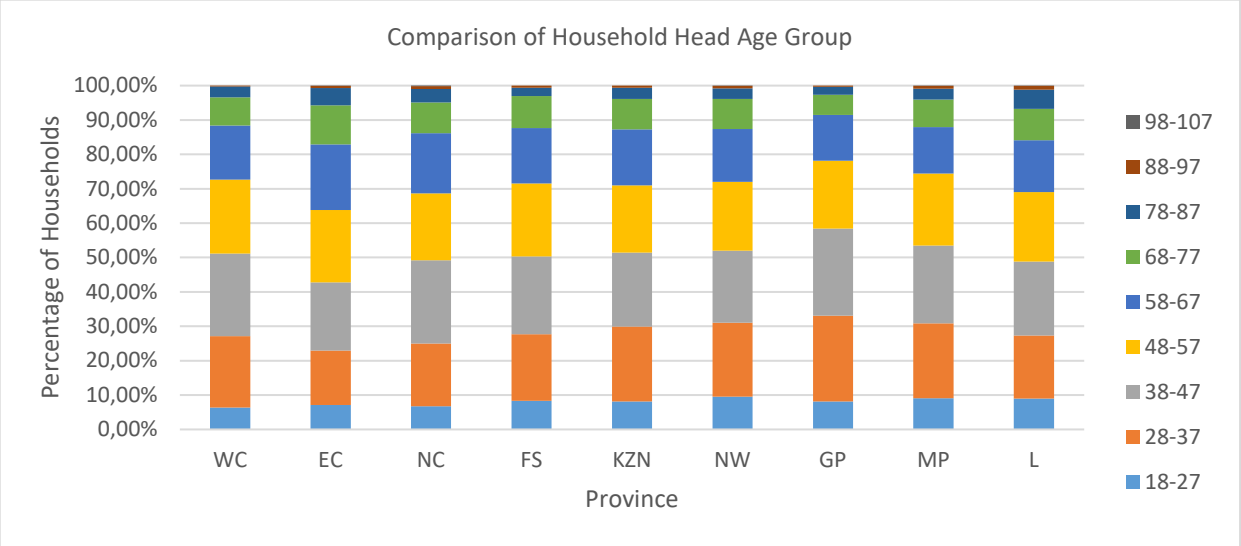

Figure 4-2: Comparison of household age percentages per province

The results in Figure 4-2 show that household heads with the highest percentage are found in the age group of 38 and 47 in Gauteng Province (GP) with a percentage of 24.96 percent while the Eastern Cape (EC) has the least percentage of 19.84 percent in this age group. Based on inspecting Figure 4-2, one could start by indicating that those aged 28-57 dominate household headship across all provinces, while those aged 88-107 and 18-27, feature least in this regard.

4.2.5 Households type of main dwelling

The Q51MainD variable represents the type of main dwelling of the households, where 1 is a dwelling/ house or brick/ concrete block structure on a separate stand or yard or farm; 2 is traditional dwelling/ hut/ traditional materials; 3 is flat or apartment in a block of flats; 4 is cluster house in complex; 5 is town house (semi-detached house in complex); 6 is semi-detached house; 7 is dwelling/ house/ flat/ room in backyard; 8 is informal dwelling/shack in backyard; 9 is informal dwelling/ shack not in backyard, for example, in an informal/ squatter settlement or on farm; 10 is room/ flat let on a property or a larger dwelling servants' quarters/granny flat; 11 is caravan/ tent; and 12 is other types of dwelling. According to the results in Figure 4-3, type 1 dwelling (dwelling/ house or brick/ concrete block structure on a separate stand or yard or on farm) has the highest percentage of households in all provinces where Limpopo (L) has the highest percentage of 83.83 percent and Western Cape (WC) has the least percentage of 56 percent.

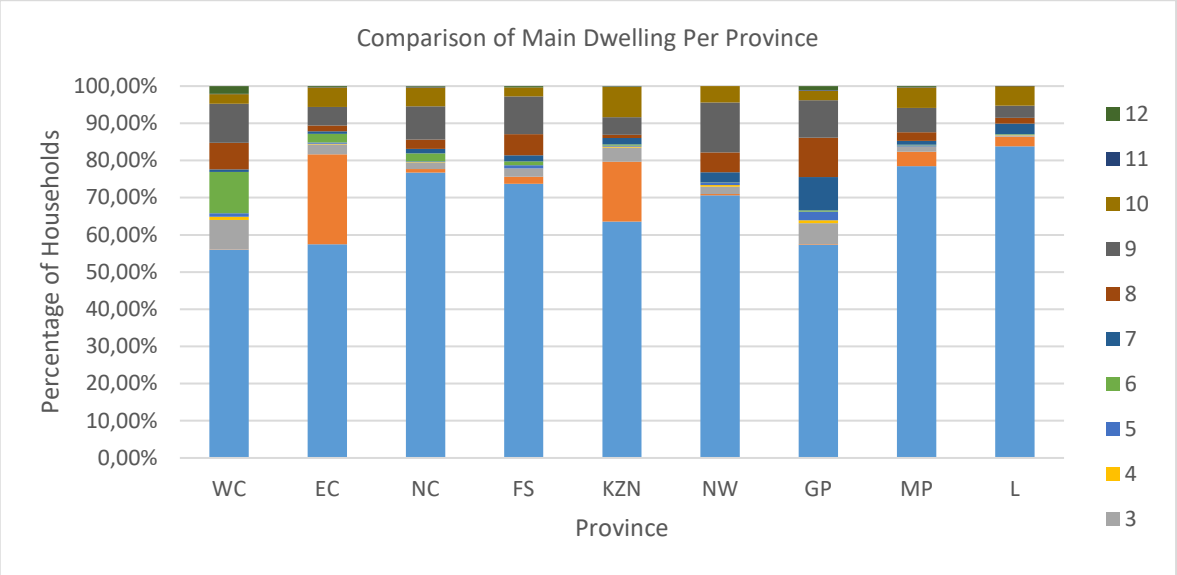

Figure 4-3: Comparison of type of main dwelling percentages per province

4.2.6 Home ownership

The Q56Owner variable seeks to establish whether a household owns the property or they are renting it, where 1 means rented from private individual; 2 indicates rented from other (incl. municipality and social housing institutions); 3 indicates owned but not yet paid off (loan from bank/ financial institution); 4 indicates owned, but not yet paid off (loan from private banker); 5 indicates owned and fully paid; 6 indicates occupied rent-free; 7 indicates other; and 8 indicates do not know. The results in Figure 4-4 reveal that the highest percentage of households own their property and it is fully paid-up. The province with the highest percentage of households whose property is owned and fully paid is in Limpopo (L) province with 75.38 percent, while GP has the least percentage (34.91 percent).

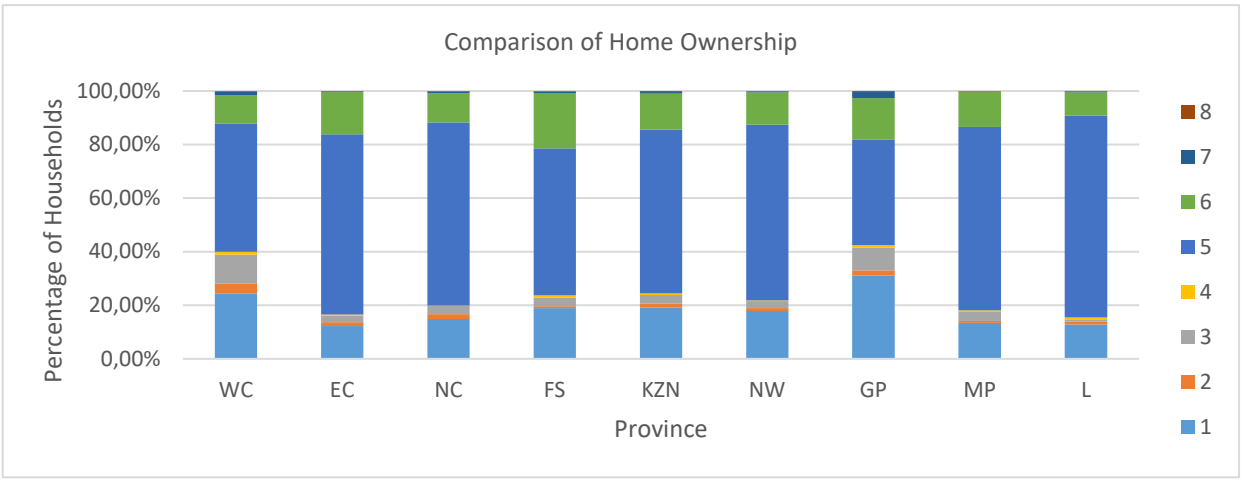

Figure 4-4: Comparison of home ownership percentages per province

4.2.7 Household property estimated market value

The variable Q58Val represents the estimated market value at which the household property could be sold. The options are: 1 (less than R50 000); 2 (R50 001-R250 000); 3 (R250 001-R500 000); 4 (R500 001-R1 000 000); 5 (R1 000 001-R1 500 000); 6 (R1 500 001-R2 000 000); 7 (R2 000 001-R3 000 000); 8 (more than R3 000 000); 9 (do not know); and 99 (unspecified). According to the results in Figure 4-5, the majority of household properties are valued less than or equal to R250 000.

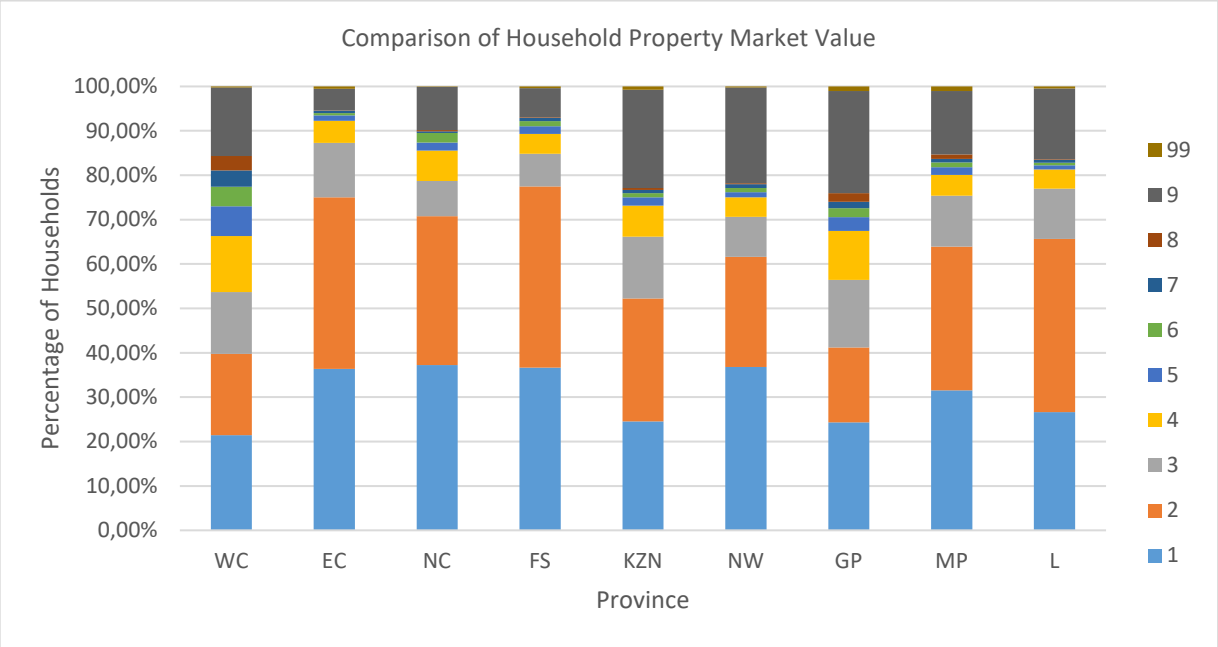

Figure 4-5: Comparison of household property market value percentages per province

4.2.8 RDP or state-subsidised dwelling

The variable Q510aRDP represents whether a household is living in an RDP or state-subsidised dwelling, where option 1 represents yes; 2 no; 3 do not know; and 4 unspecified. The results in Figure 4-6 show that the highest selected option is 2 where households do not live in an RDP or state-subsidised dwelling. EC has the highest percentages of 85.15 percent and FS has the lowest percentage of 71.81 percent for this response.

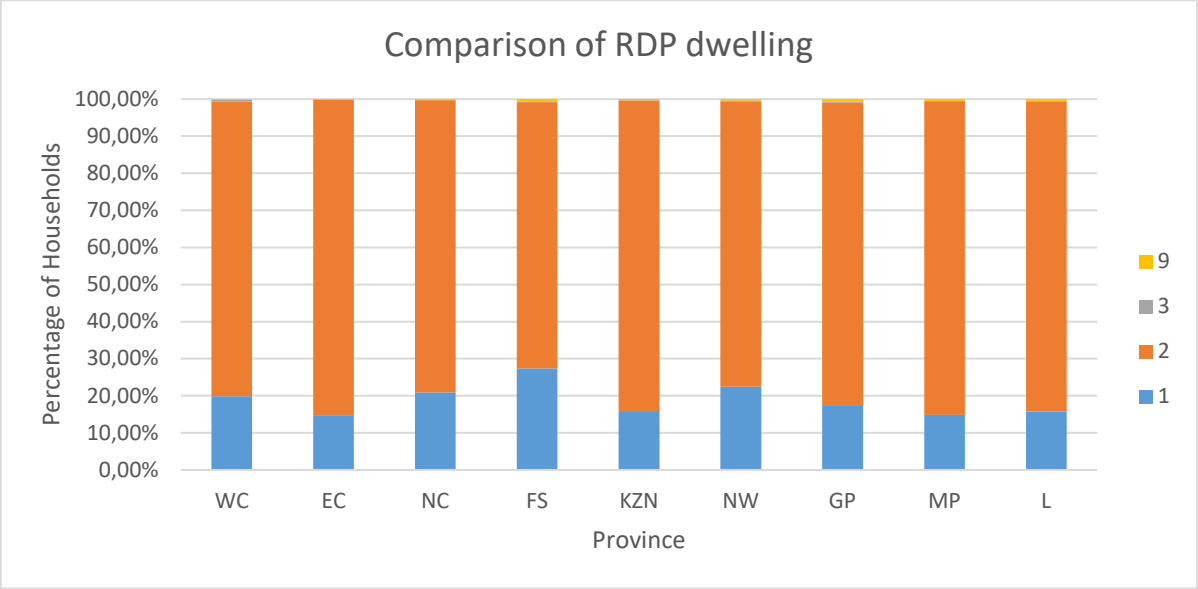

Figure 4-6: Comparison of RDP dwelling percentages per province

4.2.9 Government housing subsidy

The Q511Subs variable seeks to establish any kind of government subsidy received by the household, but excludes subsidies for government employees, where option 1 is yes; 2 no; 3 do not know; and 9 unspecified. The results in Figure 4-7 show that the highest percentage of households do not receive any government subsidy where the Eastern Cape (EC) has the highest percentage of 87.64 percent and Free State (FS) has the lowest percentage of 73.7 percent.

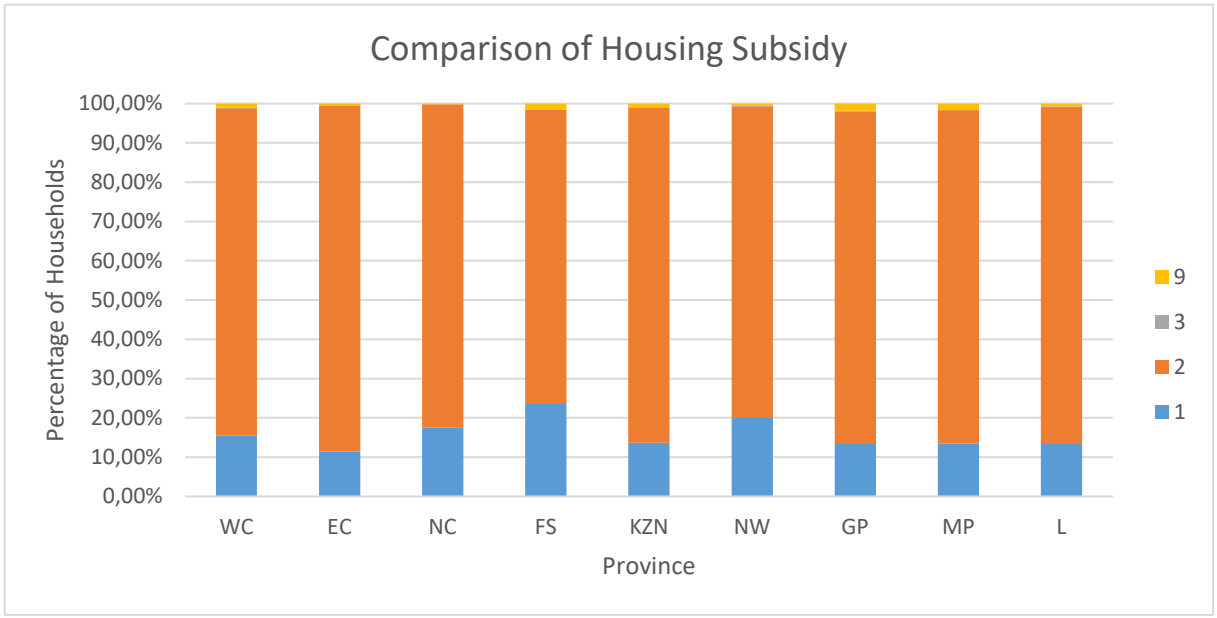

Figure 4-7: Comparison of government housing subsidy percentages per province

4.2.10 Landline telephone

The Q61Phon variable seeks to determine whether households have any landline telephone, where option 1 is yes; 2 no; 3 do not know; and 9 unspecified. The results in Figure 4-8 show that the highest percentage of households do not have any telephone landline, with Limpopo (L) having the highest percentage of 96.88 percent and Western Cape (WC) the lowest percentage of 78.72 percent.

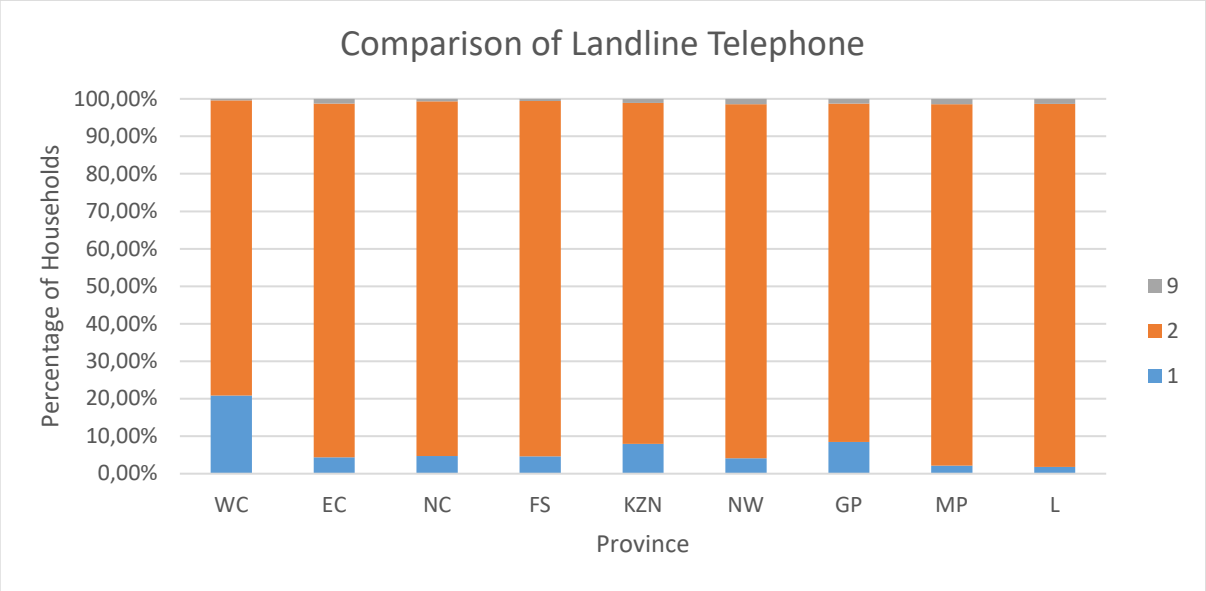

Figure 4-8: Comparison of landline telephone percentages per province

4.2.11 Cellphones

The Q62Cell variable seeks to determine whether households own any cellphones, where option 1 is yes; 2 no; 3 do not know; and 9 unspecified. The results in Figure 4-9 show that the majority of households own a cellphone, where GP has the highest percentage of 98.21 percent and NC has the lowest percentage of 89.30 percent.

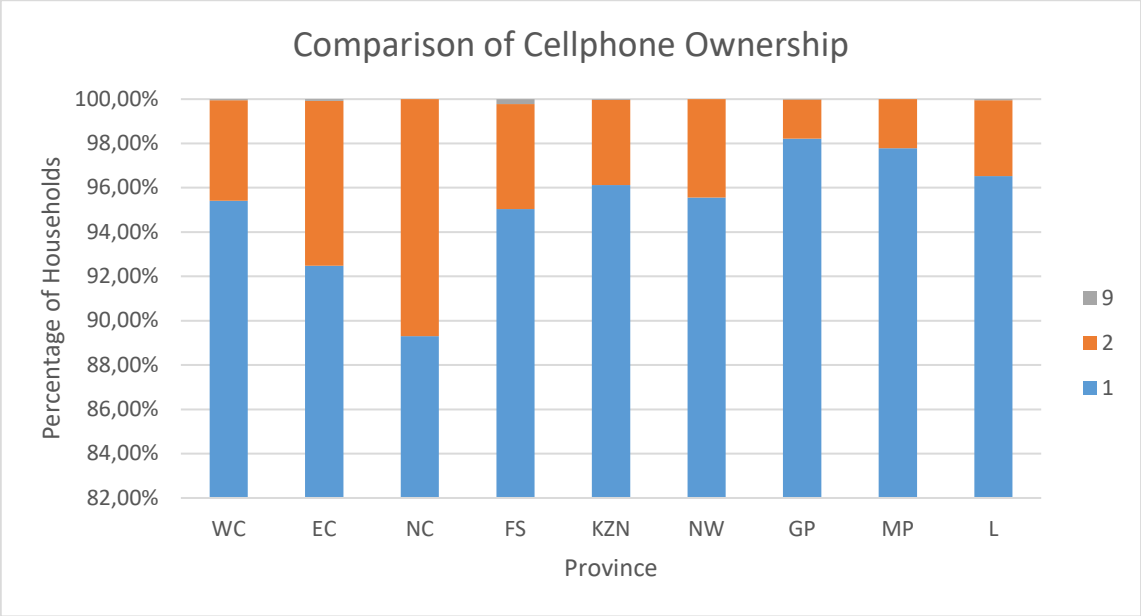

Figure 4-9: Comparison of cellphone ownership percentages per province

4.2.12 Number of cellphones

The Q63NrCell variable seeks to determine the total number of cell phones in a household where the number ranges from 1 to 35, 88 denotes not applicable and 99 unspecified. The descriptive statistics did not include responses that are not applicable and unspecified. The results in Table 4-3 show that the average number of cellphones in a household is 2 in all provinces.

Table 4-3: Number of cellphones households' descriptive statistics

	Q63NrCell
Mean	2,465127848
Standard Error	0,01134277
Median	2
Mode	2
Standard Deviation	1,608195248
Sample Variance	2,586291957
Kurtosis	33,57681893
Skewness	3,162305501
Range	34
Minimum	1
Maximum	35
Sum	49554
Count	20102

4.2.13 Internet connection

The Q65Int1 variable seeks to determine whether a household has internet, where option 1 is yes; 2 no; 3 do not know; and 9 unspecified. The results in Figure 4-10 show that the highest percentage of households do not have any internet connection where Limpopo (L) has the highest percentage of 97.62 percent and WC has the lowest percentage of 74.56 percent.

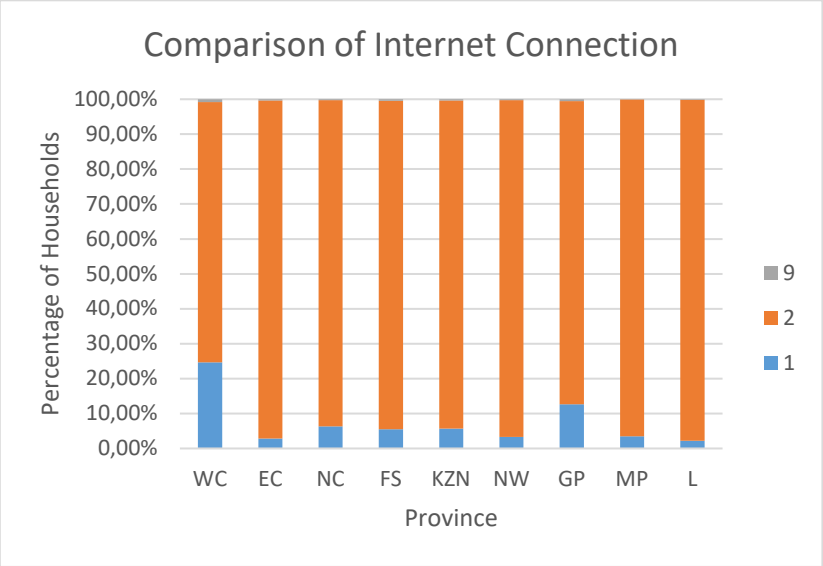

Figure 4-10: Comparison of internet connection percentages per province

4.2.14 Mobile cellphones with internet connection

The Q65Int7 variable seeks to determine whether the household has internet connection to any place through their mobile cellphones, where option 1 is yes; 2 no; 3 do not know; and 9 unspecified. The results in Figure 4-11 show that the highest percentage of households that have internet connection via their cellphones are in GP with a percentage of 62.74.

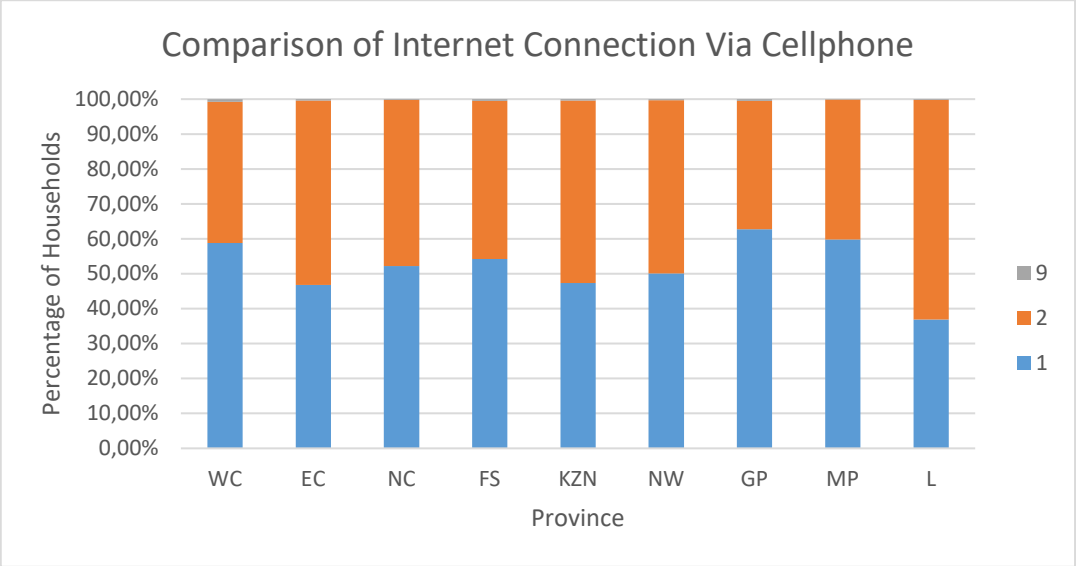

Figure 4-11: Comparison of internet connection via cellphone percentages per province

4.2.15 Any other mobile access internet connection

The Q65Int8 variable seeks to determine whether the household has internet connection to any place via any other mobile access, where option 1 is yes; 2 no; 3 do not know; and 9 unspecified. The results in Figure 4-12- show that the highest percentage of households do not have internet connection via any other mobile access where MP has the highest percentage of 98.3 percent and WC has the lowest percentage of 90.58 percent.

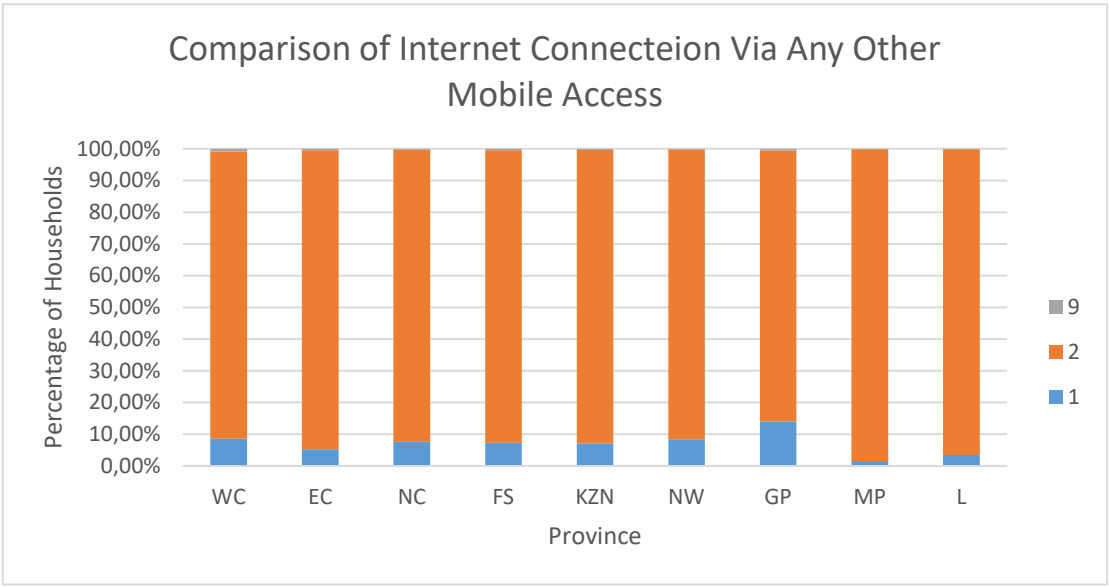

Figure 4-12: Comparison of internet connection via any mobile access percentages per province

4.2.16 Main source of income

The Q89bMain variable seeks to establish the main source of income in households, where nine options are given. Option 1 salaries/ wages/ commission; 2 incomes from a business; 3 remittances; 4 pensions; 5 grants; 6 salaries of farm products and services; 7 other income sources; 8 no income; and 9 unspecified. The results in Figure 4-13 reveal that all provinces selected salaries/ wages/ commission as their main source of income, where WC has the highest percentages of 69.15 percent and EC has the lowest percentage of 39.10 percent. In the EC, grants are the second highest main source of income with a difference of 0.03 percent from salaries/ wages/ commission.

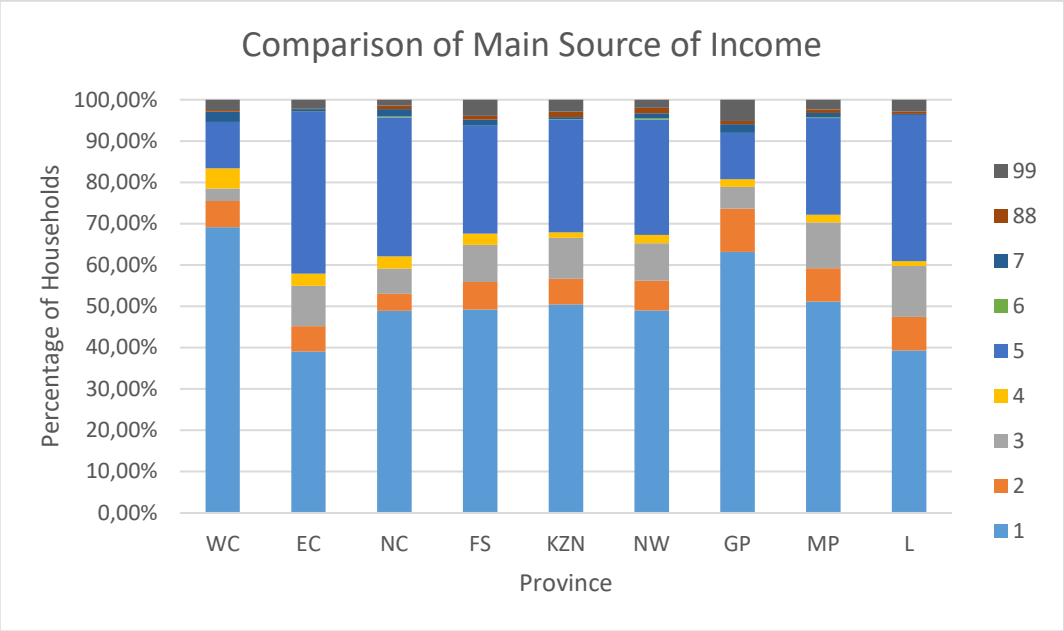

Figure 4-13: Comparison of main sources of income percentages per province

4.2.17 Household expenditure

The Q814Exp variable seeks to establish the expenditure of households where 13 options are given. Option 1 (R0); 2 (R1-R199); 3 (R200-R399); 4 (R400-R799); 5 (R800-R1 199); 6 (R1 200-R1 799); 7 (R1 800-R2 499); 8 (R2 500-R4 999); 9 (R5 000-R9 999); 10 (R10 000 or more); 11 (do not know); 12 (refuse) and 99 (unspecified). The results in Figure 4-14 reveal that the highest expenditure of households is in WC, where 32.9 percent spend R10 000.00 or more.

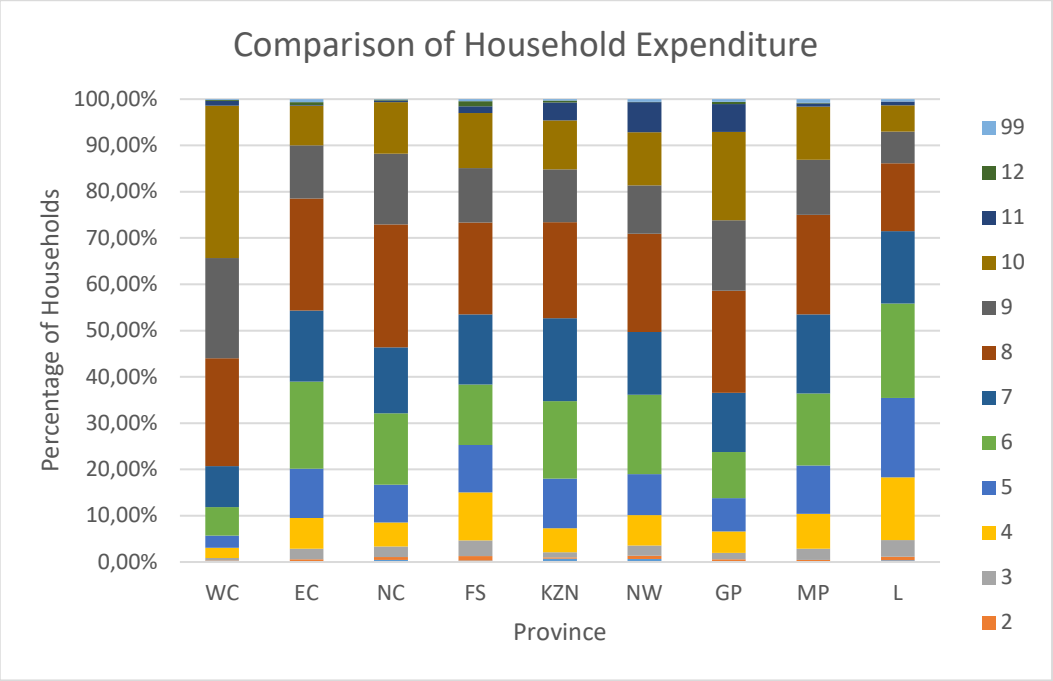

Figure 4-14: Comparison of household expenditure percentages per province

4.2.18 Metro types

The metro variable shows the classification of metro type settlement, where option 1 is metro and 2 non-metro. This variable was not included in the building of the logistic regression model. However, it has been included in this analysis to have a better view of the percentage of people that reside in metro and those who do not. The results in Figure 4-15 show that the provinces with no metros are Northern Cape (NC), North West (NW), Mpumalanga (MP) and Limpopo (L). However, the province with the highest percentage of household in metro is Gauteng Province (GP) with 84.18 percent, while the province with the least percentages is Free State (FS) with 27.96 percent.

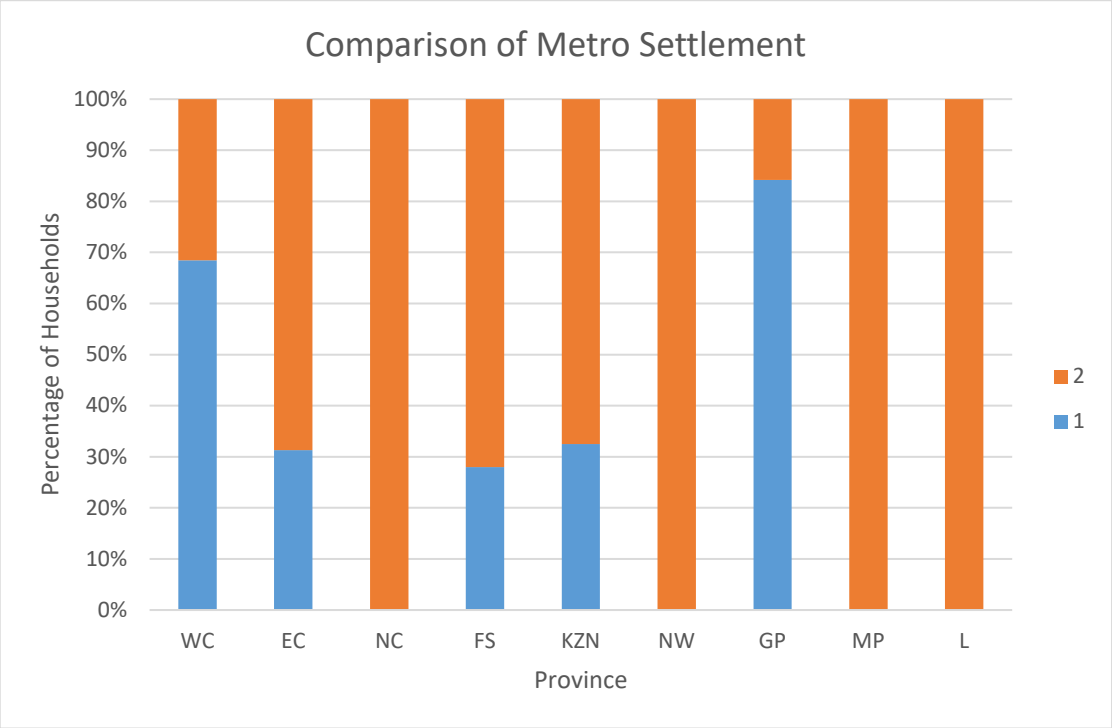

Figure 4-15: Comparison of metro settlement percentages per province

4.2.19 Happiness in life of households

The Q820Happy variable seeks to measure the relative poverty of households, and in this instance the happiness in life of households, where option 1 indicates happier; 2 indicates the same; 3 indicates less happy, 4 indicates responders who refuse to answer; 5 indicates do not know; and 9 indicates unspecified. The results in Figure 4-16 show that the highest percentage of happier households is the Northern Cape (NC) with 39.17 percent, the highest percentage of households that feel the same in life is Limpopo (L) with 45.19 percent and the highest percentage of less happy in life is found in KwaZulu Natal (KZN) with 42.25 percent.

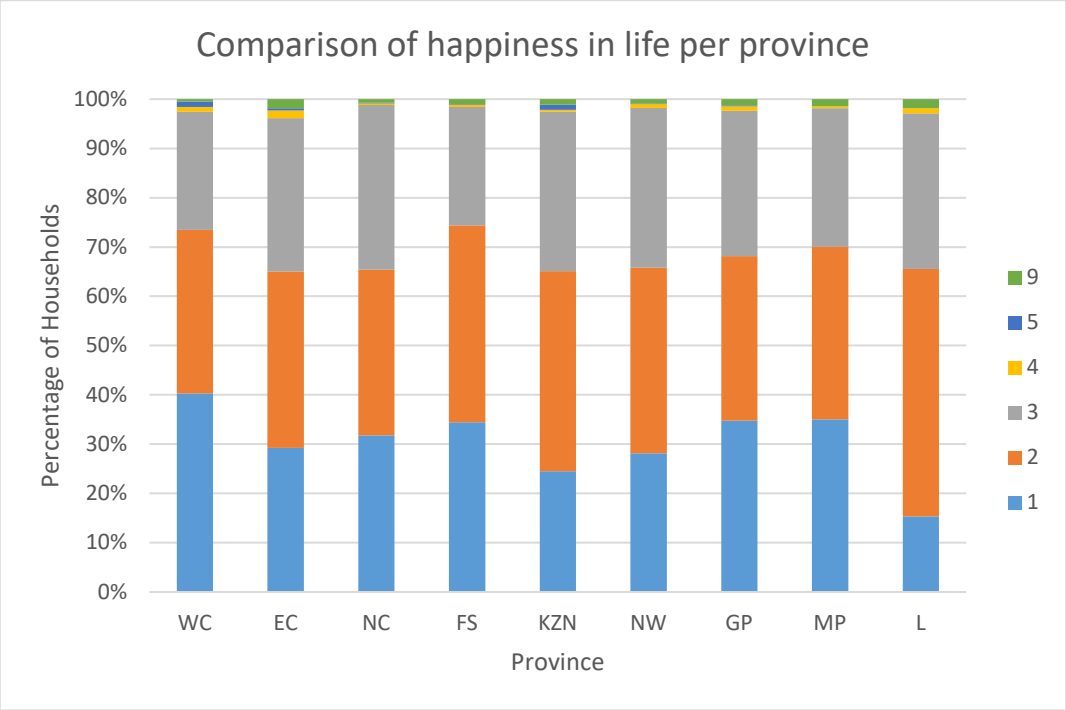

Figure 4-16: Comparison of happiness in life per province

4.2.20 TV ownership

The Q821TV variable seeks to establish whether a household owns a television (TV), where option 1 is yes; 2 no; and 9 unspecified. The results in Figure 4-17 show that the majority of households own TV is yes. The results show that the highest percentage of households that have TV are found in the Western Cape (WC) with 90.00 percent while the Eastern Cape (EC) has the lowest with 72.79 percent.

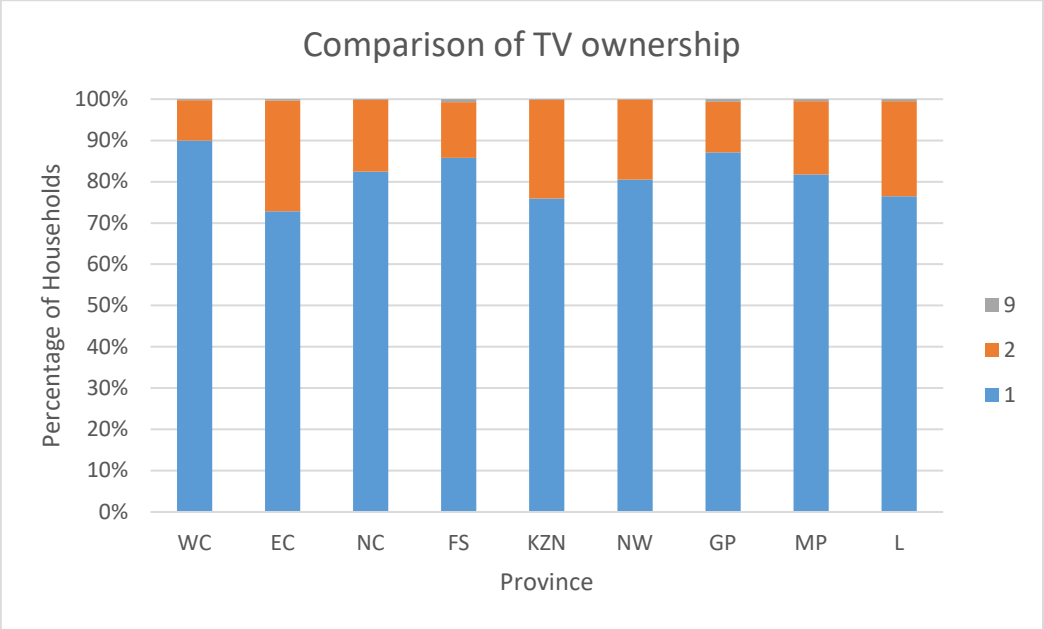

Figure 4-17: Comparison of TV ownership percentages per province

4.2.21 DVD ownership

The Q821DVD variable seeks to establish whether a household owns a DVD, where option 1 is yes; 2 no; and 9 unspecified. The results in Figure 4-18 show that the highest percentage of households that have DVD are found in the Western Cape (WC) with 63.41 percent, while the Eastern Cape (EC) has the lowest with 42.66 percent.

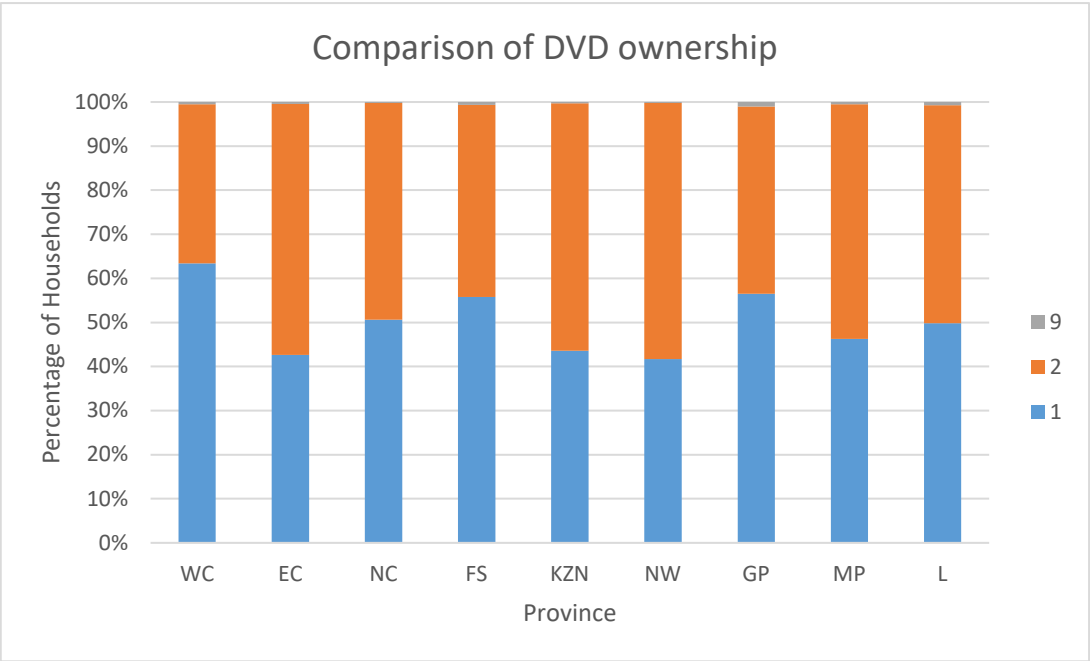

Figure 4-18: Comparison of DVD ownership percentages per province

4.2.22 Computer ownership

The Q821Comp variable seeks to establish whether a household owns a computer, where option 1 is yes; 2 no; and 9 unspecified. The results in Figure 4-19 show that, generally, the highest percentage of household do not own a computer. The highest percentage of households that do have computers are found in the Eastern Cape (EC) with 89.99 percent while the Western Cape (WC) has the lowest with 63.46 percent.

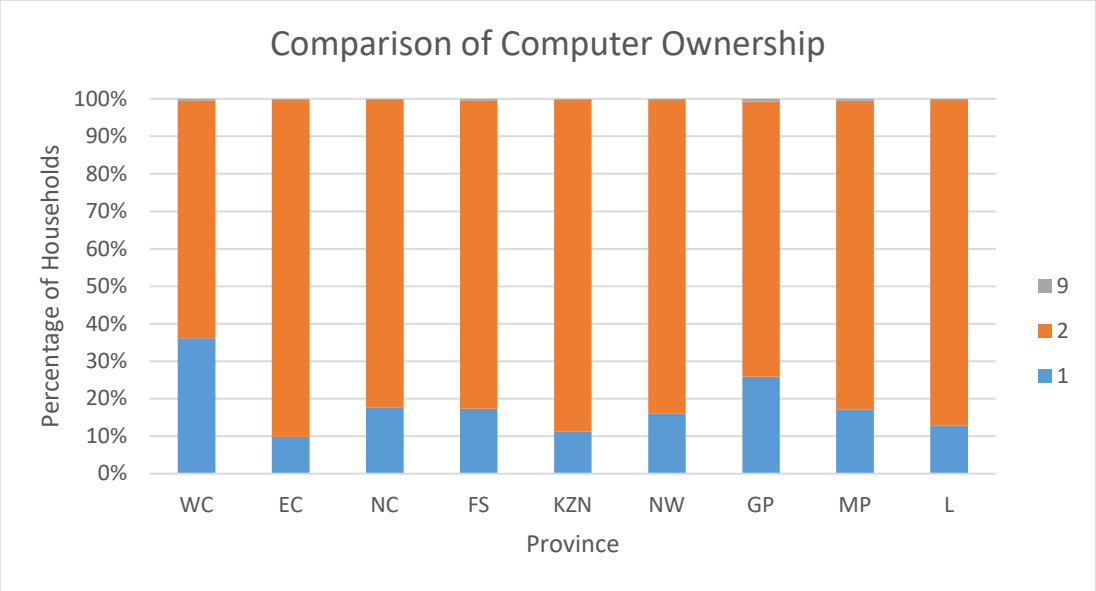

Figure 4-19: Comparison of computer ownership percentages per province

4.2.23 Washing machine

The Q821WashM variable seeks to establish whether a household owns a washing machine, where option 1 is yes; 2 no; and 9 unspecified. The results in Figure 4-20 show that the highest percentage of households that own a washing machine are found in the Western Cape (WC) with 60.74 percent, while KwaZulu Natal (KZN) has the lowest with 14.12 percent.

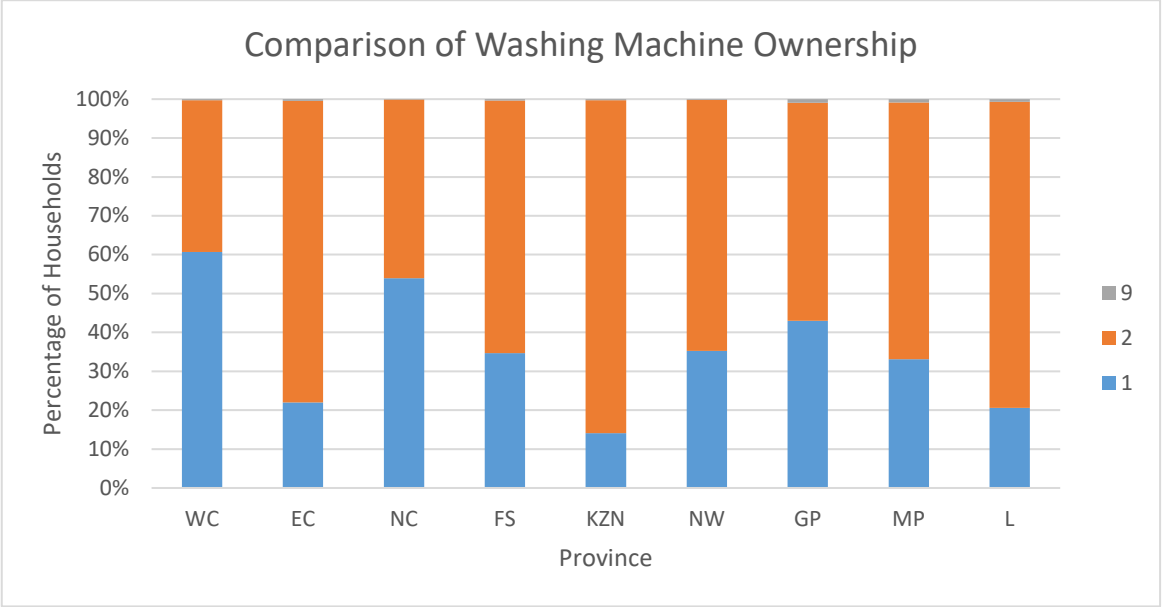

Figure 4-20: Comparison of washing machine ownership percentages per province

4.2.24 Fridge ownership

The Q821Fridge variable seeks to establish whether a household owns a fridge, where option 1 is yes; 2 no; and 9 unspecified. The results show that the response with the highest percentage is yes. The results in Figure 4-21 reveal that the highest percentage of households that own a fridge is found in the Western Cape (WC) with 87.52 percent while the Eastern Cape (EC) has the lowest with 66.95 percent.

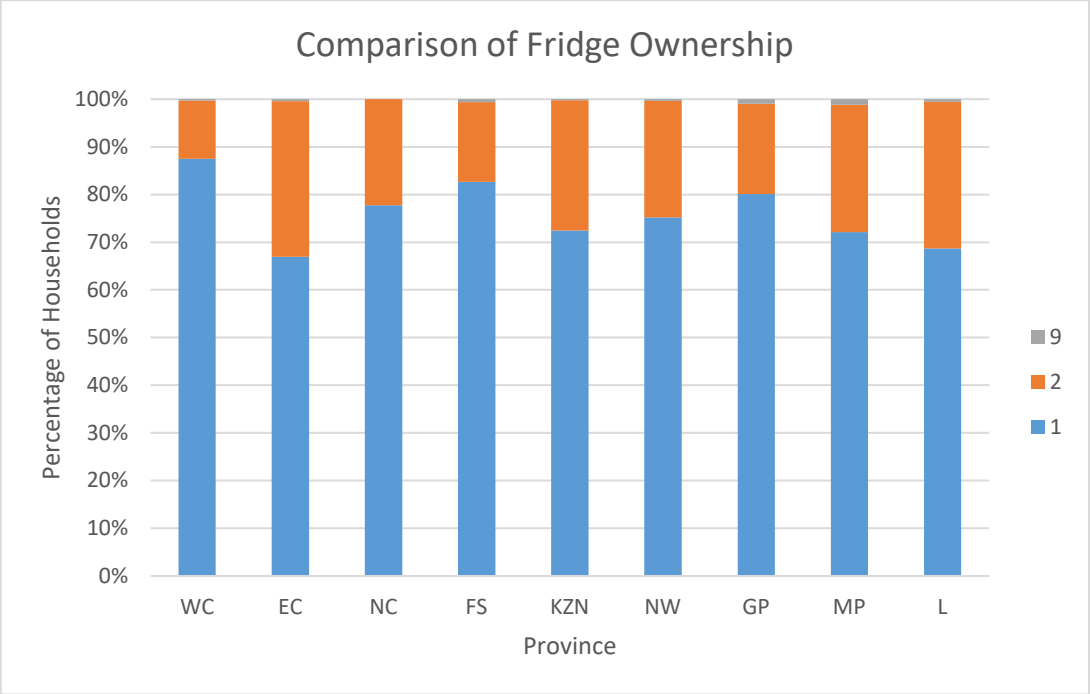

Figure 4-21: Comparison of fridge ownership percentages per province

4.2.25 Electric stove ownership

The Q821EStove variable seeks to establish whether a household owns an electronic stove, where option 1 is yes; 2 no; and 9 unspecified. The results in Figure 4-22 show that the response with the highest percentage is yes. The highest percentage of households that own a fridge are found in the Western Cape (WC) with 95.98 percent, while Limpopo (L) has the lowest with 81.32 percent.

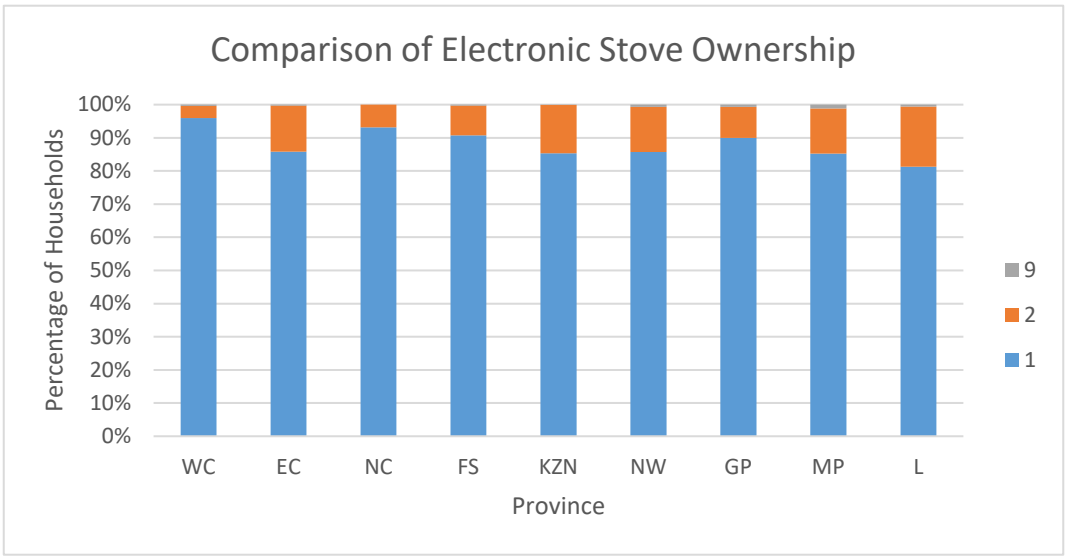

Figure 4-22: Comparison of electronic stove ownership percentages per province

4.2.26 Microwave

The Q821MicroW variable seeks to establish whether a household owns a microwave, where option 1 is yes; 2 no; and 9 unspecified.

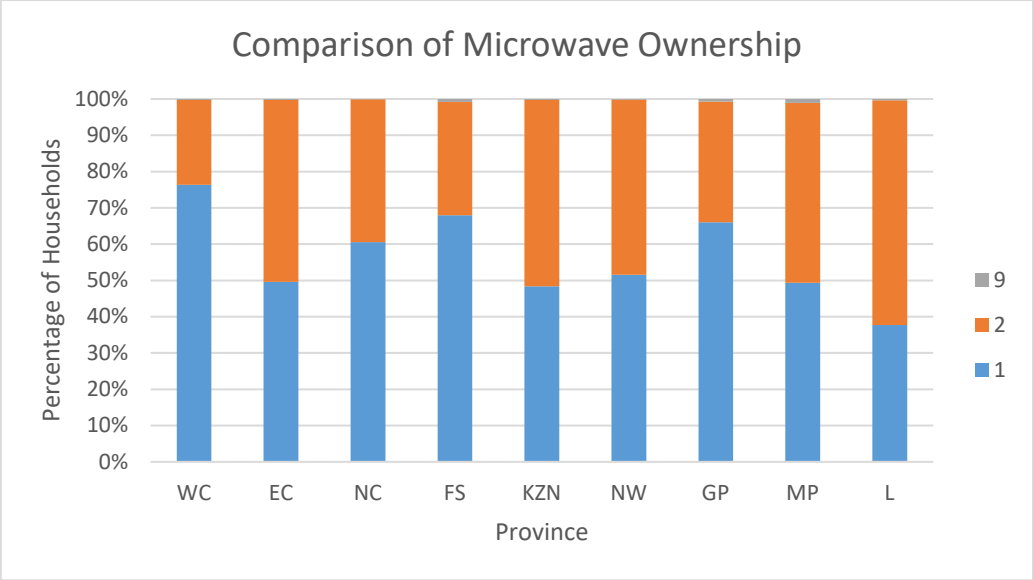

Figure 4-23: Comparison of microwave ownership percentages per province

The results in Figure 4-23 show that the highest percentage of households that own a microwave is found in the Western Cape (WC) with 76.37 percent, while the Limpopo (L) has the lowest with 37.71 percent.

4.2.27 Home theatre system ownership

The Q821HomeTH variable seeks to establish whether a household owns a home theatre system, where option 1 is yes; 2 no; and 9 unspecified. The results in Figure 4-24 show that response with the highest percentage is no. The highest percentage of households that do not own a home theatre system is found in the Limpopo (L) with 92.54 percent, while the Gauteng Province (GP) has the lowest with 75.30 percent.

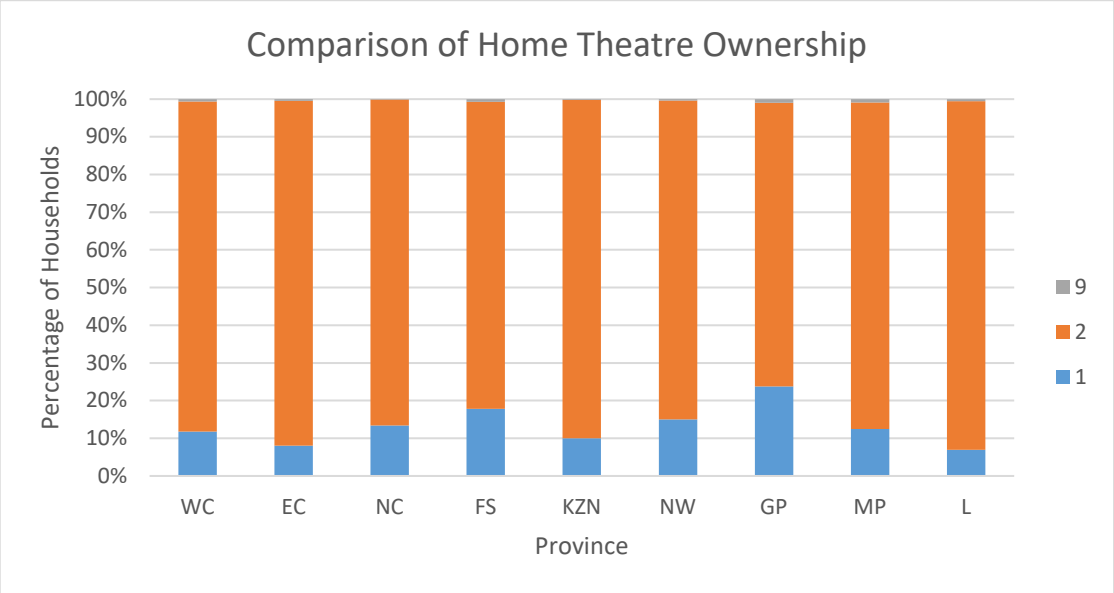

Figure 4-24: Comparison of home theatre ownership percentages per province

4.2.28 Number of household members

The hholdsz variable seeks to identify household members sharing resources in the same household. The number of household members ranges between 1 and 23. The results in Table 4-4 reveal that provinces that have the average of 3 household members that share the same resources are Western Cape (WC), Free State (FS), North West (NW) and Gauteng Province (GP) while the rest of the provinces have an average of 4.

Table 4-4: Number of household members' descriptive statistics

	WC	EC	NC	FS	KZN	NW	GP	MP	L
Mean	3	4	4	3	4	3	3	3	4
Standard Error	0,04	0,05	0,08	0,05	0,05	0,06	0,03	0,06	0,05
Median	3	3	3	3	3	3	3	3	3
Mode	2	1	1	2	1	1	1	1	1
Standard Deviation	1,93	2,45	2,39	1,99	2,83	2,23	2,01	2,40	2,36
Sample Variance	3,74	6,00	5,72	3,95	8,00	4,95	4,05	5,76	5,55
Kurtosis	4,47	2,40	2,30	1,89	3,34	1,61	2,29	2,17	1,52
Skewness	1,35	1,33	1,29	1,12	1,48	1,25	1,26	1,26	1,09
Range	18	17	15	15	21	13	17	18	15
Minimum	1	1	1	1	1	1	1	1	1
Maximum	19	18	16	16	22	14	18	19	16

	WC	EC	NC	FS	KZN	NW	GP	MP	L
Sum	6919	9858	3299	4267	12800	4524	15777	6089	8204
Count	2091	2108	916	1309	3358	1442	5037	1762	2307

4.2.29 Number of economically active household

The econact_hh seeks to establish the number of economically active household members. The variable ranges between 0 to 6, and 99 denotes unspecified. The results in Table 4-5 show that the average number of economically active members in a household is 1 across all provinces.

Table 4-5: Number of economically active household members' descriptive statistics

	WC	*EC*	*NC*	*FS*	*KZN*	*NW*	*GP*	*MP*	*L*
Mean	1	1	1	1	1	1	1	1	1
Standard Error	0,02	0,01	0,03	0,02	0,02	0,02	0,01	0,02	0,02
Median	1	1	1	1	1	1	1	1	1
Mode	1	1	1	1	1	1	1	1	1
Standard Deviation	0,97	0,79	0,86	0,81	0,87	0,76	0,83	0,79	0,76
Sample Variance	0,94	0,62	0,75	0,65	0,76	0,58	0,69	0,62	0.58
Kurtosis	1,44	0,91	0,58	1,37	1,69	0,98	1,15	0,48	0,78
Skewness	0,78	0,93	0,75	0,87	1,02	0,77	0,66	0,66	0,87
Range	7	5	5	5	5	5	7	4	4
Minimum	0	0	0	0	0	0	0	0	0
Maximum	7	5	5	5	5	5	7	4,00	4
Sum	2801	2146	871	1162	3163	1220	5839	1636	1745
Count	2091	2806	916	1309	3358	1442	5036	1762	2306

4.2.30 GeoType

The GeoType variable shows the classification of geography type settlement, where option 1 is urban formal; 2 traditional; and 3 farms. The results in Figure 4-25 reveal that the province with the highest percentage of urban settlement is Gauteng Province (GP) with 97.20 percent, the province with the highest percentage of traditional settlement is Limpopo (L) with 73.21 percent and the province with the highest percentage of farm settlements is KwaZulu Natal (KZN) with 7.39 percent.

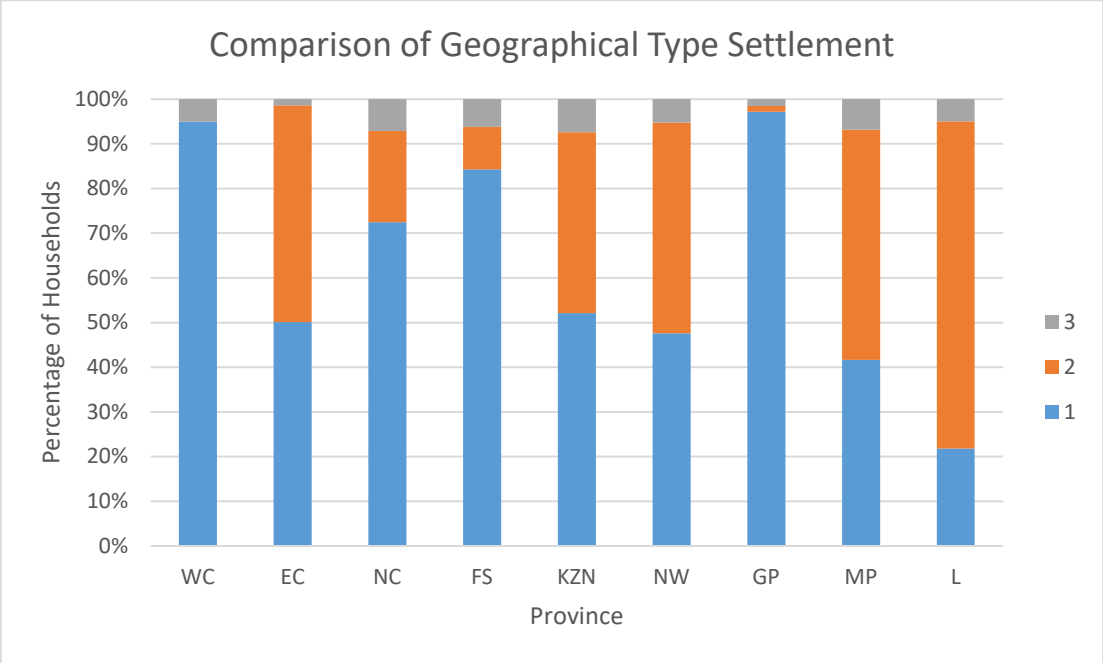

Figure 4-25: Comparison of geographical type settlement percentages per province

4.2.31 Metro name

The metro_code seeks to establish the number of households that stay in the metros. The results in Table 4-6 show that the Nelson Mandela Bay has the highest number of households of 1752 while eThekwini has the least percentage of 13.80 percent.

Table 4-6: Number of households per metro

Metro_Name	*Households*
City of Cape Town	1431
eThekwini	383
City of Johannesburg	496
Mangaung	366
Ekurhuleni	1091
City of Tshwane	1420
Nelson Mandela Bay	1752
Buffalo City	1068

4.3 Binary logistics regression

4.3.1 Introduction

The logistic regression model that is discussed in this section is developed for the South African population taking into account all the nine provinces. The response variable is Response_Y with two levels denoted as 0 and 1. A cut-off value of R 9 908,00, which is the average value of household net income in South Africa in 2015 as reflected in Table 4-1, has been used to divide the response variable into two groups. The responses with values less than or equal to this cut-off value are regarded as high risk and responses greater that this cut-off value are regarded as low risk. The model is a binary logit generated from Fisher's scoring optimising technique. The probability that consumers that apply for credit without any credit history are categorised as low risk is being modelled.

4.3.2 Forward selection results

In forward selection, an attempt is made to select variables that are significant into the model as explained in Section 3.3.4. Each addition of a variable to a model is listed in separate steps in a displayed output, and at each step a new model is fitted. Details of the forward selection steps are shown in Appendix A, Table A-2.

In Step 0, the intercept only model is fitted and score statistics for the potential variables are evaluated. In Step 1, the variable Q814Exp_10 is selected into the model. The model fit statistics section shows the AIC, SC and -2LogL for both the intercept only model and fitted model, and there seems to be a small difference in the values of these statistics. AIC and SC are used to compare different models, and models with smaller values are the preferred ones. Results of the likelihood ratio test and the efficient score test for testing the joint significance of the explanatory variables are contained in the testing global null hypothesis: B=0 section. Section 3.3.7 has already explained the likelihood ratio, score and Wald tests. In these tests, p-value is compared with $\mathrm{p} = 0.05$ to determine whether the hypothesis that all slope parameters are equal to zero is rejected. The p-values are smaller than $\mathrm{p} = 0.05$, therefore, the hypothesis that all slope parameters are equal to zero is rejected. In the forward selection, since the entry of the effect has been specified at 0.05 then the effect into the model, one at the time, is added until the residual chi-square becomes insignificant. That is, until the p-value of the residual chi-square exceeds the entry value of 0.05. The residual chi-square test is where the entered variable is evaluated. The p-value is less than $\mathrm{p} = 0.05$, therefore, the residual chi-square is significant. The same logic of interpretation applied for step 1 can be applied in all steps.

Results of the forward selection method are summarised in Table 4-7. In the study, it is assumed that when a specific predictor (independent) variable's effect on the response (dependent) variable is explained, all other variables are constant. The variables are sorted in order of importance based on their score chi-square value. This summary is consistent with the previous analysis of the forward selection method in which in all steps the selected variables are found to be statistically significant.

Table 4-7: Forward selection summary

Summary of Forward Selection					
Step	***Effect Entered***	***DF***	***Number In***	***Score Chi-Square***	***Pr > ChiSq***
1	Q814Exp_10	1	1	3014.3074	<.0001
2	Q814Exp_9	1	2	646.2212	<.0001
3	Q814Exp_11	1	3	339.8954	<.0001
4	Q814Exp_8	1	4	178.6897	<.0001
5	metro_code_13	1	5	158.5154	<.0001
6	Q814Exp_12	1	6	128.6345	<.0001
7	metro_code_14	1	7	66.9257	<.0001
8	metro_code_4	1	8	71.0288	<.0001
9	metro_code_7	1	9	67.8025	<.0001
10	Q821MicroW_2	1	10	51.2625	<.0001
11	MP	1	11	48.9213	<.0001
12	metro_code_9	1	12	48.8813	<.0001
13	Q89bMain_5	1	13	41.7608	<.0001
14	metro_code_5	1	14	33.5479	<.0001
15	Q58Val_1	1	15	31.7143	<.0001
16	Q814Exp_99	1	16	25.9815	<.0001
17	Sex	1	17	23.7145	<.0001
18	Q510aRDP_2	1	18	21.2865	<.0001
19	metro_code_15	1	19	20.3613	<.0001
20	Q821HomeTH_1	1	20	17.7679	<.0001
21	Q89bMain_3	1	21	17.2580	<.0001
22	Q65Int1_2	1	22	14.3655	0.0002
23	hholdsz	1	23	16.2061	<.0001
24	metro_code_1	1	24	12.2063	0.0005
25	Q814Exp_4	1	25	11.6763	0.0006
26	Q58Val_6	1	26	11.5382	0.0007
27	Q820Happy_3	1	27	10.5411	0.0012
28	Q56Owner_6	1	28	9.6614	0.0019
29	econact_hh	1	29	9.9018	0.0017
30	Q58Val_3	1	30	6.8424	0.0089
31	Q89bMain_1	1	31	6.6928	0.0097

Summary of Forward Selection					
Step	**Effect Entered**	**DF**	**Number In**	**Score Chi-Square**	**Pr > ChiSq**
32	Q58Val_8	1	32	6.9276	0.0085
33	Q821WashM_2	1	33	6.2462	0.0124
34	Q814Exp_7	1	34	5.4907	0.0191
35	Q820Happy_1	1	35	5.2795	0.0216
36	I_A	1	36	4.8238	0.0281
37	GeoType_2	1	37	5.4262	0.0198
38	Q58Val_5	1	38	4.2393	0.0395
39	Q58Val_4	1	39	5.1951	0.0227
40	Q56Owner_1	1	40	3.8560	0.0496

4.3.3 Goodness-of-fit tests

Table 4-8: Deviance and Pearson goodness-of-fit statistics

Deviance and Pearson goodness of fit statistics				
Criterion	**Value**	**DF**	**Value/DF**	**Pr > ChiSq**
Deviance	6847.8930	1E4	0.6538	1.0000
Pearson	10153.3537	1E4	0.9694	0.9873

Table 4-8 show the deviance and Pearson chi-square statistic. The goodness-of-fit statistic is statistically significant since p-value is greater than $p = 0.05$,as explained in Section 3.3.9. Therefore, the model fits the data. Table 4-9 shows the effect of variables. All variables are statistically significant at 0.05 level because they all have p-values that are less than 0.05 except GeoType_2.

Table 4-9: Analysis of Effects

Type 3 Analysis of Effects			
Effect	*DF*	*Wald Chi-Square*	*Pr > ChiSq*
hholdsz	1	15.0261	0.0001
econact_hh	1	5.1880	0.0227
MP	1	57.1213	<.0001
I_A	1	6.3695	0.0116
Sex	1	14.8295	0.0001

Type 3 Analysis of Effects			
Effect	*DF*	*Wald Chi-Square*	*Pr > ChiSq*
Q56Owner_1	1	3.8535	0.0496
Q56Owner_6	1	6.2740	0.0123
Q58Val_1	1	23.8058	<.0001
Q58Val_3	1	1.7341	0.1879
Q58Val_4	1	6.2479	0.0124
Q58Val_5	1	7.4630	0.0063
Q58Val_6	1	16.1744	<.0001
Q58Val_8	1	9.4435	0.0021
Q510aRDP_2	1	14.9566	0.0001
Q65Int1_2	1	4.1214	0.0423
Q89bMain_1	1	7.9579	0.0048
Q89bMain_3	1	5.3074	0.0212
Q89bMain_5	1	17.4948	<.0001
Q814Exp_4	1	6.6000	0.0102
Q814Exp_7	1	5.1542	0.0232
Q814Exp_8	1	62.1372	<.0001
Q814Exp_9	1	212.4397	<.0001
Q814Exp_10	1	593.9687	<.0001
Q814Exp_11	1	180.2132	<.0001
Q814Exp_12	1	29.8104	<.0001
Q814Exp_99	1	16.4488	<.0001
Q820Happy_1	1	6.5814	0.0103
Q820Happy_3	1	4.4051	0.0358
Q821WashM_2	1	5.3125	0.0212
Q821MicroW_2	1	14.9908	0.0001
Q821HomeTH_1	1	10.5658	0.0012
GeoType_2	1	2.9297	0.0870
metro_code_1	1	16.1048	<.0001
metro_code_4	1	82.4348	<.0001
metro_code_5	1	32.4125	<.0001
metro_code_7	1	88.2026	<.0001
metro_code_9	1	51.2533	<.0001
metro_code_13	1	178.0180	<.0001
metro_code_14	1	65.7473	<.0001
metro_code_15	1	11.5245	0.0007

Table 4-10 shows that all parameters are statistically significant at 0.05 level because they all have p-values that are smaller than 0.05, except GeoType_2.

Table 4-10: Analysis of maximum likelihood estimates

Analysis of Maximum Likelihood Estimates						
Parameter		*DF*	*Estimate*	*Standard Error*	*Wald Chi-Square*	*Pr > ChiSq*
Intercept		1	-5.4516	0.6307	74.7183	<.0001
hholdsz		1	-0.0577	0.0149	15.0261	0.0001
econact_hh		1	-0.0339	0.0149	5.1880	0.0227
MP	0	1	0.4387	0.0580	57.1213	<.0001
I_A	0	1	-0.2364	0.0936	6.3695	0.0116
Sex	0	1	0.1303	0.0338	14.8295	0.0001
Q56Owner_1	0	1	0.0824	0.0420	3.8535	0.0496
Q56Owner_6	0	1	-0.1328	0.0530	6.2740	0.0123
Q58Val_1	0	1	-0.2190	0.0449	23.8058	<.0001
Q58Val_3	0	1	-0.0623	0.0473	1.7341	0.1879
Q58Val_4	0	1	0.1377	0.0551	6.2479	0.0124
Q58Val_5	0	1	0.2439	0.0893	7.4630	0.0063
Q58Val_6	0	1	0.4964	0.1234	16.1744	<.0001
Q58Val_8	0	1	0.4699	0.1529	9.4435	0.0021
Q510aRDP_2	0	1	0.1757	0.0454	14.9566	0.0001
Q65Int1_2	0	1	-0.1099	0.0541	4.1214	0.0423
Q89bMain_1	0	1	0.1215	0.0431	7.9579	0.0048
Q89bMain_3	0	1	-0.2009	0.0872	5.3074	0.0212
Q89bMain_5	0	1	-0.2766	0.0661	17.4948	<.0001
Q814Exp_4	0	1	-0.3840	0.1495	6.6000	0.0102
Q814Exp_7	0	1	0.1608	0.0708	5.1542	0.0232
Q814Exp_8	0	1	0.4743	0.0602	62.1372	<.0001
Q814Exp_9	0	1	0.9258	0.0635	212.4397	<.0001
Q814Exp_10	0	1	1.6991	0.0697	593.9687	<.0001
Q814Exp_11	0	1	1.1104	0.0827	180.2132	<.0001
Q814Exp_12	0	1	1.0915	0.1999	29.8104	<.0001
Q814Exp_99	0	1	0.7619	0.1879	16.4488	<.0001
Q820Happy_1	0	1	-0.0944	0.0368	6.5814	0.0103
Q820Happy_3	0	1	0.0839	0.0400	4.4051	0.0358
Q821WashM_2	0	1	-0.0919	0.0399	5.3125	0.0212
Q821MicroW_2	0	1	-0.1619	0.0418	14.9908	0.0001
Q821HomeTH_1	0	1	0.1328	0.0409	10.5658	0.0012
GeoType_2	0	1	-0.0815	0.0476	2.9297	0.0870
metro_code_1	0	1	-0.3728	0.0929	16.1048	<.0001
metro_code_4	0	1	0.9337	0.1028	82.4348	<.0001
metro_code_5	0	1	-0.9105	0.1599	32.4125	<.0001
metro_code_7	0	1	0.6806	0.0725	88.2026	<.0001
metro_code_9	0	1	0.4131	0.0577	51.2533	<.0001
metro_code_13	0	1	0.7690	0.0576	178.0180	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
metro_code_14	0	1	0.4398	0.0542	65.7473	<.0001
metro_code_15	0	1	0.2257	0.0665	11.5245	0.0007

Table 4-11 shows the p-value of 0.2013, which is greater than $p = 0.05$. Therefore, as explained in Section 3.3.9 the null hypothesis that the model is adequate is not rejected.

Table 4-11: Hosmer and Lemeshow goodness of fit test

Hosmer and Lemeshow Goodness of Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
11.0063	8	0.2013

Table 4-12: Partition for Hosmer and Lemeshow test

Partition for the Hosmer and Lemeshow Test					
Group	**Total**	**Response_Y = 0**		**Response_Y = 1**	
		Observed	**Expected**	**Observed**	**Expected**
1	1052	183	160.87	869	891.13
2	1052	404	433.39	648	618.61
3	1052	715	715.34	337	336.66
4	1052	852	859.43	200	192.57
5	1052	936	933.40	116	118.60
6	1052	976	974.92	76	77.08
7	1053	1004	1002.41	49	50.59
8	1052	1020	1019.31	32	32.69
9	1052	1040	1031.69	12	20.31
10	1046	1037	1036.25	9	9.75

The classification table in Table 4-13 shows the classification of the input binary response one and zero according to whether the predicted event or non-event probabilities are above or below specified cut-off values (“Prob level”) 0.3, 0.4, 0.5, 0.6 and 0.7. This table indicates that the non-event is where Response_Y = 0 (high risk) and the event is where Response_Y = 1 (low risk). The results of the classification table reveal that at a cut-off value of 0.5, only 7 689 events are correctly classified, 1 401 non-events are correctly classified and the correct classification rate is 86.4 percent, which is the highest rate compared to the rates given by other cut-off values. For

the cut-off value of 0.3, only 7 940 events are correctly classified, 950 non-events are correctly classified and the correct classification rate is 84.5 percent, which is the lowest rate compared to the rates given by other cut-off values. Therefore the cut-off value of 0.5 provides the overall best correct classification rate.

Table 4-13: Classification table

Classification Table									
Prob Level	***Correct***		***Incorrect***		***Percentages***				
	Event	***Non-Event***	***Event***	***Non-Event***	***Correct***	***Sensi-tivity***	***Speci-ficity***	***False POS***	***False NEG***
0.300	7940	950	1398	227	84.5	97.2	40.5	15.0	19.3
0.400	7812	1205	1143	355	85.8	95.7	51.3	12.8	22.8
0.500	7689	1401	947	478	86.4	94.1	59.7	11.0	25.4
0.600	7512	1551	797	655	86.2	92.0	66.1	9.6	29.7
0.700	7192	1731	617	975	84.9	88.1	73.7	7.9	36.0

Figure 4-26 shows the results of overlaid curves of all models in the same plot. In Section 3.2.9 this curve was explained. The final model has an area under the curve of 0.894 as reflected on the c statistic in Table 4-14, which is the estimate by the statistic explained in 3.2.9 .The value of 0.894 means that the model has a high predictive power and is a good classifier since it is close to 1.

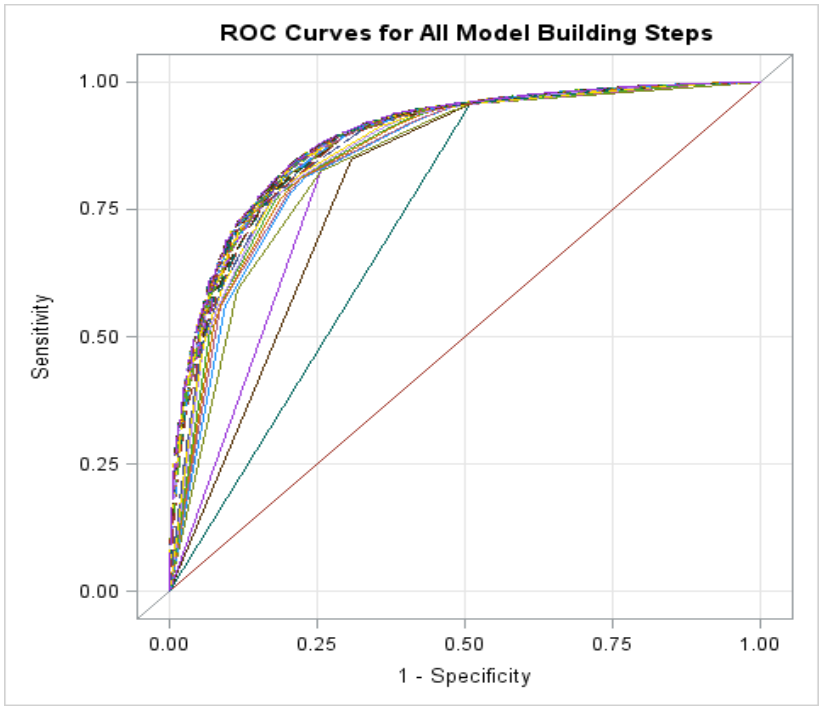

Figure 4-26: ROC curves for all model building steps

Table 4-14: Association of predicted probabilities and observed responses

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.4	Somers' D	0.787
Percent Discordant	10.6	Gamma	0.787
Percent Tied	0.0	Tau-a	0.273
Pairs	19176116	c	0.894

4.4 Validation of the model

In Section 3.3.4 it is mentioned that for the purpose of validation, quantities of the model build with 50 percent of the training dataset and the model fitted on the testing dataset, will be compared to check if there is consistency in their values. The values in the confusion matrix are taken from the classification in Table 4-13. All the results of the fitted model are recorded in Appendix A from Table A2 to Table A9.

Table 4-15: Confusion matrix for training dataset

	Observed	
Predicted	94.1	11
	25.4	59.7

Table 4-16: Confusion matrix for testing dataset

	Observed	
Predicted	94.3	11
	24.2	60.2

The results of the confusion matrix for both training and testing in Table 4-15 and Table 4-16 show that there is little difference in the rates between the observed and predicted observations. The precision rate (PVV) is $\frac{94.1}{94.1+11}*100=89.53$ and $\frac{94.3}{94.3+11}*100=89.55$ for the training dataset and the testing dataset, respectively.

The accuracy rate (ACC) is $\frac{94.1+59.7}{94.1+11+59.7+25.4}*100=80.86$ and $\frac{94.3+60.2}{94.3+11+60.2+24.2}*100=81.44$ for the training dataset and the testing dataset respectively.

The F-score (F) is $\frac{2\times 94.1}{2\times 94.1+11+25.4}*100=83.79$ and $\frac{2\times 94.3}{2\times 94.3+11+24.2}*100=84.27$

for the training dataset and the testing dataset, respectively.

The results from these calculated quantities show that there is 0.02 percent precision rate difference, 0,58 percent accuracy difference and 0.48 percent F-score difference between the training dataset and the testing dataset, therefore, the model is adequate for predicting Response_Y= '0'.

Furthermore, the conclusion of model adequacy reached from the results of the deviance and Pearson goodness-of-fit statistics in Table 4-8, is also confirmed for the fitted model in Table A-3 since the deviance and Pearson statistics do not reject the null hypothesis of good model fit. The Hosmer and Lemeshow statistics in both training and testing models reach the same conclusion that these models are adequate, that is, the null hypothesis that the model is adequate is not rejected as revealed in Table 4-11 and Table A-8.

4.5 Summary

This chapter provided an analysis of all variables in all nine provinces presented in Chapter 3. Illustration of the application of the binary logistic regression model was given and results were interpreted. Significant variables were identified at a specified 95 percent significance. The model was validated and confirmed adequate for prediction.

CHAPTER 5: CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

The problem statement of this study is that consumers with no credit history are denied access to credit, and if they do access credit, they pay high amount of interest rates. A model was to be developed to score credit applicants without credit history.

The study addressed different objectives, namely to provide a literature review of existing methodologies for credit scoring models developed in logistic regression and comparing it with other statistical methods, to develop the credit scoring model to categorise credit applicants into high risk and low risk, to test the relevance of the proposed methodology and to construct the datasets needed to address these objectives. In Chapter 2, the literature review was addressed. The results and interpretation of the developed methodology that was explained in Chapter 3 are documented in Chapter 4. The key findings and conclusions are highlighted in this chapter and recommendations and suggestions for future research are provided.

5.2 Key findings

Chapter 2 documented 11 examples of credit scoring development methodologies and other information regarding credit scoring methodologies relevant to the study. Insight gained from the discussion includes general knowledge regarding the type of data, variables and statistical methods that are considered in credit scoring methodologies. This knowledge was used to identify variables that can be used to develop the credit scoring model. Furthermore, the logistic regression model was compared to other statistics methods. The logistic regression model was able to classify people into specified groups.

The data that were used to develop the credit scoring model was discussed in Chapter 3. Our study used the General Household Survey (GHS) data sourced from Stats SA. The GHS is an annual survey introduced in 2002, and collects information from South African households on: education, health and social development, housing, household access to services and facilities, food and security and agriculture. Attention was given to variables selected from this dataset. It should be noted that for the sake of the study, discriminatory variables were included and used to build the logistic regression model. However, it is advisable not to include them when scoring applicants as stated in the National Credit Act 34 of 2005. The act states that assessment based on the ground of unfair discrimination is prohibited based on section 9(3) of the Constitution. A continuous dependent variable used was used, which was then converted to a binary variable based on a specified cut-off value of R10 169,12 which is the average net income household of our data. The categorised variables were then quantified such that they represent a dummy

variable. For example, the province variable has nine categories, representing the nine provinces in South Africa. If for example, Mpumalanga were to be chosen, 1 would denote the event that Mpumalanga is true and 0 would denote the event that Mpumalanga is false. The same resemblance has been applied in all variables.

In Chapter 4, the methodology used to develop the logistic regression model is documented as well as the results of the model developed. The data were divided into 50:50 for training and testing to validate the model. The training data were then used to build the model and the forward selection method provided a summary of significant variables that can be used to score applicants. Interestingly, included in the summary of significant variables are discriminatory variables sex (gender), Indian and Asian (race), expenditure, property value, number of household members, number of economic active household members and happiness, to name a few. Furthermore, the model was tested for adequacy and its predictive probability. The model was discovered to be adequate.

5.3 Recommendations

It is recommended that the models should only be used for consumers that have never had a credit history. A consumer should be investigated to clarify this issue.

Possible future research may involve the application of the knowledge gained from this dissertation to develop binary logistic regression models for consumers who do not have credit history.

Due to the nature of the data used, future studies should be aimed at collecting more data from consumers who do not have credit history and compare statistical models or a hybrid of these models to improve the accuracy of the scoring models by taking into account factors that were not considered in this research. This will assist in understanding the behaviours of these consumers.

Future studies should be aimed at improving the selection of variables before the model build so that there will be few variables used.

In addition, future studies aimed at investigating the reduction of significant variables when a model is fitted on a new data should be carried out and the implication thereof should be studied.

BIBLIOGRAPHY

ABDOU, H., POINTON, J. & EL-MASRY, A. 2008. Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Systems with Applications,* 35**,** 1275-1292.

ABDOU, H. A. & POINTON, J. 2011. Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management,* 18**,** 59-88.

ABDOU, H. A., TSAFACK, M. D. D., NTIM, C. G. & BAKER, R. D. 2016. Predicting creditworthiness in retail banking with limited scoring data. *Knowledge-Based Systems,* 103**,** 89-103.

ABID, L., MASMOUDI, A. & ZOUARI-GHORBEL, S. 2016. The consumer loan's payment default predictive model: an application of the logistic regression and the discriminant analysis in a Tunisian commercial bank. *Journal of the Knowledge Economy***,** 1-15.

AGRESTI, A. 2018. *An introduction to categorical data analysis, Wiley.*

ANDREEVA, G. 2006. European generic scoring models using survival analysis. *Journal of the Operational research Society,* 57**,** 1180-1187.

AVERY, R. B., BOSTIC, R. W., CALEM, P. S. & CANNER, G. B. 1996. Credit risk, credit scoring, and the performance of home mortgages. *Fed. Res. Bull.,* 82**,** 621.

BANASIK, J. & CROOK, J. 2007. Reject inference, augmentation, and sample selection. *European Journal of Operational Research,* 183**,** 1582-1594.

BANASIK, J., CROOK, J. & THOMAS, L. 2003. Sample selection bias in credit scoring models. *Journal of the Operational Research Society,* 54**,** 822-832.

BELLOTTI, T. & CROOK, J. 2009a. Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society,* 60**,** 1699-1707.

BELLOTTI, T. & CROOK, J. 2009b. Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications,* 36**,** 3302-3308.

BENSIC, M., SARLIJA, N. & ZEKIC-SUSAC, M. 2005. Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intelligent Systems in Accounting, Finance and Management,* 13**,** 133-150.

BOLTON, C. 2010. Logistic regression and its application in credit scoring.

BORITZ, J. E. & KENNEDY, D. B. 1995. Effectiveness of neural network types for prediction of business failure. *Expert Systems with Applications,* 9**,** 503-512.

BOSLAUGH, S. 2007. An introduction to secondary data analysis. *Secondary data sources for public health: a practical guide***,** 2-10.

CAIRE, D. 2004. Building Credit Scorecards for Small Business Lending in Developing Markets. *Bannock Consulting,* *http://www.microfinance.com/English/Papers/Scoring_SMEs_Hybrid. pdf.*

CASKEY, J. P. 2002. Bringing unbanked households into the banking system.

CAVENDISH, B. 2009. The Value of a Credit Score :Developing an Equitable Model for the Use of Credit Histories in Financially Underserved Communities.

CHEN, M.-C. & HUANG, S.-H. 2003. Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications,* 24**,** 433-441.

CRONE, S. F. & FINLAY, S. 2012. Instace sampling in credit scoring: An empirical study of sample size and balancing, *International Journal of Forecasting, 28, 224-238.*

DE LA REY, T. 2007. *Two statistical problems related to credit scoring/Tanja de la Rey.* PhD in Risk Analysis, North-West University.

DESAI, V. S., CROOK, J. N. & OVERSTREET, G. A. 1996. A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research,* 95**,** 24-37.

EISENBEIS, R. A. 1978. Problems in Applying Discriminant Analysis in Credit Scoring Models.
FLETCHER, D. Forecasting with neural networks. *Information & Management,* 24**,** 59-167.
FOURIE, E. 2015. *Review of subnational credit rating methodologies and their applicability in South Africa.* Philosophiae Doctor in Risk Analysis, North-West University.
GREENE, W. 1998. Sample selection in credit-scoring models. *Japan and the world economy,* 10**,** 299-316.
HAND, D. J., SOHN, S. Y. & KIM, Y. 2005. Optimal bipartite scorecards. *Expert Systems with Applications,* 29**,** 684-690.
HSIEH, N.-C. 2004. An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert systems with applications,* 27**,** 623-633.
HUBER, P. J. From large to huge: A statistician's reactions to KDD & DM. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, 1997. AAAI Press, 304-308.
HUNG, S.-Y., YEN, D. C. & WANG, H.-Y. 2006. Applying data mining to telecom churn management. *Expert Systems with Applications,* 31**,** 515-524.
JUPITER, D. C. 2013. Logistic models—an odd (s) kind of regression. *The Journal of Foot and Ankle Surgery,* 52**,** 279-280.
KIVEU, C. W. 2015. *Appraisal of credit application using logistic regression and linear discriminant models with pricnipal component analysis.* MASTERS IN SOCIAL STATISTICS, UNIVERSITY OF NAIROBI.
LANDAJO, M., DE ANDRÉS, J. & LORCA, P. 2007. Robust neural modeling for the cross-sectional analysis of accounting information. *European Journal of Operational Research,* 177**,** 1232-1252.
LEE, T.-S. & CHEN, I.-F. 2005. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications,* 28**,** 743-752.
LEE, T.-S., CHIU, C.-C., LU, C.-J. & CHEN, I.-F. 2002. Credit scoring using the hybrid neural discriminant technique. *Expert Systems with applications,* 23**,** 245-254.
LENARD, M. J., ALAM, P. & MADEY, G. R. 1995. The application of neural networks and a qualitative response model to the auditor's going concern uncertainty decision. *Decision Sciences,* 26**,** 209-227.
LIMSOMBUNCHAI, V., GAN, C. & LEE, M. 2005. An Analysis of Credit Scoring for Agricultural Loans in Thailand *American Journal of Applied Sciences,* 2**,** 1198-1205.
LIU, Y. 2001. New issues in credit scoring application. *Work report No,* 16.
MANSOURI, S. & DASTOORI, M. 2013. Credit Scoring Model for Iranian Banking Customers and Forecasting Creditworthiness of Borrowers. *International Business Research,* 6**,** 25.
MESTER, L. J. 1997. What's the point of credit scoring? *Business review,* 3**,** 3-16.
NETER, J. 2004. *Applied linear statistical models,* McGraw-Hill/Irwin.
ONG, C.-S., HUANG, J.-J. & TZENG, G.-H. 2005. Building credit scoring models using genetic programming. *Expert Systems with Applications,* 29**,** 41-47.
ORGLER, Y. E. 1970. A credit scoring model for commercial loans. *Journal of Money, Credit and Banking,* 2**,** 435-445.
POTTS, W. J. & PATETTA, M. J. 2000. Predictive Modeling Using Logistic Regression: Course Notes. *SAS Institute*.
POWERS, D. M. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.
ŠARLIJA, N., BENŠIĆ, M. & BOHAČEK, Z. Multinomial model in consumer credit scoring. 10th International Conference on Operational Research (10; 2004), 2004.

SAS/STAT 2010. SAS/STAT 9.22 User's Guide: The LOGISTIC Procedure (book Excerpt). Cary, NC, USA: Books24x7. com.
SCHREINER, M. 2004. Benefits and pitfalls of statistical credit scoring for microfiinace/Ventajas y Desventajas del Scoring Estadístico para las Microfinanzas/Vertus et Faiblesses de l'Évaluation Statistique (Credit Scoring) en Microfinance. *Savings and Development***,** 63-86.
SIBANDA, W. & PRETORIUS, P. 2012. Comparative study of the application of box behnken design (BBD) and binary logistic regression (BLR) to study the effect of demographic characteristics on HIV risk in South Africa. *Journal of Applied Medical Sciences,* 1**,** 15-40.
SPEAR, N. A. & LEIS, M. 1997. Artificial neural networks and the accounting method choice in the oil and gas industry. *Accounting, Management and Information Technologies,* 7**,** 169-181.
STATSSA 2017. General Household Survey 2017. *In:* AFRICA, S. S. (ed.) 2017 ed. Pretoria: Statistics South Africa.
STEENACKERS, A. & GOOVAERTS, M. J. 1989. A credit scoring model for personal loans. *Insurance: Mathematics and Economics,* 8**,** 31-34.
STOLTZFUS, J. C. 2011. Logistic regression: a brief primer. *Academic Emergency Medicine,* 18, 1099-1104.
ŠUŠTERŠIČ, M., MRAMOR, D. & ZUPAN, J. 2009. Consumer credit scoring models with limited data. *Expert Systems with Applications,* 36**,** 4736-4744.
THABISO, P. M. 2014. Credit Scoring Techniques: A Survey.
The National Credit Act 34 of 2005.
VARTANIAN, T. P. 2010. *Secondary data analysis,* Oxford University Press.

APPENDIX A

Table A-1: Modified proposed variables for building the credit scoring model

i	Name	Label of Response	New Quantitative Variable (X_i)
1	Q511Subs	1 - Yes	Q511Subs_1
2		2 - No	Q511Subs_2
3		3 - Do not know	Q511Subs_3
4		9 - Unspecified	Q511Subs_9
5	Geotype	1 - Urban formal	GeoType_1
6		2 - Traditional	GeoType_2
7		3 - Farms	GeoType_3
8	head_popgrp	1 - Black	B
9		2 - Coloured	C
10		3 - Indian / Asian	I_A
11		4 - White	W
12	head_sex	1 - Male	0 if sex=male, 1 if sex=female
13		2 - Female	
14	Metro_code	WC – Non-Metro	metro_code_1
15		WC - City of Cape Town	metro_code_2
16		EC - Non-Metro	metro_code_3
17		EC - Buffalo City	metro_code_4
18		EC - Nelson Mandela Bay	metro_code_5
19		NC - Non-Metro	metro_code_6
20		FS - Non-Metro	metro_code_7
21		FS - Mangaung	metro_code_8
22		KZN - Non-Metro	metro_code_9
23		KZN - eThekwini	metro_code_10
24		NW - Non-Metro	metro_code_11
25		GP - Non-Metro	metro_code_12
26		GP - Ekurhuleni	metro_code_13
27		GP - City of Johannesburg	metro_code_14
28		GP - City of Tshwane	metro_code_15
29		MP - Non-Metro	metro_code_16
30		LP - Non-Metro	metro_code_17
31	Q510aRDP	1 - Yes	Q510aRDP_1
32		2 - No	Q510aRDP_2
33		3 - Do not know	Q510aRDP_3
34		9 - Unspecified	Q510aRDP_9
35	Q51MainD	1 - Dwelling/house or brick/concrete block structure on a separate stand or yard or on farm	Q51MainD_1
36		2 - Traditional dwelling/hut/traditional materials	Q51MainD_2
37		3 - Flat or apartment in a block of flats	Q51MainD_3
38		4 - Cluster house in complex	Q51MainD_4
39		5 - Town house (semi-detached house in complex)	Q51MainD_5
40		6 - Semi-detached house	Q51MainD_6

i	Name	Label of Response	New Quantitative Variable (X_i)
41		7 - Dwelling/house/flat/room in backyard	Q51MainD_7
42		8 - Informal dwelling/shack in backyard	Q51MainD_8
43		9 - Informal dwelling/shack not in backyard	Q51MainD_9
44		10 - Room/flat let on a property or a larger dwelling servants' quarters/granny flat	Q51MainD_10
45		11 - Caravan / tent	Q51MainD_11
46		12 - Other types of dwelling	Q51MainD_12
47	Q56Owner	1 - Rented from private individual	Q56Owner_1
48		2 - Rented from other	Q56Owner_2
49		3 - Owned but not yet paid off to bank/financial institution	Q56Owner_3
50		4 - Owned but not yet paid off to private banker	Q56Owner_4
51		5 - Owned and fully paid	Q56Owner_5
52		6 - Occupied rent-free	Q56Owner_6
53		7 - Other	Q56Owner_7
54		8 - Do not know	Q56Owner_8
55	Q58Val	1 - Less than R50 000	Q58Val_1
56		2 - R50 001 - R250 000	Q58Val_2
57		3 - R250 001 - R500 000	Q58Val_3
58		4 - R500 001 - R1 000 000	Q58Val_4
59		5 - R1 000 001 - R1 500 000	Q58Val_5
60		6 - R1 500 001 - R2 000 000	Q58Val_6
61		7 - R2 000 001 - R3 000 000	Q58Val_7
62		8 - More than R3 000 000	Q58Val_8
63		9 - Do not know	Q58Val_9
64		99 - Unspecified	Q58Val_99
65	Q61Phon	1 - Yes	Q61Phon_1
66		2 - No	Q61Phon_2
67		9 - Unspecified	Q61Phon_9
68	Q62Cell	1 - Yes	Q62Cell_1
69		2 - No	Q62Cell_2
70		9 - Unspecified	Q62Cell_9
71	Q65Int1	1 - Yes	Q65Int1_1
72		2 - No	Q65Int1_2
73		9 - Unspecified	Q65Int1_9
74	Q65Int7	1 - Yes	Q65Int7_1
75		2 - No	Q65Int7_2
76		9 - Unspecified	Q65Int7_9
77	Q65Int8	1 - Yes	Q65Int8_1
78		2 - No	Q65Int8_2
79		9 - Unspecified	
80	Q814Exp	1 R0	Q814Exp_1
81		2 R1 - R199	Q814Exp_2
82		3 R200 - R399	Q814Exp_3
83		4 R400 - R799	Q814Exp_4

i	Name	Label of Response	New Quantitative Variable (X_i)
84		5 R800 - R1 199	Q814Exp_5
85		6 R1 200 - R1 799	Q814Exp_6
86		7 R1 800 - R2 499	Q814Exp_7
87		8 R2 500 - R4 999	Q814Exp_8
88		9 R5 000 - R9 999	Q814Exp_9
89		10 R10 000 or more	Q814Exp_10
90		11 Do not know	Q814Exp_11
91		12 Refuse	Q814Exp_12
92		99 Unspecified	Q814Exp_99
93	Q820Happy	1 - Happier	Q820Happy_1
94		2 - The same	Q820Happy_2
95		3 - Less happy	Q820Happy_3
96		4 - Refuse to answer	Q820Happy_4
97		5 - Do not know	Q820Happy_5
98		9 - Unspecified	Q820Happy_9
99	Q821Comp	1 - Yes	Q821Comp_1
100		2 - No	Q821Comp_2
101		3 - Unspecified	Q821Comp_9
102	Q821DVD	1 - Yes	Q821DVD_1
103		2 - No	Q821DVD_2
104		3 - Unspecified	Q821DVD_9
105	Q821EStove	1 - Yes	Q821EStove_1
106		2 - No	Q821EStove_2
107		3 - Unspecified	Q821EStove_9
108	Q821Fridge	1 - Yes	Q821Fridge_1
109		2 - No	Q821Fridge_2
110		3 - Unspecified	Q821Fridge_9
111	Q821HomeTh	1 - Yes	Q821HomeTH_1
112		2 - No	Q821HomeTH_2
113		3 - Unspecified	Q821HomeTH_9
114	Q821MicroW	1 - Yes	Q821MicroW_1
115		2 - No	Q821MicroW_2
116		3 - Unspecified	Q821MicroW_9
117	Q821TV	1 - Yes	Q821TV_1
118		2 - No	Q821TV_2
119		3 - Unspecified	Q821TV_9
120	Q821WashM	1 - Yes	Q821WashM_1
121		2 - No	Q821WashM_2
122		3 - Unspecified	Q821WashM_9
123	Q89bMain	1 - Salaries/wages/commission	Q89bMain_1
124		2 - Incomes from a business	Q89bMain_2
125		3 - Remittances	Q89bMain_3
126		4 Pensions	Q89bMain_4
127		5 - Grants	Q89bMain_5
128		6 - Salaries of farm products and services	Q89bMain_6

i	Name	Label of Response	New Quantitative Variable (X_i)
129		7 - Other income sources	Q89bMain_7
130		8 - No income	Q89bMain_8
131		9 - Unspecified	Q89bMain_9
132	Province	Western Cape	WC
133		Eastern Cape	EC
134		Northern Cape	NC
135		Free State	FS
136		KwaZulu Natal	KZN
137		North West	NW
138		Gauteng	GP
139		Mpumalanga	MP
140		Limpopo - L	LP

Table A-2: Summary of Forward selection – Testing Data

Summary of Forward Selection					
Step	**Effect Entered**	**DF**	**Number In**	**Score Chi-Square**	**Pr > ChiSq**
1	Q814Exp_10	1	1	3088.4190	<.0001
2	Q814Exp_9	1	2	575.0004	<.0001
3	Q814Exp_11	1	3	345.1406	<.0001
4	Q814Exp_12	1	4	192.1659	<.0001
5	metro_code_13	1	5	177.0650	<.0001
6	Q814Exp_8	1	6	133.0912	<.0001
7	metro_code_14	1	7	120.0273	<.0001
8	metro_code_4	1	8	133.8324	<.0001
9	Q821MicroW_2	1	9	67.4785	<.0001
10	econact_hh	1	10	59.8030	<.0001
11	metro_code_7	1	11	61.6416	<.0001
12	MP	1	12	44.9517	<.0001
13	metro_code_9	1	13	44.7726	<.0001
14	metro_code_5	1	14	28.8391	<.0001
15	Q814Exp_7	1	15	27.3348	<.0001
16	Q89bMain_5	1	16	25.5320	<.0001
17	Q820Happy_1	1	17	27.4054	<.0001
18	metro_code_15	1	18	23.2910	<.0001
19	Q65Int1_2	1	19	21.9485	<.0001
20	Q58Val_1	1	20	11.4101	0.0007
21	Q89bMain_3	1	21	10.5241	0.0012
22	Q56Owner_1	1	22	8.5651	0.0034
23	Q820Happy_3	1	23	6.8378	0.0089
24	Q821HomeTH_1	1	24	6.6974	0.0097
25	Q814Exp_99	1	25	5.2473	0.0220
26	Sex	1	26	5.2834	0.0215
27	metro_code_1	1	27	5.1706	0.0230
28	GeoType_2	1	28	5.0632	0.0244
29	Q510aRDP_2	1	29	5.6392	0.0176
30	hholdsz	1	30	6.1416	0.0132
31	Q89bMain_1	1	31	5.0922	0.0240

Table A-3: Deviance and Pearson Goodness-of-Fit Statistics – Testing Data

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	**Value**	**DF**	**Value/DF**	**Pr > ChiSq**
Deviance	6651.6675	9626	0.6910	1.0000
Pearson	9389.4769	9626	0.9754	0.9567

Table A-4: Type 3 Analysis of Effects – Testing Data

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
hholdsz	1	6.4312	0.0112
econact_hh	1	9.3625	0.0022
MP	1	61.4018	<.0001
Q56Owner_1	1	5.6311	0.0176
Q58Val_1	1	14.6616	0.0001
Q510aRDP_2	1	6.7933	0.0092
GeoType_2	1	7.9519	0.0048
metro_code_1	1	5.6517	0.0174
metro_code_4	1	151.2929	<.0001
metro_code_5	1	27.6158	<.0001
metro_code_7	1	76.8004	<.0001
metro_code_9	1	49.9427	<.0001
metro_code_13	1	202.5779	<.0001
metro_code_14	1	107.1921	<.0001
metro_code_15	1	13.9294	0.0002
Sex	1	5.9171	0.0150
Q65Int1_2	1	18.8471	<.0001
Q89bMain_1	1	5.0871	0.0241
Q89bMain_3	1	3.9968	0.0456
Q89bMain_5	1	11.9964	0.0005
Q814Exp_7	1	23.6391	<.0001
Q814Exp_8	1	72.1953	<.0001
Q814Exp_9	1	254.6359	<.0001
Q814Exp_10	1	795.6429	<.0001
Q814Exp_11	1	217.9000	<.0001
Q814Exp_12	1	67.9987	<.0001
Q814Exp_99	1	5.0399	0.0248
Q820Happy_1	1	16.2292	<.0001
Q820Happy_3	1	6.9113	0.0086
Q821MicroW_2	1	42.2747	<.0001
Q821HomeTH_1	1	5.3503	0.0207

Table A-5: Analysis of Maximum Likelihood Estimates – Testing Data

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-6.2013	0.5312	136.2939	<.0001
hholdsz		1	-0.0396	0.0156	6.4312	0.0112
econact_hh		1	-0.1420	0.0464	9.3625	0.0022
MP	0	1	0.4693	0.0599	61.4018	<.0001
Q56Owner_1	0	1	0.0923	0.0389	5.6311	0.0176
Q58Val_1	0	1	-0.1651	0.0431	14.6616	0.0001
Q510aRDP_2	0	1	0.1170	0.0449	6.7933	0.0092
GeoType_2	0	1	-0.1339	0.0475	7.9519	0.0048
metro_code_1	0	1	-0.2142	0.0901	5.6517	0.0174
metro_code_4	0	1	1.1436	0.0930	151.2929	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
metro_code_5	0	1	-0.8486	0.1615	27.6158	<.0001
metro_code_7	0	1	0.6352	0.0725	76.8004	<.0001
metro_code_9	0	1	0.3967	0.0561	49.9427	<.0001
metro_code_13	0	1	0.7917	0.0556	202.5779	<.0001
metro_code_14	0	1	0.5584	0.0539	107.1921	<.0001
metro_code_15	0	1	0.2491	0.0667	13.9294	0.0002
Sex	0	1	0.0813	0.0334	5.9171	0.0150
Q65Int1_2	0	1	-0.2267	0.0522	18.8471	<.0001
Q89bMain_1	0	1	0.0981	0.0435	5.0871	0.0241
Q89bMain_3	0	1	-0.1721	0.0861	3.9968	0.0456
Q89bMain_5	0	1	-0.2280	0.0658	11.9964	0.0005
Q814Exp_7	0	1	0.3172	0.0652	23.6391	<.0001
Q814Exp_8	0	1	0.4873	0.0574	72.1953	<.0001
Q814Exp_9	0	1	0.9624	0.0603	254.6359	<.0001
Q814Exp_10	0	1	1.8727	0.0664	795.6429	<.0001
Q814Exp_11	0	1	1.1907	0.0807	217.9000	<.0001
Q814Exp_12	0	1	1.7551	0.2128	67.9987	<.0001
Q814Exp_99	0	1	0.4791	0.2134	5.0399	0.0248
Q820Happy_1	0	1	-0.1503	0.0373	16.2292	<.0001
Q820Happy_3	0	1	0.1023	0.0389	6.9113	0.0086
Q821MicroW_2	0	1	-0.2601	0.0400	42.2747	<.0001
Q821HomeTH_1	0	1	0.0929	0.0402	5.3503	0.0207

Table A-6: Analysis of Maximum Likelihood Estimates – Testing Data

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.3	Somers' D	0.787
Percent Discordant	10.7	Gamma	0.787
Percent Tied	0.0	Tau-a	0.276
Pairs	19395794	c	0.893

Table A-7: Analysis of Maximum Likelihood Estimates – Testing Data

Partition for the Hosmer and Lemeshow Test					
Group	**Total**	**Response_Y = 0**		**Response_Y = 1**	
		Observed	**Expected**	**Observed**	**Expected**
1	1052	177	156.08	875	895.92
2	1052	392	417.92	660	634.08
3	1052	716	713.74	336	338.26
4	1052	850	853.87	202	198.13
5	1052	918	930.10	134	121.90
6	1052	980	973.03	72	78.97
7	1052	998	999.43	54	52.57
8	1052	1025	1018.04	27	33.96
9	1052	1032	1030.70	20	21.30
10	1047	1041	1036.08	6	10.92

Table A-8: Hosmer and Lemeshow Goodness-of-Fit Test – Testing Data

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
11.9327	8	0.1542

Table A-9: Classification Table – Testing Data

Classification Table									
Prob Level	**Correct**		**Incorrect**		**Percentages**				
	Event	**Non-Event**	**Event**	**Non-Event**	**Correct**	**Sensi-tivity**	**Speci-ficity**	**False POS**	**False NEG**
0.300	7896	1009	1377	233	84.7	97.1	42.3	14.8	18.8
0.400	7776	1287	1099	353	86.2	95.7	53.9	12.4	21.5
0.500	7665	1436	950	464	86.6	94.3	60.2	11.0	24.4
0.600	7490	1572	814	639	86.2	92.1	65.9	9.8	28.9
0.700	7157	1762	624	972	84.8	88.0	73.8	8.0	35.6