UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

MODEL FOR PREDICTING THE PROBABILITY OF EVENT OCCURRENCE USING LOGISTIC REGRESSION: CASE FOR A KENYAN COMMERCIAL BANK

BY

CHRISGONE OTIENO ADEDE

P58/61969/2010

SUPERVISOR

DR. ROBERT OBOKO

October, 2012

A Research project report submitted in partial fulfillment of the requirements of the Degree of Master of Science in Computer Science at the University of Nairobi.

DECLARATION

This project, as presented in this report, is my original work and has not been presented for any other award in any other University.

Name: Chrisgone Otieno Adede

Reg. No: P58/61969/2010

Signature:

Date: 24/10/2012

This project has been submitted as partial fulfillment of the requirements for the degree of Master of Science in Computer Science of the University of Nairobi with my approval as the University supervisor.

Name: Dr. Robert Oboko

Sign:

Date: 27/10/2012

ACKNOWLEDGEMENT

I express my gratitude to all those who gave me the possibility to complete this study.

I am particularly thankful to my supervisor, Dr. Robert Oboko, for his patience in leading me through the research process. I am deeply indebted to you for the help, suggestions and encouragement I received during the research and when writing this report.

To my wife, Jane Arwa, your support for this work is immeasurable and finally, we got the degree.

To my daughters, Mitchelle and Danielle, that you understood that schooling is not only for the young is appreciated. Finally, to my Parents, you laid the foundation and I am forever grateful.

ABSTRACT

Credit Scoring has been a key undertaking of lenders over the last few decades. However, the most common use of credit scoring has been as a credit application appraisal tool rather than as a credit monitoring tool for existing credit holdings. This study, in the literature, investigated the different sets of sophisticated and classical credit scoring techniques used in the areas of classification and prediction of customers defaulting on credit repayments. The main aim of this study was to build a behavioral credit scoring model from a provided dataset using Logistic Regression. The study also aimed to use data mining concepts to develop a prototype to automate the modeling process and calibrate predictions so as to formulate score cards. Our review of literature indicated the non-existence of an overall superior method for credit scoring since modeling objectives always differ and are thus suited for diverse methods. Model validation and assessments methods, variable selection and interpretation of validation results were investigated. The results of the study indicated the possibility of use of validation methods from related fields, the formulation of model deployment frameworks and formulation of a guide to trade-off between model hit-rate and false alarms. Transformation variables were also found to often offer better predictive power over raw data variables whilst the Gini-Index was shown to be a good indicator of model over-fitting. Finally, this study suggests possible future research endeavors in the effort to eliminate the requirements of statistical knowledge in the Credit Scoring process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

AI- Artificial Intelligence

ANN- Artificial Neural Network

CI- Confidence Interval

DA- Discriminant Analysis

DQC- Data Quality Configurations

GA- Genetic Algorithms

KNN- K-Nearest Neighbor

LDA- Linear Discriminant Analysis

LR- Logistic Regression

MLE- Maximum Likelihood Estimate

PD- Probability of Default

PIT- Point in Time

SEM-Standard Error of Mean

SLR- Simple Linear Regression

SVM- Support Vector Machine

TTC-Through the Cycle

CHAPTER 1: INTRODUCTION

1.1 Background

Economies and populations, over the years, have become dependent on credit with the concept of deferred payments for current consumptions gaining lots of significance. With the chance of defaulting, lenders have had to attempt to improve their odds of recovering debts.

From the 1950's, efforts to automate the process of credit risk assessment have become a core undertaking of lenders.

The developments in data warehousing and business intelligence technologies have enabled financial institutions to typically gather information on both customer characteristics and behavior in their data warehouses. In the view of Vitt et al (2002), in the current fast paced market, companies need to create competitive advantage and craft superior business strategies so as to out-smart competition. Exploitation of the power of information derivable from data should thus be a key undertaking of financial institutions that aim to be leaders in credit driven environments.

Credit scoring produces a numerical expression based on the statistical analysis of a person's credit files and aims to determine the creditworthiness of a customer and has been practically applied in making business decisions like credit limits of customers, targeted marketing campaigns, down payments, deposits and interest rates.

Recent development, in the effort to distinguish between good and bad borrowers from a pool of existing customers, has led to the importance attached to the prediction of probability of default, a tenet of credit scoring.

Credit scoring has gained prominence and ubiquity in cases which are stochastic (quantifiable in probabilities) and not in cases that have distinct certainties.

The current trends involve the automated use of computers to combine the usage of scores and strategies to make decisions thereby providing for a form of artificial intelligence (AI).

1.2 Problem Statement

The classical definition of the problem of credit scoring is that of classifying credit applicants or holders into either good or bad customer groups based on a set of attributes (X1, X2,... ,Xm). This therefore lies in the domain of classification problems and is based on the capture of the relationship between historical information and future credit-worthiness.

The Basel II accord formulated by the banking regulators of the group of ten countries (G10), in 2004 outlined the need to identify factors and create internal rating models that have the ability to differentiate risk and have predictive and discriminatory power. The full adoption of this accord by the Central Bank of Kenya will see Kenyan financial institutions move with speed to build in-house models that will then be used to rank customers.

The usage of credit scoring models in the Kenyan market has however concentrated on the appraisal of credit at the point of application for credit facilities. This is regardless of whether such applications are repeat applications from existing customers. The score cards and models used at application for credit (application score cards) are based on innate customer characteristics like age, gender, marital status, number of dependants and residence. The formulation of the application score cards equally relies on lender experience, customer repayment capacity and capital requested. This makes such score cards subjective since most of the characteristics are what customers have no control over for the purposes of access to credit.

The ranking of customers based on their behavior as depicted on how they conduct their accounts and how they have and continue to manage both previous and current credit facilities is referred to as behavioral credit scoring. Behavioral credit scoring is a growing area of undertaking by several banks, both locally and internationally. The behavioral score cards act as tools for the continuous management of the relationship between a lender and existing borrowers. This is as opposed to the current practice of dependence on ranking of credit applicants that has the limitation of not taking into account past and historical conduct of customers.

With the recent troubles of the world financial institutions as a result of customers defaulting on credit payments, behavioral scoring methods that will facilitate banks to predict the probability of an existing customer with credit facilities defaulting on credit payments are desired. Several techniques for the development of such credit scoring models exist, but with varying infrastructural, statistical, Machine Learning and computational knowledge requirements.

This study aims to search for a behavioral modeling method that gives a prediction of the probability of defaulting on credit payments but within a framework that reduces requirements of technical knowledge of predictive analytics. Such a modeling framework and method should, however, be easy to use and to interpret but with few assumptions even as it remains verifiable and robust.

1.3 Objectives

The overall reason for carrying out the research is to use data mining and statistical techniques to building a behavioral model for the prediction of probability of defaulting on credit card payments.

This involves the following specific objectives:

i. Build a behavioral model for the prediction of probability of defaulting for credit card customers which orderly ranks the customers on the basis of their future credit risk using statistical techniques.

ii. Apply data mining concepts in the modeling process for the prediction of probability of defaulting with the view to prototyping a data mining and validation framework deployable in any typical credit scoring environment.

iii. Calibration of the generated probabilities of defaulting for the formulation of score cards for the credit customer portfolio of study with the aim of grouping the customers based on the similarity or closeness of risk levels.

1.4 Research Questions

Based on the above research objectives, the research questions include:-

i. Can model building be automated and reproducibility of modeling results guaranteed in a data mining based credit scoring modeling process?

ii. What are the variables, in the provided dataset that most determine the probability of defaulting?

iii. Can the calibration of statistical models be automated to reduce human intervention?

iv. What are the benefits of automating the model building process based on the effort required for such automation?

v. Is it possible to integrate the application development, database and statistical tool environments in the modeling process?

vi. Is it feasible to automate model validation and assessment and are such model validation methods related?

1.5 Rationale for the study

The need to rank customers based on their creditworthiness has always been a key undertaking of financial institutions, not only to reduce on risk but also to identify possibilities for profit maximization.

The Basel Committee, a grouping initially formed by the banking supervisory authorities of the group of ten countries in 1975, recommends the use of the probability of defaulting to rank credit holders. However, despite the advancements in data warehousing and the vast volumes of data available for credit card customers and the Basel Committee recommendations on banking supervision (Basel II Accord), most financial institutions in Kenya still use human and expert based judgments at credit application (application models and score cards) instead of objective models based on the history of customer account and credit management (behavioral models and score cards).

In the light of reduced collateral requirements from banks, the unsecured nature of most credit card lending and the mobility of card customers coupled with the fast changing legal frameworks, accuracy in the ranking of customers using statistically and computationally verifiable methods are ever more important.

The current scenario involves the determination of how much credit a customer qualifies for based on inherent and intuitive characteristics like those of age, gender, marital status, number of dependants. Some of these characteristics are deemed discriminatory in some legal environments.

Credit scoring calls for use of methods guided by customer behavior and based on past transactional data. Such methods should be practical, with documented steps, statistically sound and computationally verifiable.

Scientific verification for correctness and reproducibility of results are mostly a problem in commonly used expert judgmental models. This calls for the development and deployment of behavioral models, in-line with the banking regulations of the Basel committee (Basel II accord). The drivers to use of behavioral models are the need for accuracy, speed, consistency and objectivity in credit decisions for existing customers while expected benefits include risk based credit pricing, decision automation and prioritization of collection efforts coupled with better loss forecasting.

1.6 Scope of the study

The scope of the project was limited to activities that lead to data preprocessing, model building, validation and assessment for the data set provided by a local Kenyan commercial bank.

The flexibility of the proposed model and framework should, however, be reproducible in similar settings with a high degree of success for any other exercise that models for the prediction of probability of occurrence using the defined methodology.

Data pre-processing activities included data extraction from source systems, data validation and other data pre-processing operations prior to this were included though they are not bound to be the basis of the study.

The prototype tool to be developed as part of the study includes a score cards generation component that therefore makes score cards generation a remit of this investigation.

The study will undertake and recommend a statistics based model deployment method for any typical productive environment.

1.7 Assumptions of the research

The assumptions in this study are as outlined here-under:-

i. The dependent variable assumption provides that the dataset to be used must have a single variable (or a transformed combination of variables) as the definition of defaulting on credit payments. It is this variable that acts as the dependent variable. The dependent and independent variables used are all assumed to be categorical in nature for the formulation of score cards.

ii. The study assumes a data mining approach on data already collected for other purposes. The proposed validation framework for the correctness of the data although proposed is noted to be just an indicator of expected trends and does not have data correction capabilities. The data is thus assumed to be a true reflection of the actual customer behaviors as they occurred.

iii. The study assumes a given set of transformations in the development of the modeling tool to be delivered. The provided list of variable transformations is by no means authoritative and complete, but sound enough to be used as the basis of doing transformations that outline customer behavior. The model deployment method proposed and the validation methods used are presented in no order of superiority.

1.8 Definition of the important terms

The following terms assume the definitions bounded in the following context.

Basel II accord: Basel II refers to the second of the Basel Accords, which are recommendations on banking laws and regulations issued by the Basel Committee on Banking Supervision

Credit worthiness: This refers to trustworthiness with money as based on a person's credit history and is a general qualification for borrowing

Logistic Regression (LR): This is a method for determining the relationship between predictor variables and a dichotomously coded dependent variable by contrasting different theoretical sets of predictor variables

R: R is a free software environment for statistical computing and graphics that compiles and runs on a wide variety of platforms (UNIX, Windows and MacOS).

Scorecard: This is a mathematical model which attempts to provide a quantitative estimate of the probability that a customer will display a defined behavior (default).

Variables: The element, feature, characteristic or factor that is liable to change is referred to as a variable and could either be dependent (default in this case) or independent (exploratory/explanatory).

Defaulting/ Default: This is the non-repayment of credit for a period of greater than 90 days from the last repayment date.

Probability of Defaulting: A measure within the [0, 1] range of the chance of a borrower failing to repay credit for a period greater than 90 days.

Ubiquity: This is the impression of being simultaneously present in multiple places.

G10 countries: A group formed by counties that accepted to be in General Arrangements to Borrow of 1962 and includes 8 IMF Members (Belgium, Canada, France, Italy, Japan, Netherlands, UK, US) and Germany and Sweden.

Obligor: An entity that owes another entity a certain debt or duty, alternatively called debtor or credit holder.

CHAPTER 2: LITERATURE REVIEW

In this section, a review of existing literature on credit scoring is done. A description is made of the data mining, machine learning and statistical techniques that are used in the prediction of the probability of defaulting. A review of the application of various techniques for credit scoring is also done.

This chapter includes a brief definition of the credit scoring problem, a review of application of data mining and machine learning in credit scoring, a description of types of scorecards, a review of both statistical and machine learning techniques in credit scoring and finally an outline of the benefits, challenges and limitations of credit scoring.

2.1 The Credit scoring problem

Credit scoring is a statistical technique that combines several financial characteristics to form a single score to represent a borrower's credit worthiness. In the view of Lahsasna et al (2010), credit scoring is defined as the capture of the relationship between historical information and future credit risk that is mathematically definable as:

$$Y_i=f(X1, X2,..., Xm)\ldots\ldots\ldots\ldots(1)$$

Where X1, X2,..., Xm are each customer's attributes, while Y_i is the denotation of the customer's resultant defaulting status (defaulted or not defaulted). The prediction of Y_i is the task of credit scoring.

Credit scoring was first proposed in the 1950's and reinforced by the Basel II accord of the Basel Committee on Banking Supervision (2004). This is based on the need to maintain regulatory capital based on financial institutions own rating of customers (internal rating models).

The target of credit scoring is to build models which rank order credit customers based on their future credit risk.

2.2 Data Mining and Machine Learning in Credit scoring

Data mining, in the view of Shmueli et al (2007), is the process of discovering hidden patterns and relationships in data in order to make better and more informed decisions. Data mining tools are thus used by businesses to make knowledge driven decisions.

The role of data and information in credit related decision making of most firms in the service sector has been noted. According to Anderson (2007), without data, modern commercial opportunities would be very limited since data and information are fundamental to the success of any business today. Data and information increasingly provide a commercial competitive edge to most institutions that formulate strategies for their exploitation.

The advancements in data mining over the years have seen them gain importance even in other areas like Machine Learning. Maclennan et al (2009) hold the view that the process of analyzing data to find hidden patterns using automatic methodologies is sometimes referred to as predictive analytics.

In the opinion of Rajaraman et al (2010), Machine Learning has use in the typical case when there is a little idea of what is being looked for from vast volumes of data. However, Machine Learning as a method is rather unsuitable for use in the solution of problems in which what is sought for is known in advance (straight goal problems). Straight goal problems are better solved using data mining techniques, especially in the presence of vast amounts of data.

It is our considered opinion that the objective in credit scoring is always know before-hand and in our opinion is more of a data mining rather than a machine learning problem.

It is the view of Shmueli et al (2007) that data mining techniques are classified as either supervised or unsupervised methods. In unsupervised methods, the feature to be predicted (target variable) does not exist before hand. Data mining algorithms are thus used in the search for patterns and structures among all the variables. Unsupervised data mining is in our opinion comparable to the popular uses of Machine Learning techniques as advanced by Rajaraman. It is, however, the case that in supervised data mining, the algorithms are provided with a vast set of examples containing the target variable from which associations between particular values of the target variable with values of the predictor (exploratory) variables are derived.

It is our assertion that Machine Learning techniques are the true data mining techniques. This is since they can be used for unsupervised as well as supervised data mining). Statistical techniques are, however, best suited for only supervised data mining techniques.

2.3 Classes of score cards

There are two alternative classifications for scorecards. The first classification is based on the duration of application of a score card (Jakimova et al (2009)). In this classification, score cards that assess ability to repay over a short time horizon (usually one year) are referred to as Point in Time (PIT) score cards. However, score cards that have a longer time horizon (usually 5 years) are referred to as Through the Cycle (TTC) score cards. The major difference between PIT and TTC score cards is that PIT score cards are based on immediate circumstances and are mostly quantitative whilst TTC score cards take adverse and unforeseen economic events into account and at time use past data into account.

The alternative classification of score cards puts score cards basically in to two types of application and behavioral score cards. Past practical studies and application of credit scoring have mostly concentrated on the development of application score cards. Application score cards have, as their basis, the characterization of new (entrant) customers into the basic good or bad classes mainly for credit appraisal. These have not incorporated the Basel II accord that recommends as an output, a prediction of probability of defaulting. Behavioral score cards are those based on customer behavior and are thus a true reflection of the customers interaction with the lender. Behavioral models are based on transactional information as they happened and are derived from an organizations data warehouse or transactional systems.

The information used for application scoring, as outlined by Anderson (2007), is reliant on the usage of information on a customer's credit application. Such data include: socio-demographic information (age, gender, marital status, number of dependants etc), financial data (current account balance, retuned cheques etc), work status (contract, permanently employed, non-employed etc) and data sourced externally from institutions that hold industry credit defaulting information (credit reference bureaus) data. Such data, with the exception of bureau data, are thus hard to validate. This is as opposed to behavioral scores that are based on customer initiated account behavior, adverse indicators including returned cheques and past delinquencies, the information of which is obtainable from the organizations data warehouse and transactional systems, extracted via data mining techniques.

2.4 Review of credit scoring techniques

Credit scoring, as a process, uses the concept of prediction of probability of occurrence of an event and could be achieved by a variety of methods, ranging from machine learning approaches to statistical and data mining approaches, with the most common implementations built upon statistical approaches (predictive analytics).

The most common Machine Learning techniques in credit scoring include Support Vector Machines (SVM), Artificial Neural Networks (ANN), Genetic Algorithms (GA) and K-Nearest Neighbor (KNN) while statistical methods include Discriminant Analysis (DA), Probit models, Linear Regression and Logistic Regression (LR).

2.4.1 Statistical techniques in credit scoring

Gosling (1995) and Anderson (2007) assert that predictive statistics is divided into either parametric or non-parametric modeling procedures. Parametric models assume a precise form of the probability of distribution and include methods such as Linear Regression, Discriminant Analysis and Logistic regression. Non parametric models do not assume a type of distribution for variables and are thus termed as being distribution free and are similar to the Machine Learning techniques such as Artificial Neural Networks.

According to Anderson (2007), the tools for credit scoring are mostly from data mining, statistical and operational research. Statistical techniques and are some of the most successful and profitable applications of statistical theory in the last two decades.

In the opinion of Joao (2008), Linear Discriminant Analysis (DA) was the first parametric technique suggested for credit scoring, having been developed by R.A. Fisher in 1936. DA is best applied in classification problems where the variable to be predicted is categorical in nature. This would be limited in application in instances where raking of customers is the aim. DA is not based on the search for the relationship between variables but on classification. The best application of DA in our opinion would be in linear classification and reduction of dimensionality in the modeling process.

Regression analysis, in the assertion of Montgomery et al. (2003), is a collection of techniques that focus on the relationship between dependent and independent variables and are mostly used in the fields of prediction and forecasting. The variants of Regression that have been used in prediction and forecasting include Linear Regression, Probit models and Logistic Regression.

Linear Regression models involve the prediction of a scalar variable from one or multiple predictor variables. The case of one predictor variable is called Simple Linear Regression (SLR) while that of multiple predictor variables is called Multiple Regression. In credit scoring, Linear Regression lies on the assumption that the dependent variable and predictor variables are linearly related and their extrapolations well beyond observed variable combinations to correct for differences between learning, test and performance data sets (Domingue et al. (2009)). Yang (2005) asserts that Linear

Regression has the limitation of an output of a probability that is possibly out of the [0, 1] interval thereby making them hard to interpret.

Probit Analysis (PA) models are regression models where the dependent variable assumes only two values (binary dependent variable). Abdou et al. (2007) asserts that in Probit models, a linear combination of the independent variables is transformed into its cumulative probability value from a normal distribution. This twin reliance of PA models on normal distribution of independent variables and the binary nature of the target variables serves as their limitation in use in credit scoring scenarios that desire real valued output or in cases where cumulative probability values do not have a normal distribution. It is the view of Anderson (2007) that Probit models are equally not easy in implementation due to their requirement for numerical calculation of integrals that makes for high computational complexities.

Logistic Regression (LR) is a variant of Linear Regression and is considered to be a special class of Linear Regression that is currently one of the most popular methods in building credit scoring models. Yang (2005) outlines the LR model as using the following decision model:

$$\ln(p / 1 - p) = W_0 + W_1 X_1 + \ldots + W_k X_k \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where p is the probability that the outcome *y* will occur and $p /(1 - p)$ is the odds that *y* will occur. The probability of occurrence p of *y* can be computed using:-

$$p = 1 /(1 + \exp(- W_0 - W_1 X_1 - \ldots - W_k X_k) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Where W_0 is the intercept while W_1. W_2_ W_k are the regression coefficients of the independent variables X_1. X_2_ X_k respectively. It is the observation of Yang that the Logistic Regression model is similar to the Single Layer Perceptron (SLP) Artificial Neural Network, the difference being on being that LR uses MLE for the estimation of weights while SLP uses the Perceptron Learning Rule outlined in Mitchell (1997).

The popularity of Logistic Regression is due to the fact that the output predictions of the models are within the [0, 1] interval. This is guaranteed by the assumption that the input variables times their weights are linearly correlated to the natural log of the odds that the outcome event will occur (Yang (2005)).

Logistic Regression does not assume the distribution of predictor variables and handles, with a great measure of success, continuous and categorical variables. This is as opposed to the opinion in Mugambi (2011) that advances the limitation of LR in handling categorical variables when in fact the complexity of the method is in conversion from continuous to categorical variables.

The output of a Logistic Regression model is a Maximum Likelihood Estimate (MLE). MLE in statistics is used in determining the parameters of a model that maximizes the probability of the sample data. The earlier limitations of computational complexity in the use of MLE methods for prediction, in the opinion of Anderson (2007), have been overcome as time progressed and computing power increased. MLE have therefore become more feasible and practical in credit scoring.

2.4.2 Machine Learning techniques in credit scoring

Support Vector Machines are non-probabilistic binary linear classifiers that have been used and shown to produce results that are comparable to the statistical methods and with better performance on out of sample test sets (Gestel et al. (2003)). Bellotti et al. (2009) documented the improvement in performance of SVM models over the traditional techniques of Logistic Regression and Discriminant Analysis but equally notes the limitation in the requirements of a large number of learning tasks (lots of data requirements) for the marginal gains in performance.

Artificial Neural Networks have the ability to learn complex non-linear structures from data sets. Lahsasna et al. (2010) asserts that of the non-traditional methods, ANN's are a serious alternative to the traditional methods of credit scoring, a fact vindicated by other past studies. Desai et al. (1996) and West (2000) have surveyed the use of and efficiency of ANNs in credit scoring. Desai and West note that the superiority of ANN's to other methods is based on their nonlinear nature and elimination of priory assumptions on the distribution of the data. The biggest limitations of ANNs, as explained by Mitchell (1997) and Lahsasna et. al. (2010), include the absence of explanatory capabilities and the non-existence of a structured way of selection of parameters.

Research in use of Genetic Algorithms (GA) in credit scoring has been given prominence following the success of ANNs. Desai et al. (1997) found that Genetic Algorithms are superior to Discriminant Analysis, Logistic Regression and some variants of Artificial Neural Networks but are constrained by their computation costs and lack of comprehensibility.

K-Nearest Neighbor Algorithms base their classification on closeness to provided learning examples. Though not widely documented in credit scoring, this technique has been extensively used in hybrid environments of credit scoring for variable ranking for prediction of churn rates (Desai et al. (1997)). This is, in our opinion, a deployment of KNN in an area which rests in the domain of prediction of probability of event occurrence.

2.4.3 Selection of credit scoring technique

Anderson (2007) gives the criteria for the evaluation and choice of the specific techniques to be used. The criteria are based on the following modeling considerations, amongst others:-

- Suitability

  This takes into consideration the appropriateness for the data to be modeled.

- Development speed

  This is based on how fast the model can be developed and the time lines given to the modeler. It is widely observed that complexities of modeling techniques and tools deployed in credit do not directly translate into superior and more predictive models. Anderson (2007), in reference to studies in economics, advocates for the use of simple and times naive techniques that have been proven to consistently outperform more sophisticated alternatives. Naivety in our opinion does not, however, equate uninformed or arbitrary techniques but those that are informed by theory.

- Adaptability

  The flexibility of the modeling technique as compared to the changes in modeling environment will dictate choice of techniques.

- Output transparency

  The desired simplicity in interpretation of the outputs of a modeling process dictates the choice of techniques to be used. This will see a method that gives superior results discarded for those that give less accurate results but that are easier to interpret.

- Data considerations

  Data considerations such as interactions (relationship between variables), type of variable to be predicted (continuous with un-bound limits or discrete with countable unique values) and frequency of events (how many times the event occurs in the data) dictate the choice of modeling technique to be deployed in any environment

It is our considered opinion that the choice of what method to use for credit scoring for financial institutions formerly depended on the ease of calculation, transparency and suitability to the problem. There, however, has been a shift based on the Basel II accord for risk compliance that encourages the use of modeling techniques that give an estimated probability of defaulting on credit payments.

The limitation of methods that do not compute the probability of defaulting in credit scoring, in the assessment of Thomas et al. (2002), is that of the mistaken assumption of creditworthiness as a personal attribute (height, weight, or eye color) or as a directly measurable quantity. This is without regard to the fact that creditworthiness should be considered in the context of other factors (like amounts borrowed and credit limits) and that these vary across lenders and sectors.

In a comparative study of Artificial Neural Network methods versus the traditional statistical methods, West (2000) investigated the accuracy of ANN based models for credit scoring (Multilayer Perceptron, mixture-of-experts, Radial Basis Function, learning vector quantization, and fuzzy adaptive resonance) against traditional methods of Discriminant Analysis and Logistic Regression. The conclusion of the study was that back propagation Neural Networks are the most accurate non-traditional models whilst Logistic Regression is the most accurate of the traditional model techniques.

2.5 Requirements, drivers and expected benefits of credit scoring

As the amount of data continues to grow within financial organizations, there arises a greater need to convert these massive datasets into useful information. The reasons for this include competitive advantage in terms of faster decision making, process control and information management.

According to Caire et al. (2006), the benefits of credit scoring models should be viewed in the long term as opposed to the short term. In costing, and only considering use of internal expertise in the context of micro-finance institutions, Caire estimates that it requires at least the intermittent time of a senior manager and an in-house information technology (IT) specialist for anywhere from 6 to 24 months to develop a score card. Outside expertise, however, could charge from $10,000 to $65,000 to develop a scorecard. Specialized software to develop and deploy a scorecard could also cost tens of thousands of dollars. These estimates are by all means a modest illustration of why there is need for easy to develop, inexpensive and easy to deploy credit scoring models.

Quittner (2003) holds the view that current development of the sub-prime lending industry for customers who have poorly documented credit histories, and thus fall short of traditional credit acceptance and risk levels, is attributable to the benefits of credit scoring.

The sub-prime lending market has not only been supported by automated underwriting but also by the usage of credit scoring technologies that dictated the initial success of specialized financial institutions in this market. Quittner predicts a growth of the lending market as technology in credit scoring advances, a trend that is currently being witnessed in the almost universal use of score cards to manage credit applications. The recent credit scoring advances could be used to grow the loan portfolios and profitability of banks and other financial institutions.

The implicit benefits and thus drivers for credit scoring include but are not limited to accuracy (through the reduction of biased selection), speed (through the automation of credit processes), consistency (from the resultant standardization of procedures across all branches and loan officers), and objectivity (implying defendable decisions). The use of such scores to produce inputs for analysis for more decision making processes like risks and portfolio management is an explicit benefit.

2.6 Challenges and Limitations of building credit scoring models

2.6.1 Challenges of Credit scoring

The challenges of credit scoring include over-fitting of models, data limitations, the erosions of model predictive power, variable selection and model validation.

According to Mitchell (1997), the greatest limitations to use of Machine Learning techniques are the risks of over-fitting and sample size limitations. These problems are noted to be similar to those in data mining and even to those in statistical methods.

The problem of over-fitting results in a model that performs well in the learning data-set but not in the application data set. This is due to over generalization at the model building stage and calls for automated detection of over-fitting during model building. This remains a challenge even though methods of detection are widely published.

With time and changes in economic factors, the predictive powers and performances of scorecards are bound to deteriorate. Scorecards, therefore, have to be constantly redeveloped - around every 18 months to 2 years in order to overcome this drift in the population as articulated by Thomas (2000, p.164).

Variable selection and model validations is in most modeling environments a manual process. Automation of this based on rankings would make it easy on the decision of what variables to include in the models.

Statistical credit scoring, despite the challenges of model validation and continued assessment, are exhaustive in quality assurance and objectivity hence superior to judgmental models. Thomas (2000, p.167) asserts that once an organization takes up statistical and Operational Research based credit scoring, it hardly ever returns to judgmental based models.

2.6.2 Limitations of Credit scoring

The fundamental question in the usage of credit scoring is the extent of reliance on them that is considered reasonable. The use of score cards is to be done with caution as advocated for by Anderson (2007). He asters that over-reliance on score cards can becomes their limitation. There still is need for end users to still maintain and exercise their knowledge in special cases, in the form of overrides, since transactional systems do not hold all information on a customer.

The sampling process, as advanced by Hand et al. (2001), may lead to the building of biased models with the rejection of the data of potential customers. Biased sample result when the chosen population does not conform to the general population. This is however solved through the use of sampling methods that minimize bias.

One of the most salient assumptions of credit scoring is the linkage between the past and the future and is based on the view that the past will always mirror the future (Berry et al., 2000). A shift in data patterns over time if rapid enough will invalidate models. This is solved by continued refresh of models and continued assessment of their predictive power.

The limitations of credit scoring should not be a reason for their non-deployment since they guarantee a competitive advantage in their usage.

CHAPTER 3: METHODOLOGY

3.1 Introduction

In this chapter, the research methodology is outlined. An indication of model type, techniques and tools, data collection and collation techniques is also done. Variable transformation, model validation and assessment, score cards generation and model deployment methods are also discussed.

The study, based on the classification of research methods in computer science as advanced by Dodig-Crnkovic (2002), falls under modeling and involves the simplification of a phenomenon (defaulting on credit payments) so as to study it. Abstraction is thus a large part of the methodology of this study.

The Sample, Explore, Modify, Model, Assess (SEMMA) methodology is used in the modeling process but with extensions for it to support iterations at each of the stages. The extensions have led to the proposed Explore, Transform, Sample, Re-categorize, Model and Assess (ETSRMA) methodology that provides an iterative modeling methodology but with finer implementation steps between two main steps.

3.2 Model Type Developed

This study is based on the development of behavioral models. Behavioral models are based on customer initiated account behavior, adverse indicators including returned cheques and past delinquencies, the information of which is obtainable from the organizations data warehouse and transactional systems, extracted via data mining techniques. This is as opposed to application models that are comparable to customized rule-based models build out of a banks experience of what is considered to be the best descriptors of credit risk and are thus judgmental.

The Behavioral models developed are Point in Time (PIT) models since the basis of the models is the prediction of probability of defaulting on credit payments with a time horizon of one year in the future.

3.3 Modeling Techniques, Tools and Modeling Infrastructure Used

The chosen statistical modeling technique is Logistic Regression (LR). The choice is based on the need for the local Kenya banks to be compliant to the Basel II regulatory framework that stipulates the development of internal rating models that deliver prediction of the probability of default as a real valued output for every customer. The output, Maximum Likelihood Estimation (MLE), is the recommendation of the regulatory framework.

The concept behind the use of MLE based methods is the fact that the generated probabilities are usable in multiple application domains and is the basis for development of product offering for existing good customers with low possibilities of defaulting on credit payments.

The statistical tool used in the manipulation of the data is the open- source statistical software-R that is downloadable from the World Wide Web and used under the GNU license. The presentation of results and the development of the prototype are done using C# and DevExpress while the data is stored in an SQL Server 2008 database.

The prototype infrastructure is as illustrated In Figure 3.1

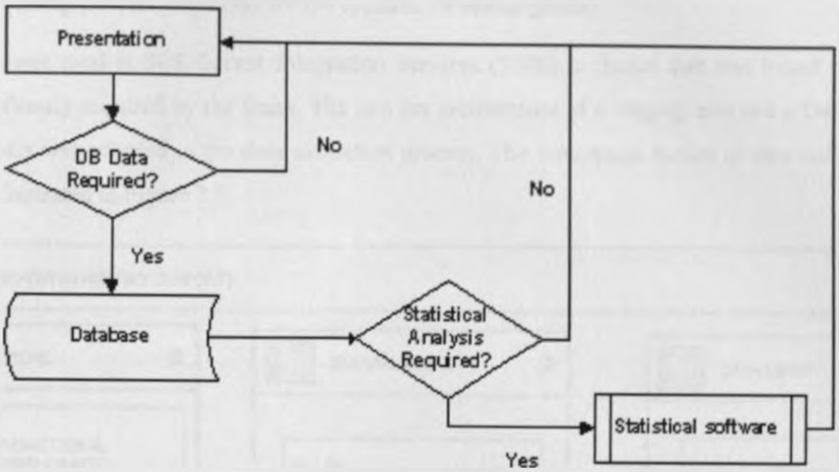


Figure 3.1 The integrated prototype modeling environment

3.4 Data Collection and Collation

For credit scoring using Logistic Regression, at least 18 months of data is required. The data was divided in to two subsets of observation (learning) period and performance (validation) period. The observation period is 6 months of behavioral information whilst the remaining period of 12 months is the performance period. This is as illustrated in Figure 3.2.

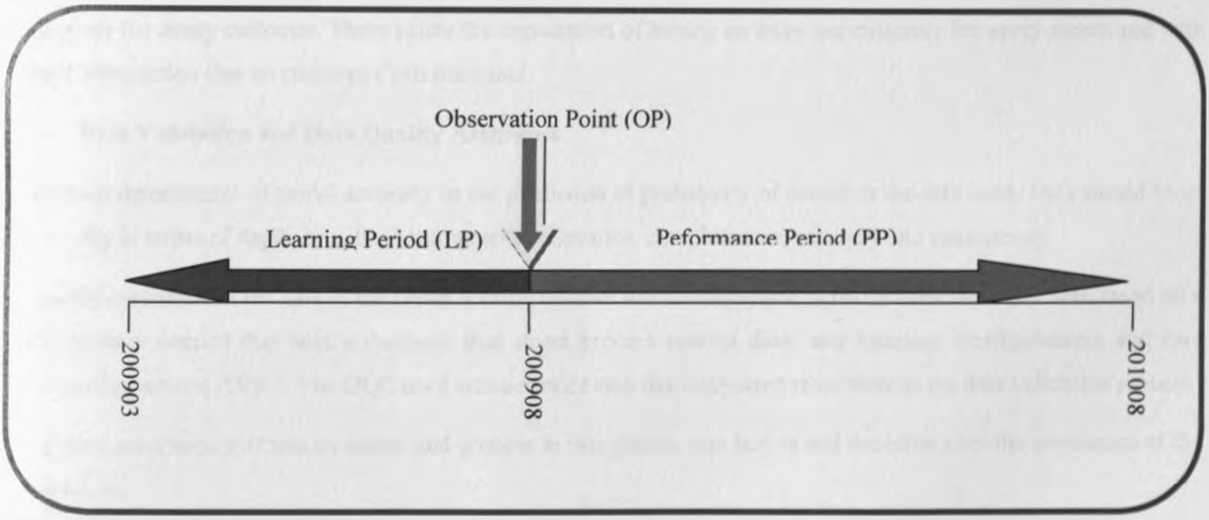


Figure 3.2 Illustration of Observation versus Performance Period

ETL based data extraction techniques were then used to extract the data from the Bank's credit card transactional systems.

The back dating of the data extraction process was done to make the model building exercise an immediate undertaking rather than that of waiting for data collection for the required 18 month period.

The ETL tool that was used is SQL Server Integration Services (SSIS), a choice that was based on the existence of licenses that were already acquired by the Bank. The two tier architecture of a Staging area and a Data Mart area for the credit scoring process was adopted in the data extraction process. The conceptual model of data collection process that was adopted is as illustrated in Figure 3.3.

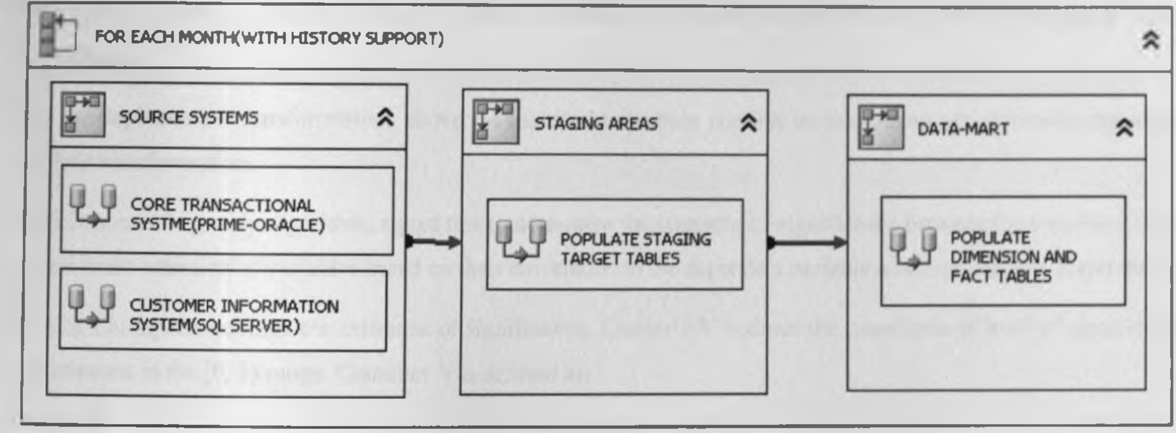

FOR EACH MONTH(WITH HISTORY SUPPORT)
SOURCE SYSTEMS
CORE TRANSACTIONAL SYSTME(PRIME-ORACLE)
CUSTOMER INFORMATION SYSTEM(SQL SERVER)
STAGING AREAS
POPULATE STAGING TARGET TABLES
DATA-MART
POPULATE DIMENSION AND FACT TABLES

Figure 3.3 Conceptual model of data collection process

The Panel data format was adopted. This involves the observation of multiple behavioral characteristics over multiple time periods for every customer. There exists the expectation of having an entry per customer for every month and with the basic assumption that no customers exit the panel.

3.5 Data Validation and Data Quality Assurance

The greatest determinant of model accuracy in the prediction of probability of default is the data used. Data should be of good quality in terms of depth, breadth, homogeneity, relevance, completeness, accuracy and consistency.

The quality assurance of the data in the credit scoring process was implemented in two phases that were both based on a generic process control that held a database that stores process control data, user interface configurations, and data quality configurations (DQC). The DQC used was a service task that supported reusability in the data validation process.

Data quality assurance will was an automated process in two phases, one before and the other after the population of the staging tables.

The first phase was for validation of field level data and the calculation of basic statistics on the data like Minimum values, Maximum values, Average, Sum and Standard deviation of the monthly data to ensure reasonable distribution before population of the data mart. A set of user set up rules on trends was used in this phase based on expected standard deviations.

The second phase. however, was a precursor to the model building phase and involved inter-period trend analysis using SQL queries on the data and excel pivot tables on the data to identify any trend violation and spikes on the data.

3.6 Variable Handling: Transformation and Selection of Variables

The variables from the panel data are not necessarily of the most statistical relevance in the indication and measure of behavior. Transformations are done to realize new variables that could be of increased predictive power. Transformations implemented in this study are either static or overtime and includes aggregations (SUM. MIN. MAX, and AVG), moving averages, averages without extremes and ratios that are bound to be of increased statistical significance.

All the implemented transformations above are assumed to be only possible on continuous variables with the exception of AsIs transformations.

The Crammer V greedy algorithm, a post test to determine the strengths of significance between the variables, was used to guide the selection of variables based on their correlation to the dependent variable and inter-variable correlations.

Whilst Chi-square outlines the existence of significance, Cramer's V outlines the magnitude of level of significance for correlations in the [0, 1) range. Crammer V is defined as:

$$Cramer\ 's V = \sqrt{\chi 2 /(N (k - 1))} \text{ ....} (4)$$

Where: χ^2 is the Chi-Square value, N the total number of observations and k the number of rows or columns, whichever is less.

Other statistical indicators that are used in variable selection include:-

- P-value and level of significance

  In this study, the Null hypothesis (H_o) was that a variables coefficient is not significant (does not add to the predictive power of the model). This null hypothesis is then rejected if the P-Value is less than the preset level of significance that was assumed to be at 0.05 (95% Confidence Interval).

- Standard Error of the Mean(SEM)

  The coefficients of the LR model are mean values. SEM is the ratio of sample standard deviation and the square root of the sample size.

$$SEM = s / \sqrt{n} \text{ ....} (5)$$

- Odds Ratio

The Odds Ratio is used to assess the strength of association of probability of event occurrence between any two groups. It is an indication of the number of times an event is likely to occur in one group compared to another group. For the LR models developed, the odds ratio is exponential of the coefficients.

The ranges of the Odds Ration are given by Formula (6), assuming a 95% confidence interval:-

$$OddsRatioCI = Exp(Coeff \pm (1.96 * SEM)) \quad \text{.................................} \ (6)$$

The chosen modeling method, Logistic Regression (LR), for the ease of management of the resultant scorecards, is based on categorical variables.

To ensure all variables are categorical, necessary conversions are supported in this modeling process. These conversions (Re-categorizations) are done on both categorical and continuous variables so as to generate new categories for the variables.

Re-Categorization is thus straight forward for categorical variables since it involves the combination of one or more variable categories. However, the algorithm outlined below was used for the conversion of continuous variables to categorical variables:-

i. Select the variable to convert from continuous to categorical and determine the number of desired groups to generate (G).

ii. Get the list of all records that are marked to be in sample and obtain their count (Cnt). Add a ranking (R1) to the records based on the variable. The Rank is allowed to repeat for the same values and are non-consecutive if values repeat. For Example, if two similar values are both ranked x, then next available rank is x+2). Leave out exception cases from the rankings (special cases include negative values for a variable expected to have positive values and vice versa)

iii. Generate a row number (R2) on the above set of records.

iv. Obtain the group by getting the product of G and the average of R2 over each Rank (R1). Divide the resulting value by Cnt+1 and add 1 to the above value (to ensure group zero is left for exceptional cases) and get the floor.

Category=FLOOR ((Avg (R2)*G/Cnt+1) +1)

3.7 Sampling and Sample Definition

The prevalent data volumes was noted to be relatively low and thus non accommodative on learning from a sample of the data. The entire population of the data collected was thus used in the modeling process.

Sampling was however applied at the model building stage and to reduce any biases in the sampling process, random sampling was used.

The sampling methodology thus involved the generation of a random number against each individual participant (to ensure an equal chance for every individual to be selected) and the assignment of a cut off to separate those to be in-sample from those to be out of sample.

3.8 Model building, Assessment and Validation

3.8.1 Model Building

The model building and assessment supported in this study is iterative in the search for the perfect model. The model building and assessment process is thus a state space search for the perfect model.

The population for which a model is built is further divided into two sets: training (in-sample) and validation (out-of-sample). The ratio of in-sample to out of sample population is set at approximately 70%: 30% with the motivation to avoid occurrence of over-fitting and thus increase model accuracy and applicability in the performance dataset.

For model variable selection at model building, the Stepwise Selection approach is implemented. In this approach, the model development begins with no variable in the model, variables are added one (single or batch) at a time. After addition of a set of new variables and the creation of the model, all variables become candidates for removal, based on appropriate statistical measures.

3.8.2 Model Assessment and Validation

Model validation, in most cases relies on stakeholder and data based techniques. In this study, we investigate the usage and automation of the model validation and assessment process.

The following techniques are implemented and used in model validation in this study:-

i. **Compliance with business logic**

   The choice of variables and the resultant score cards are validated against business expectations to help incorporate expert advice in the modeling process. Some variables could be made part of variables even when their statistical contributions are found not to be of any significance.

ii. **Capture**

   This is used to measure the performance of the model as measured by the ratio of capture of the default cases and will serve as a measure of the sensitivity of the model.

   The capture level to be adopted follows the 20:80 rule. This has the implication that 80% of the defaults are captured at the top 20% of an ordered listing of defaults.

iii. **Capture-Response, Cumulative accuracy profile (CAP) and Gini-index (Accuracy ratio)**

The automation and use of cumulative accuracy profile (CAP) and the Gini-index (Accuracy Ratio) for model validation was assessed in this study. The rule of thumb, of accepting models with an accuracy ratio greater than 0.7, is followed in the search for the best model.

iv. **Lift charts**

The possibility of automation of the use of lift charts is investigated and implemented in the methodology of this study.

Lift is a measure of the effectiveness of a predictive model. Lift is calculated as the ratio between the results obtained with and without the predictive model.

It is a plot of how much better it is to use a model than it is to randomly classify the occurrence of an event.

v. **Receiver Operator Characteristic curve**

With its origin in Engineering (signal detection theory), the ROC curve is a graphical plot of the sensitivity (true positive rate) versus 1-specificity (false positive rate). ROC curve is derivable from a Confusion Matrix.

The quadrants of a Confusion Matrix, as used for categorical classification, are as illustrated in Table 3.1.

Table 3.1 Binary Classification Confusion Matrix

Predicted Class	**Actual Class**	
	1 (Defaulted)	0 (Not Defaulted)
1 (Defaulted)	**True Positive (TP)**	**False Positive (FP)**
0 (Not Defaulted)	**False Negative (FN)**	**True Negative (TN)**

The ROC curve is implemented as a graduated threshold selection process and is plot of TP versus FP.

This model validation and assessment measure can be used to set an appropriate threshold for the conversion of the real valued model output to a Boolean classifier. Swets et al (2002) documents the use of ROC curves to depict trade-off between hit rates (sensitivity) and false alarms (false positives) of classifiers.

3.9 The General Process of data mining

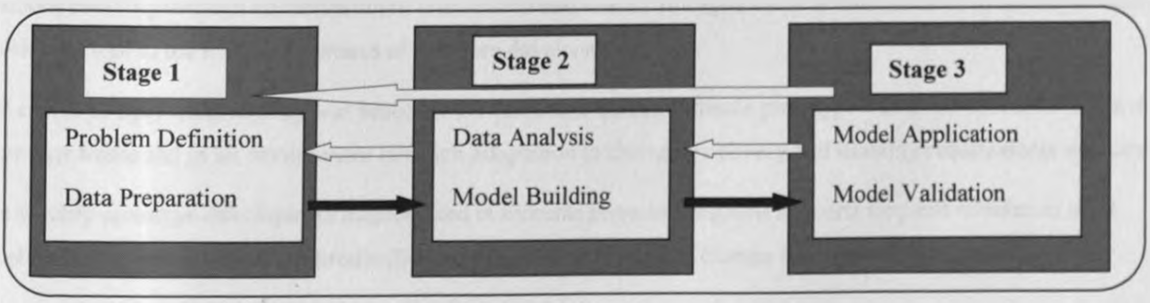


Figure 3.4 High Level data mining framework

A high level general view of the data mining process for credit scoring is as shown in figure 3.4 above. This study uses this framework of the data mining process with modifications to make for more sub processes that have a shorter interactive cycle.

Stage 1 for this study is clear cut since the problem of building a credit scoring model to rank customers is clear and data preparation is assumed to have been done with reasonable correctness.

Stage 2 was implemented as a modeler managed process while Stage 3 is equated to a state space search for the most predictive model within the set of all possible models based model validation parameters that are deployed.

It is worth noting that, even though the general modeling process does not provide for model deployment, this study proposed a model deployment framework.

3.10 Model Deployment

Most studies do not provide a framework for model deployment, a process that is always left at the discretion of the modelers and users of the generated score cards.

A statistically sound procedure for model deployment was investigated and is documented as a three step process involving:-

i. Score card generation
ii. Creation of model risk buckets
iii. Actual deployment of the model for credit scoring of a given set of observations.

The modeling prototype developed has the capability to support the visualization of the credit scoring process and aid the creation of risk buckets that is mostly assumed to be intuitive.

The process of score card generation and that of actual deployment are both scientific and could be automated whilst the creation of model buckets is modeler based and is thus guided by the tool developed.

3.11 Prototype Development Methodology

The study involved the delivery of a software based prototype. Agile software methodology was the methodology of choice since it supports an interactive and incremental approach to software development with requirements and solutions evolving through collaboration of cross functional teams. The agile cycle is thus made up of iterations each of which is related to the traditional phases of software development.

The choice of agile methodology was based on the fact that a unique software prototype was to be delivered within a short time frame and in an environment in which adaptation to changing delivery and usability requirements was desired.

The specific prototype development method used is Extreme programming that supports frequent releases in short development cycles, improves on productivity and is accommodative of change in customer requirements.

Further details of prototype development are presented in Chapter 4.

CHAPTER 4: SYSTEM DESIGN AND IMPLEMEMNTATION

4.1 Introduction

This chapter outlines the design process and implementation of the software prototype that was built for the purpose of experimentation in this study.

The implementation of the system in terms of the data set used, the programming strategies selected and the testing process is outlines.

4.2 Description of the Basic Dataset

The Basic Dataset in this study refers to the raw dataset that was sourced from the Banks source systems for this study. It therefore excludes any such variables as are product of any transformations.

Based on the study methodology that settled on panel data as the best data set presentation strategy, the provided data set had a participant identifier (CustID) and an observation time identifier (PeriodDate). The interval of observation adopted was one month.

The result was that each participant had an entry per month in the dataset. There equally was the assumption that no participant exited the panel one enlisted.

The variables within the dataset represent the monthly position of the measured or indicated behavior. The variables thus have varying aggregation operations ranging from final position (for the status indicators), counts and summations for the numbers and values of transactional features.

The advantages of using Panel data format includes:-

- It handles unobserved or omitted variables that remain fixed over time thus eliminating the omitted variable bias that is common in statistical learning.
- Reduces the problem of co-linearity among variable and increases efficiency gains from the use of more data
- They can be used to measure dynamics over time

The variables (features) that were provided in the basic dataset were as listed in the Table 4.1

Table 4.1 variables in the basic dataset

Variable Name	Variable Description
CustID	The participant identifier
PeriodDate	The month of observation of the participant features
BrandName	The historical best performing brand name of the participant credit card
BillingCycle	The participants billing cycle (month-month cycle). In categories of 1 (5th)...
GeneralStatus	This the indication of the activity status of the card.
FinancialStatus	This represents the mostly temporal financial status of a card.
AuthorizationStatus	The authorization status indicates the last authorization status on a card for the month of observation
NumberOfCards	The number of credit cards held by the participant.
CreditLimit	This is the maximum amount that the participant is allowed to transact within an observation period.
OverdrawnAmount	The amount owed by a participant over and above their credit limit
NumberOfTransactions	The count of transactions by a participant within the observation cycle
ValueOfTransactions	The total amount transacted by a participant within the observation cycle
NumberOfCashWithdrawals	The count of cash withdrawal at ATMs by a participant within the observation cycle
ValueOfCashWithdrawals	The total amount of cash withdrawn at ATMs by a participant within the observation cycle
AmountPaid	The total amount paid by a participant within the observation cycle
AmountPastDue	Total amount owed past stipulated payment period-Overdue amount
DaysPastDue	The number of days after the last payment was to be made-Overdue days
OutstandingBalance	The total amounts owed by a participant at the observation period
OverdueCycles	The number of months a participant has been overdue
Tenure	The number of months a participant has had a credit card relationship with the Bank.
InternalEmployee	In indication whether the customer is an internal employee or not

4.3 System design

4.3.1 The system design process

The prototype development methodology used is Extreme Programming (XP), a variant of Agile software development methodology.

The design process involved an application of the XP values of Simplicity and Feedback.

The process was iterative but at each stage commencing with the simplest possible solutions that were then subjected to refactoring for better and more complex solutions.

Feedback was achieved through unit tests that communicated the need for recoding broken and non-functional code components to the developer. The principle of rapid feedback was thus observed in the prototype development process as is a principle in XP.

The design process is as illustrated in Figure 4.1.

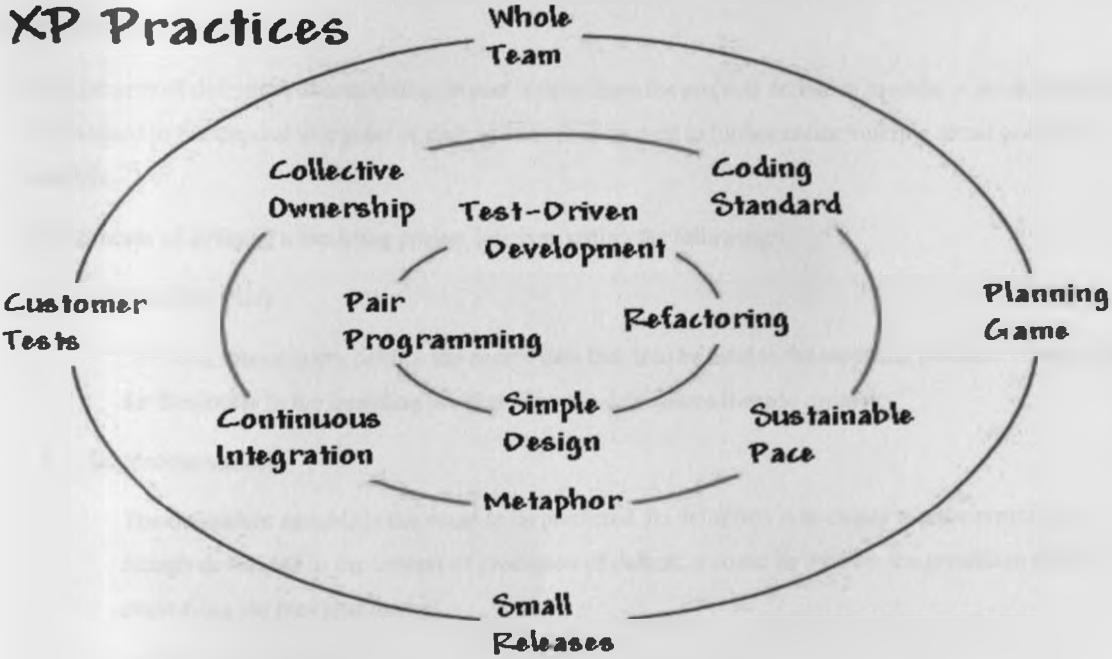


Figure 4.1 Illustration of Extreme Programming Process (XPgroup, 1999)

4.3.2 The system modules (Modeling process implementation)

The modeling methodology used in this study was the earlier proposed Explore, Transform, Sample, Re-categorize, Model and Assess (ETSRMA) methodology.

This methodology guided the design process by dictating the module and module names that were adopted in the resultant prototype. The resultant prototype modules are described here-under in section 4.3.3.

4.3.3 Description of the Models of the Prototype

The modules of the system are as described below and double up as an outline of the modeling process.

a) **Authentication to Project**

The prototype assumes that authentication to use the system includes access to one of the several listed projects. A project is thus the workspace for every successful authentication.

Authentication in this aspect does not include access administration as this is considered as out of the scope of the study.

An authenticated user can thus only work on the project thy have logged into.

b) **Systems access administration**

This module is used to configure the server and associated login details as well as to manage the system access profiles.

c) **Project Definition Module**

The process of definition of a modeling project is done from the projects definition module. A modeling project is assumed to correspond to a panel of data and can thus be used to further create multiple actual prediction models.

The process of defining a modeling project involves setting the following:-

- Data source query

  The data source query defines the source data that is to be used in the modeling process. This provides for flexibility in the modeling process since the data source is made dynamic.

- Dependent variable

  The dependent variable is the event to be predicted. Its definition is to ensure that the prototype, though developed in the context of prediction of default, it could be used for the prediction of any event from the provided dataset.

- Observation point

  This outlines the end of the learning period. Before the observation point (and including it) there must exist 6 months of data and a further 12 months of data after the observation point.

d) **Explore**

Explore module corresponds to the first activity in the model building process. This is the data quality verification module. It gives a structure for the determination of what type of analysis to run on the data (Categorical analysis or Continuous data analysis).

e) **Transformation**

The transformation module is used in the generation of transformed variables. The transformations result in the generation of new variables from the dataset and are of three types:-

I. **Automatic Transformations**

These are inbuilt in to the modeling process and are internal management variables. The automated transformation variables are thus mostly indicators on the data and are as listed in Table 4.2.

Table 4.2 Automatic transformation on the Basic Dataset

Variable	**Description**
CountMMback	Count of number of months of observations for each participant before and up to the Observation point.
CountMMFor	Count of number of months of observation for each participant after the Observation point.
WithHistory	An indicator of whether a participant has 6 continuous months of observations before and up to Observation point(1 If CountMMback=6,0 Otherwise)
WithHorizon	An indicator of whether a participant has 12 continuous months of observations after the Observation point or has an occurrence of the event in the future. (1 If CountMMFor =12 (or DependentVariable=1), 0 Otherwise)
PastDueAtObs	This is an indicator of whether the participant has occurrence of event at the observation point. (1 If has event at observation point, 0 otherwise).
ForModelInd	An indicator whether a participant qualifies to be learnt from. This is defined as a participant who has history-WithHistory=1(CountMMback=6), has Horizon (CountMMFor=12 or DependentVariable=1) and does not have occurrence of event at the Observation point (PastDueAtObs=0)

II. Script Transformations

Script transformations are an advance set of transformations that require the knowledge of Structured Query Language (SQL). The equally lead t the definition of new variables on the initial data set. They are optional in the process but could be used to defined better measure of behavior that those in the Basic dataset.

An example of a script transformation is as illustrated in Table 4.3.

Table 4.3 Example of Script Transformation

Variable	Script	Description

UtilizationVoT	CASE WHEN CreditLimit<>0
	THEN
	1.00*ValueOfTransactions/CreditLimit
	ELSE 0 END
		Ratio of the value of transactions within an observation period to the credit limit of a participant. This gives a measure of credit limit usage behavior.

Script variables, with creativity in their definition, could turn out to be a better measure of behavior over time as compared to the individual fields.

III. Extended Transformation

Extended transformations can be done on both Initial dataset variables and those from script transformations.

For this study, the following extended transformations are implemented in the prototype:-

- Entry at Observation Point (AsIs)
- Average for Last Six Months (Avg6) and Average for Last Three Months (Avg3)
- Average without Extremes (WoExt)
- Ratio of the first month to that of the sixth month (Ratio1_6), Ratio3_6, Ratio12_56 and Ratio123_456

AsIs Transformations are included as transformations by default but could be dropped by the modeler, whilst all the others are considered optional and are included by the modeler, on demand.

The new variables created have a name format similar to Concatenation (InitialVariableName, Transformation) E.g. AmountPaidAvg6 for the average for last six months for the initial AmountPaid variable.

f) Flattening

The design of the flattening phase assumes the conclusion of all the transformations and involves creating a single entry for every participant.

The design provides for a single flattening action by the modeler.

g) Sampling

Random sampling is inbuilt in every sampling phase of the modeling process. The implementation of random sampling is by the use a random number generation system that uses the algorithm outlined below.

i. Get the list of all participants who qualify for the Learning phase of the modeling project.

ii. For each of the above, generate a random number ranging from 0 to 1.

iii. If the random number generated in ii above is less or equal to 0.7, then mark the record as in-sample, otherwise mark as out-of sample.

h) Variable Re-Categorization

Re-Categorization in prototype implementation is for both Categorical and Continuous variables.

The Re-Categorization process is guided by the use of and dependence on similarity of even occurrence percentages within the different categories.

The aim in Re-Categorization is to group together participants with similar event occurrence profiles.

The implementations are as follows:-

I. Re-categorization for Categorical variables

Categorical variables can have their categories merged to form new categories. The new categories created are thus larger as units but make for an overall fewer count of categories.

II. Re-categorization for Continuous Variables

The Re-Categorization of continuous variables is, however, not as straight forward as that of Categorical variables.

The Re-categorization algorithm outlined in section 3.6 is automated for use in the Re-Categorization process.

i) Variable Rankings

The ranking of variables in the system design phase is implemented as a one action point. The rankings use the Crammer's V greedy algorithm.

The implementation in the prototype involves the creation of two Correlation Matrices.

I. Correlations with Dependent Variable (WithY)

This is a listing of all the independent variables and the associated Cramer V values against the Dependent (Y) variable.

II. Correlations between Independent Variables (BetweenX)

This is a listing of all the independent variables and the associated Cramer V values against the set of all independent variables.

It is thus an n X n matrix showing the correlation between the independent variables.

j) Model Building and Assessment

The design of the model building and assessment module is as a variable selection into a model, the automatic creation of the selected model and an outline of the related assessment criteria.

k) Model deployment

The design of model deployment follows from the methodology and is in the trio of steps.

- Generate Score cards
- Group to Buckets
- Production (Credit Score)

The model deployment steps above are designed as a sequential and parameterized process but with support for iteration.

4.3.4 System testing

Unit testing and User Acceptance Testing (UAT) were the most commonly used software testing strategies used within the Extreme Programming (XP) methodology adopted.

XP, in its very nature does not follow the traditional software development stages of formal analysis and design. This is however, made up for by intensive system testing specifications.

Unit testing involved the testing of each of the modules outlined above as if they were complete systems on their own. In this study, unit tests were developed in parallel with the coding process. A set of unit test was developed and after each change to any of the modules, all the unit tests were re-run to ensure that together with the iterative builds with small increments approach, no major faults were introduced (regression testing).

User Acceptance Testing (UAT) was implemented by the developer role playing a user. The user (developer in this case) would come up with a list of all desired functionalities and system behavior and ensure these are met. At each development milestone, the system was evaluated against mock user requirements prior to delivery.

4.3.5 Limitations of the system design

The greatest limitation of the design is that it does not take care of such transformations as moving averages and both scalar and vector variable transformations. This limitation is however solved by the use of transcript transformations that are however, complex and thus difficult to use.

4.3.6 System installation and deployment strategy

The system is deployable from a windows installer that is provided with an installation kit. The installation procedure is bundled in the provided installation kit.

4.4 Design of Experiment

Whereas experimental designs involve manipulation of one or more Independent Variables (IV) and can therefore have causality potentially inferred, Correlational designs examine the relationship between two or more existing and non-manipulated variables.

This study adopts the Non-experimental, Correlational design approach. This is since the study does not involve the manipulation of the participants in the study. The experimentation process is thus designed as a Correlational study, in which the correlation between defaulting on credit repayments (the dependent variable) and other provided variables in a panel dataset is sought.

It is therefore imperative in this study that causality cannot be inferred. Creative interpretation of data and results is thus key to interpretation of the results of this study.

The process of experiment design process and results of the actual experimentation are as discussed in Chapter 5.

CHAPTER 5: MODELING RESULTS AND ANALYSIS

5.1 Introduction

This chapter outlines the results of the modeling process. The experimentation phase of this study was carried out using the developed and deployed prototype. The prototype was given the working name of ***Logistica®***. For this study, therefore, the modeling process is equally the experimentation process. The terms experimentation and modeling process are thus be used interchangeably.

5.2 The experimentation process setup

The requirements of the modeling process and as outlined below:-

i. Panel Basic Dataset in SQL Server 2008 database table (CsCreditCard) and in the requisite structure as outlined in the design section (4.2) above.

ii. An operational deployment of the modeling prototype (Logistica®) and associated infrastructure.

iii. Requisite access privileges set up by the administrator.

The research results presented here are a reflection of the results generated in the course of the development of prediction models using *Logistica®*. The models used in the experimentation process are bundled in the prototype infrastructure.

5.3 Settings used for experimentation and generation of results

The creation of models using *Logistica®* follows the ETSRMA methodology outlined in section 3.1. The process of building a model to predict the probability of default from the credit card data uses the settings in Table 5.1.

Table 5.1 Settings for Experimentation

Setting	Definition	Description
ProjectName	Model2	The Name of the modeling project. Could be used for multiple models
ProjectDescription	Credit Card Default Prediction Model	This is brief description of the modeling project.
DataSourceQuery	Select * From CsCreditCard	Query that defines the set of columns to be used in the modeling project. This defaults to all the columns
DependentVariable	Max(Case When Validitydate Between @mm1f And @mm12f And Dayspastdue>90 Then 1 Else 0 End)	This is the definition of what is to be predicted. This has the default setting value of predicting default 12 months in the future
ObservationPoint	200908	The end point of learning

5.4 Modeling results

One set up as outlined in 5.2 above, the following results were obtained from the modeling process:-

a) Explore Results for dataset

Each variable (both from the provided dataset and the transformed variables) can be analyzed using Logistica®. The types of analysis that are supported include:-

I. Categorical analysis (CAT)

This form of analysis only achieves the display and plot of frequencies for the selected variable.

The output for Categorical analysis is either by

i. The plot of the overall variable analysis over time or

ii. The plot of a specific variable value over time

Illustrations of these are as outlined in Figure 5.1 and Figure 5.2.

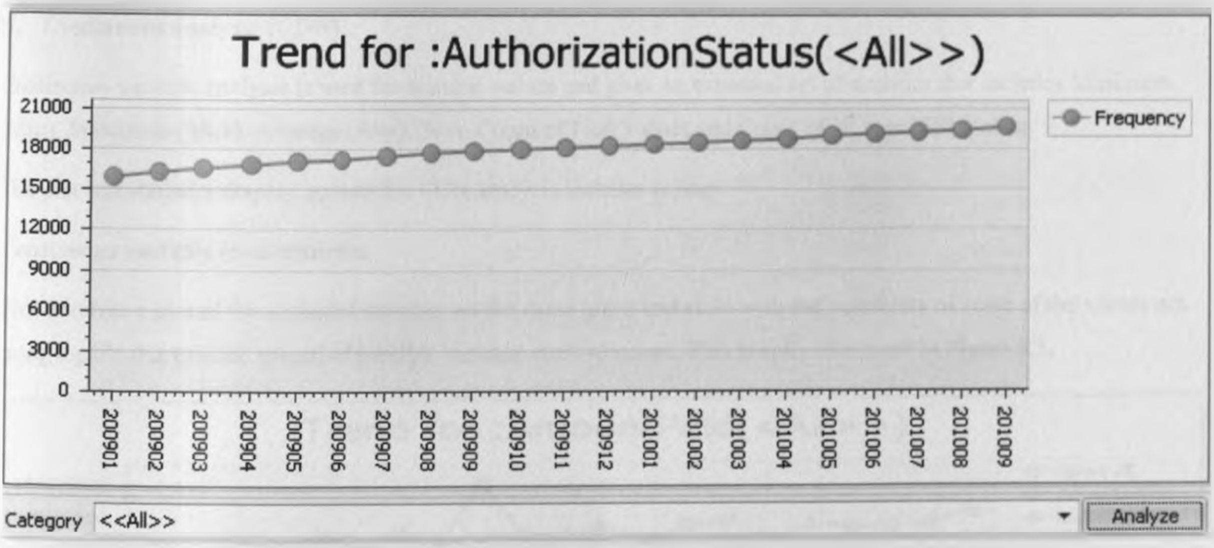
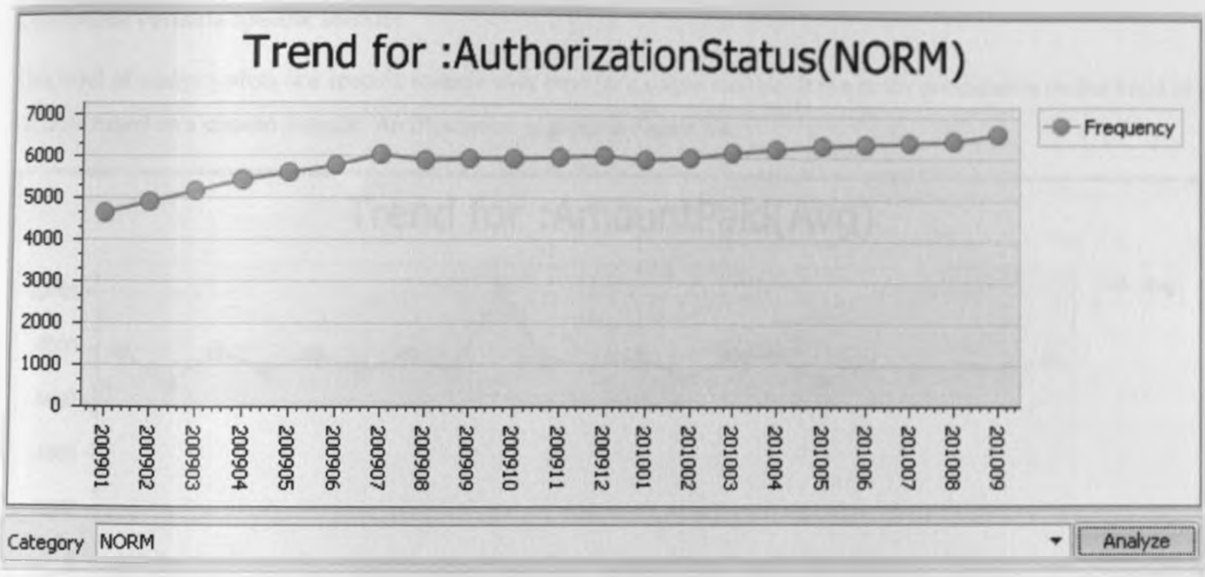


Figure 5.1 Categorical variable analysis

Categorical analysis (CAT) can be used even on numeric values but does not make for the best analysis where a variable value assumes an unbounded range. Interpretation in such cases and comprehension are increasingly impaired.



Category NORM

Figure 5.2 Categorical variable value analysis

II. Continuous analysis (CON)

Continuous variable analysis is used for numeric values and gives an extended set of statistics that includes Minimum (Min), Maximum (Max), Average (Avg), Sum, Count of Null Values and Count of all Non-Null Values.

The plot and statistics display options for CON analysis includes either:-

Continuous variable level statistics

This involves a plot of the extended statistics on the same graph and scale with the possibility of some of the values not being legible due to wide spread of possible variable statistic values. This is aptly illustrated in Figure 5.3.

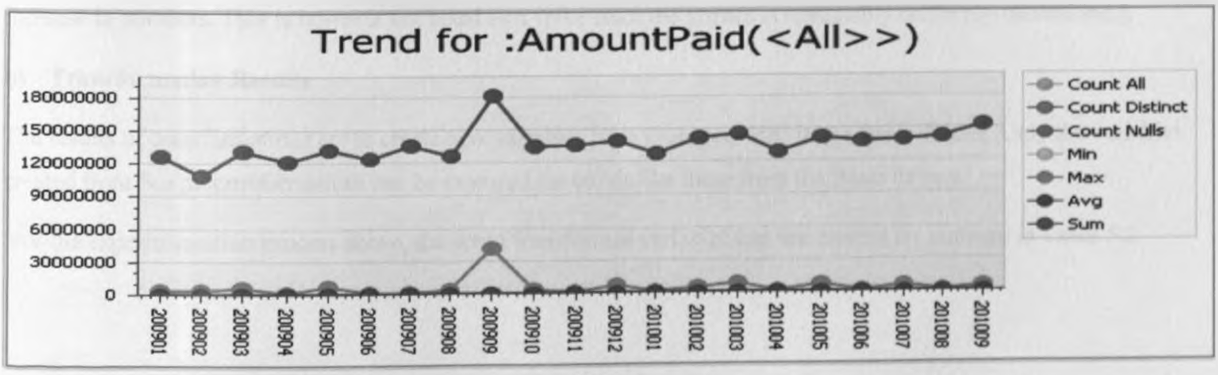


Figure 5.3 Continuous variable analysis

Continuous variable specific statistic

This level of analysis plots one specific statistic over time for a single variable. It is a better presentation on the trend of a variable based on a chosen statistic. An illustration is given in Figure 5.4.

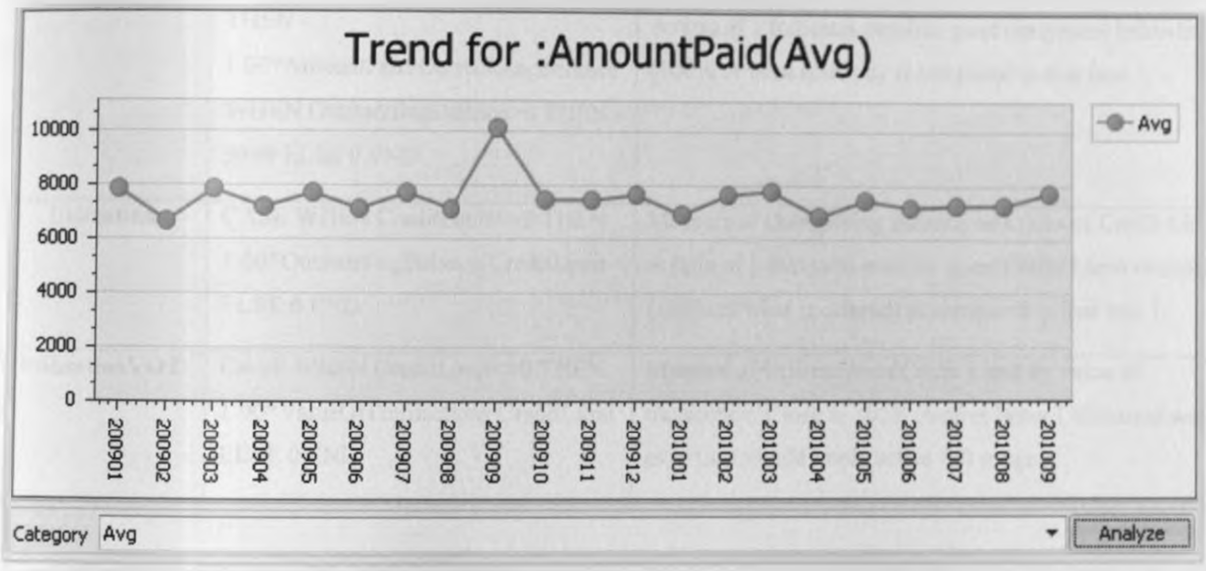


Figure 5.4 Continuous variable specific statistic analysis

The results of the explore phase revealed an acceptable growth trend for most of the variables except for AmountPaid and the BrandName value Local Classic. The business explanation to the notable spikes on AmountPaid , in Figure 5.3, was the intensification of collection efforts within the month that saw more people paying up on their credit card debts. The decline in the counts of credit cards by the BrandName value Local classic was attributed to a change in policy that was converting them to mostly to International Classic cards (International Classic has a corresponding growth due to the increase in numbers. This is however not noted as a spike since the growth is reasonably uniformly maintained.).

b) Transformation Results

The results of transformations are to create new variables from those provided in the Basic dataset. Only the variables created from Script transformations can be explored for trends like those from the Basic dataset.

For the experimentation process above, the script transformed variables that are created are outlined in Table 5.2.

Table 5.2 Script Transformations

Variable	Transformation	Description
AmtOfOsbal	CASE WHEN OutstandingBalance<0 THEN -1.00*AmountPaid/OutstandingBalance WHEN OutstandingBalance>0 THEN -9999 ELSE 0 END	Measure of AmountPaid as a ratio of Outstanding Balance. A ratio of 1 Indicates possible good repayment behavior (100% of what is owed) as compared to that less 1.
Utilization	CASE WHEN CreditLimit<>0 THEN 1.00*OutstandingBalance/CreditLimit ELSE 0 END	Measure of Outstanding Balance as a ratio of Credit Limit. A ratio of 1 Indicates possible good Credit Limit utilization (100% of what is offered) as compared to that less 1.
UtilizationVOT	CASE WHEN CreditLimit<>0 THEN 1.00*ValueOfTransactions/CreditLimit ELSE 0 END	Measure of Utilization of Credit Limit by value of transaction. Close to 100% implies better Utilization and over time would imply active full usage.

The data exploration results of the UtilizationVOT variable, based on averages over time for the whole population is as outlined in Figures 5.5.

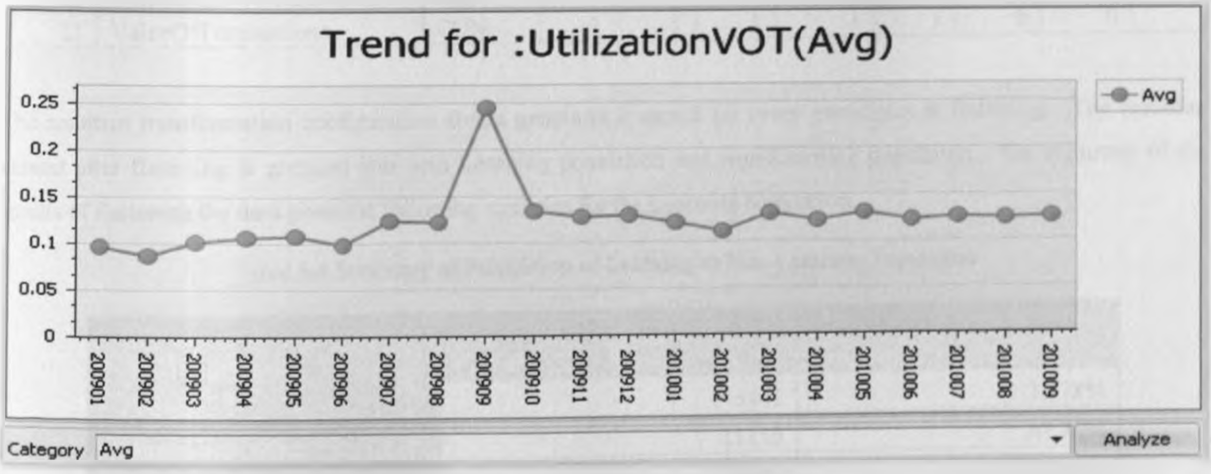


Figure 5.5 Trend for Average Utilization by Value of Transactions (a script transformed variable)

c) Data Reduction (Flattening) Results

From the experimental modeling process, the flattened data had as many variables as had been selected from the transformation point.

The following table shows the result of variable configure for transformation.

Table 5.3 Resultant Transformation Configuration Matrix

Serial No	Name	Analysis Type	AsIs	Avg6	Avg3	Avg Woext	Ratio 1_6	Ratio 3_6	Ratio 12_56	Ratio 123_456
1	AmountPaid	CON	1	1	1	1	1	0	0	0
2	AmountPastDue	CON	1	1	1	1	1	0	0	0
3	AmtOfOsbal	CON	1	0	0	0	0	0	0	0
4	AuthorizationStatus	CAT	1	0	0	0	0	0	0	0
5	BillingCycle	CAT	1	0	0	0	0	0	0	0
6	BrandName	CAT	1	0	0	0	0	0	0	0
7	CreditLimit	CON	1	1	0	0	1	0	0	0
8	DaysPastDue	CON	1	1	1	0	0	0	0	0
9	FinancialStatus	CAT	1	0	0	0	0	0	0	0
10	GeneralStatus	CAT	1	0	0	0	0	0	0	0
11	InternalEmployee	CAT	1	0	0	0	0	0	0	0
12	NumberOfCards	CON	1	1	0	0	0	0	0	0
13	NumberOfCashWithdrawals	CON	1	1	0	0	0	0	0	0
14	NumberOfTransactions	CON	1	1	0	0	0	0	0	0
15	OutstandingBalance	CON	1	1	1	1	1	0	0	0
16	OverdrawnAmount	CON	1	1	0	1	0	0	0	0
17	OverdueCycles	CAT	1	0	0	0	0	0	0	0
18	Tenure	CON	1	0	0	0	0	0	0	0
19	Utilization	CON	1	1	1	1	0	0	0	0
20	UtilizationVOT	CON	0	0	0	0	0	0	0	0
22	ValueOfCashWithdrawals	CON	1	1	1	1	0	0	0	0
23	ValueOfTransactions	CON	1	1	1	1	1	0	0	0

The resultant transformation configuration above generates a record for every participant at flattening. The resulting dataset after flattening is grouped into both Learning population and Non-Learning population. The summary of the results of flattening the data gives the following statistics for the Learning population.

Table 5.4 Summary of Proportion of Learning to Non-Learning Population

PopulationDetails	Count	Proportion
Learning	5441	32.78%
Out of Learning	11159	67.22%
Summary(Total)	**16600**	**100.00%**

The resultant summation of the values of the Transformation Configuration Matrix (Table 5.3) gives the total number of variables that make past the Flattening phase to the Sampling phase.

d) Sampling Results

The Learning population from the dataset (a total of 5441 participants) is considered the population f the Learning Phase of the Modeling process.

The results of the Sampling phase were as shown in the Table5.5.

Table 5.5 Sampling Statistics

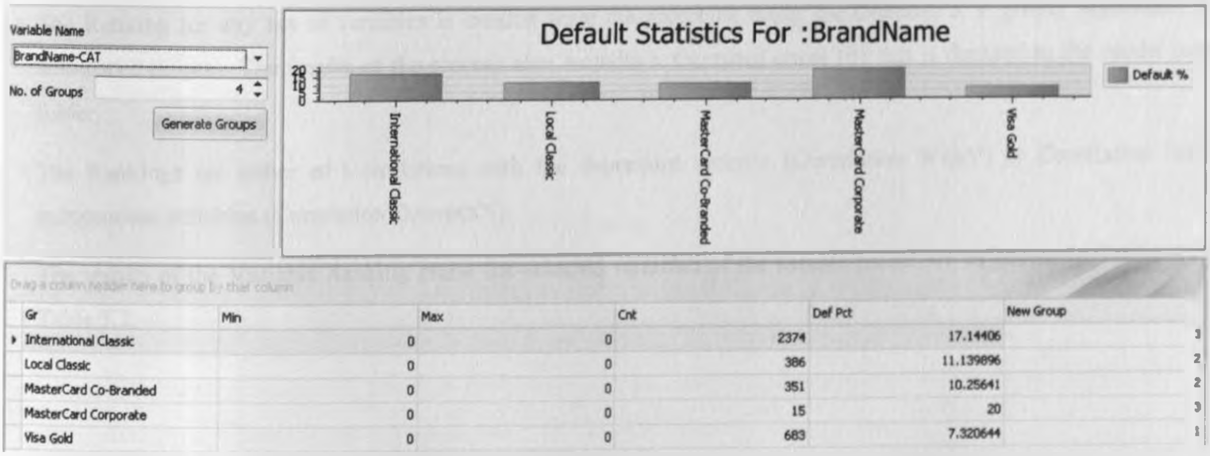
Population	Details	Count	Proportion (%)	Default (%)
In-Sample	To be used for Learning(Training Set)	3,809	70.01	16.48
Out of Sample	To be used for Validation(Validation Set)	1,632	29.99	14.29
Summary(Total)		**5,441**	**100.00**	**15.82**

e) Variable Re-Categorization Results

The Variable Re-Categorization

The Variable Re-Categorization for the sample model created assumes a maximum of four groups for every variable that was selected.

Sample results for Re-Categorization of Categorical variables are shown in Figure 5.6, whilst that for a Continuous variable is shown in Figure 5.7.



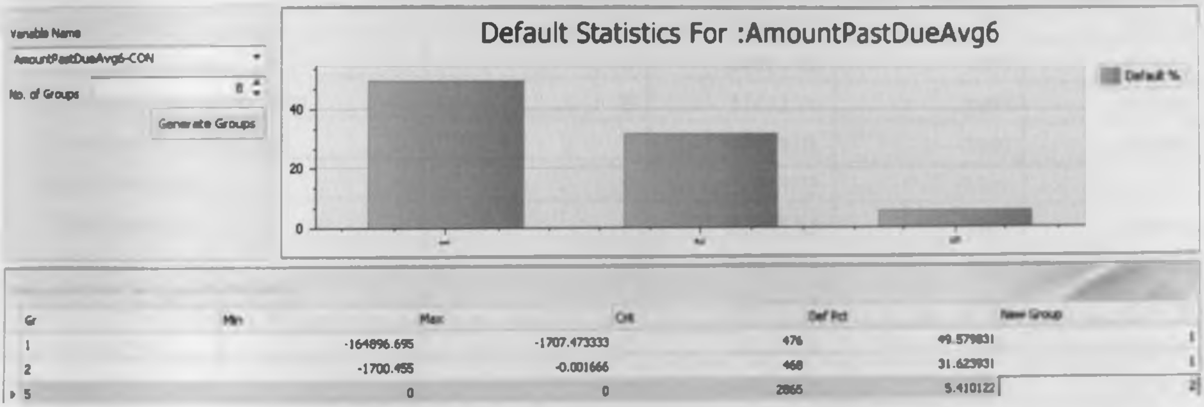
Variable Name: BrandName-CAT

No. of Groups: 4

Gr	Min	Max	Cnt	Def Pct	New Group
International Classic	0	0	2374	17.14406	3
Local Classic	0	0	386	11.139896	2
MasterCard Co-Branded	0	0	351	10.25641	2
MasterCard Corporate	0	0	15	20	3
Visa Gold	0	0	683	7.320644	1

Figure 5.6 Re-Categorization for Categorical Variable (BrandName)

From the results of the above Re-Categorization, it is clear business wise that Visa Gold cards are the least defaulters and thus least risky cards (7.3%), followed by the pair of Local Classic and MasterCard Co-Branded (approximately 11%) and finally the riskiest of all Card Brands is the pair of International Classic and Master-Card Corporate. It is therefore evident that Categorical variables can be grouped based on difference and similarity of behaviors to generate new variables.



Variable Name: AmountPastDueAvg6-CON

No. of Groups: 8

Gr	Min	Max	Cnt	Def Pct	New Group
1	-164896.695	-1707.473333	476	49.579831	1
2	-1700.455	-0.001666	468	31.623931	1
5	0	0	2865	5.410122	2

Figure 5.7 Re-Categorization for Continuous Variable (AmountPastDueAvg6)

The Re-Categorization for the variable AmountPastDueAvg6 creates two categories with the business result that whose average past due amount for the last six months exceed Zero (0) are riskier customers (default rates greater that 31%) as compared to those who have a Zero (0) Six (6) month average past due amount (5% default rate).

The results for Re-Categorization clearly show that Continuous variables can be transformed to Categorical variable through the choice of appropriate multiple thresholds.

f) Variable Ranking Results

The Ranking for any set of variables is created from the prototype using the Crammer's V greedy algorithm, in an automated process. The results of the process also includes a formatted excel file that is dumped to the model outputs folder.

The Rankings are either of Correlations with the dependent variable (Correlation WithY) or Correlation between independent variables (Correlation BetweenX).

The results of the Variable Ranking phase for selected variables of the sample model are as shown in Tables 5.6 and Table 5.7.

Table 5.6 Correlation between Dependent and Independent Variables

Variable	NCategories	Chisq	Cnt	Cramer'sV
OverdueCycles	4	96844.75	3809	50.42%
DaysPastDueAvg6	6	64964.86	3809	41.30%
OverdrawnAmountWOExt	6	51472.16	3809	36.76%
OverdrawnAmountAvg6	6	46645.21	3809	34.99%
UtilizationAvg6	6	37003.73	3809	31.17%
UtilizationAvg3	6	27635.05	3809	26.94%
AuthorizationStatus	4	20076.24	3809	22.96%
OutstandingBalanceAvg3	6	11213.17	3809	17.16%
AmountPaid	4	9823.54	3809	16.06%
InternalEmployee	4	9721.54	3809	15.98%
ValueOfCashWithdrawalsAvg6	4	6070.12	3809	12.62%
NumberOfTransactions	6	5539.77	3809	12.06%
BrandName	6	5120.36	3809	11.59%
NumberOfCards	4	1910.95	3809	7.08%

Table 5.7 Sample Correlation between Independent Variables

	AmountPaid	AuthorizationStatus	BrandName	DaysPastDueAvg6	Internal Employee
AmountPaid	100%	15%	9%	12%	4%
AuthorizationStatus	15%	100%	10%	9%	5%
BrandName	9%	10%	100%	12%	20%
DaysPastDueAvg6	12%	9%	12%	100%	36%
InternalEmployee	4%	5%	20%	36%	100%

g) Model Building and Assessment (Validation) Results

I. Overview of the built model

The variables in the sample model built included: AmountPaid, AuthorizationStatus, DaysPastDueAvg6, InternalEmployee, NumberOfCards, NumberOfTransactions, OverdrawnAmountAvg6, OverdueCycles and UtilizationAvg3. The model had an out-sample Gini-Index of 79.85%.

Score Card

The Score Card, based on the selected variables from the model building phase is as shown in Table 5.8. For the sample model, a total of 9 variables were used. Higher scores equate higher risks. The coefficient of each variable category corresponds to the predictive weight that is assigned to that particular variable value.

II. Model Validation Results

Model validation results have their summary outlined in Table 5.9. A distinction is given for both in-sample and out of sample model validation results and an all sample validation result as is appropriate.

Table 5.8 Sample Model Score Card

Variable	Category	Coeff	Score
AmountPaid	1	0	36.88
AmountPaid	2	-0.33	0
AuthorizationStatus	1	0	68.65
AuthorizationStatus	2	-0.62	0
DaysPastDueAvg6	1	0	0
DaysPastDueAvg6	2	0.82	90.55
DaysPastDueAvg6	3	1.67	185.59
InternalEmployee	1	0	183.97
InternalEmployee	2	-1.66	0
NumberOfCards	1	0	46.06
NumberOfCards	2	-0.41	0
NumberOfTransactions	1	0	89.09
NumberOfTransactions	2	-0.54	29.39
NumberOfTransactions	3	-0.8	0
OverdrawnAmountAvg6	1	0	0
OverdrawnAmountAvg6	2	0.44	48.93
OverdrawnAmountAvg6	3	0.9	99.74
OverdueCycles	1	0	0
OverdueCycles	2	1.2	133.23
UtilizationAvg3	1	0	156.81
UtilizationAvg3	2	-0.57	93.18
UtilizationAvg3	3	-1.41	0

Table 5.9 Model Validation Results Summary

Segment	Capture 20%	Gini-Index	Count	Count of Default	Default%
In-Sample	72.5%	77.56	1,632	204	12.50
Out-Sample	79.4%	79.85	3,809	539	14.15
All Sample	77.33%	78.71%	5,441	743	13.66

The model validation results summary gives the key statics of Capture at 20%, Gini-Index, Total section counts and counts of defaults within the sample and out of sample data sets. An indication is also given of the overall percentage of defaults in each of the in-sample and out of sample sets of data.

The key results of model variable categories and associated P-Vales, Odds Ratio and Standard Error of Mean are outlined in Table 5.10. For the sample model, all the variables are noted to have P-Vales <0.05 and are thus all considered significant. The Odds ratios of all the variables are noted not to cross the boundary of 1.

Table 5.10 Model Variable Coefficients and Selection Results

Variable	Category	Coeff	PValue	StdErr	Odds Ratio	Odds Ratio Lower	Odds Ratio Upper	Recomm
AmountPaid	2	-0.33207482	0.04667416	0.1669	0.7174	0.5172	0.9951	Accept
AuthorizationStatus	2	-0.618176029	0.000308205	0.1713	0.5389	0.3852	0.7540	Accept
DaysPastDueAvg6	2	0.815428823	1.04E-06	0.1670	2.2601	1.6293	3.1353	Accept
DaysPastDueAvg6	3	1.671227265	5.46E-23	0.1693	5.3187	3.8170	7.4113	Accept
InternalEmployee	2	-1.656636046	7.84E-13	0.2312	0.1908	0.1213	0.3002	Accept
NumberOfCards	2	-0.414734196	0.005701298	0.1500	0.6605	0.4922	0.8863	Accept
NumberOfTransactions	2	-0.537605094	0.00038101	0.1513	0.5841	0.4342	0.7858	Accept
NumberOfTransactions	3	-0.802280851	1.19E-05	0.1832	0.4483	0.3131	0.6420	Accept
OverdrawnAmountAvg6	2	0.440623864	0.012136612	0.1757	1.5537	1.1011	2.1923	Accept
OverdrawnAmountAvg6	3	0.898124323	1.38E-06	0.1860	2.4550	1.7049	3.5350	Accept
OverdueCycles	2	1.199743498	6.92E-12	0.1749	3.3193	2.3559	4.6765	Accept
UtilizationAvg3	2	-0.572975236	0.000500979	0.1646	0.5638	0.4083	0.7786	Accept
UtilizationAvg3	3	-1.412039637	1.04E-10	0.2186	0.2436	0.1588	0.3739	Accept

Model Assessment and Validation Charts

Figures 5.8-5.10 give pictorial representations of the Capture-Response chart, Lift Chart and ROC Chart of the sample model created. A discussion on the charts and their results is done in Chapter 6.

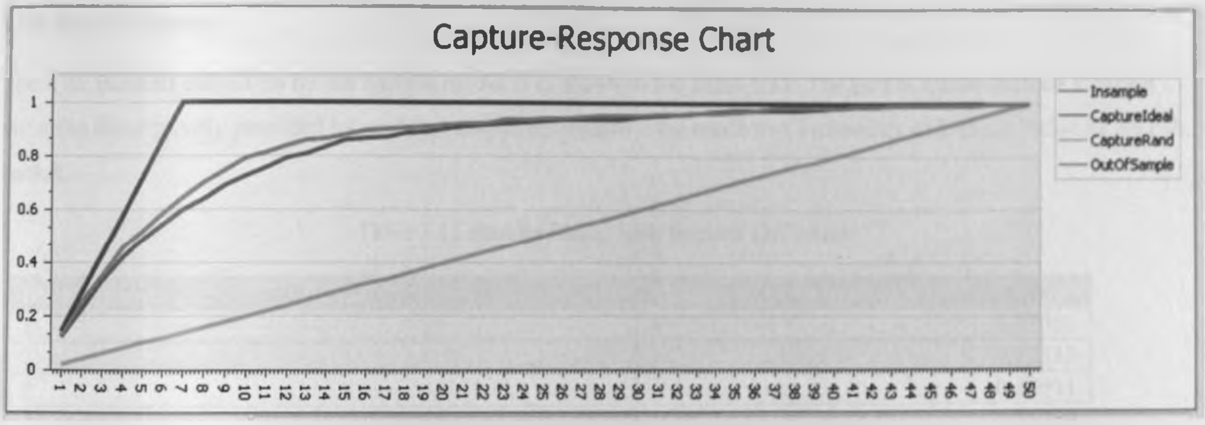


Figure 5.8 Model Capture-Response Chart

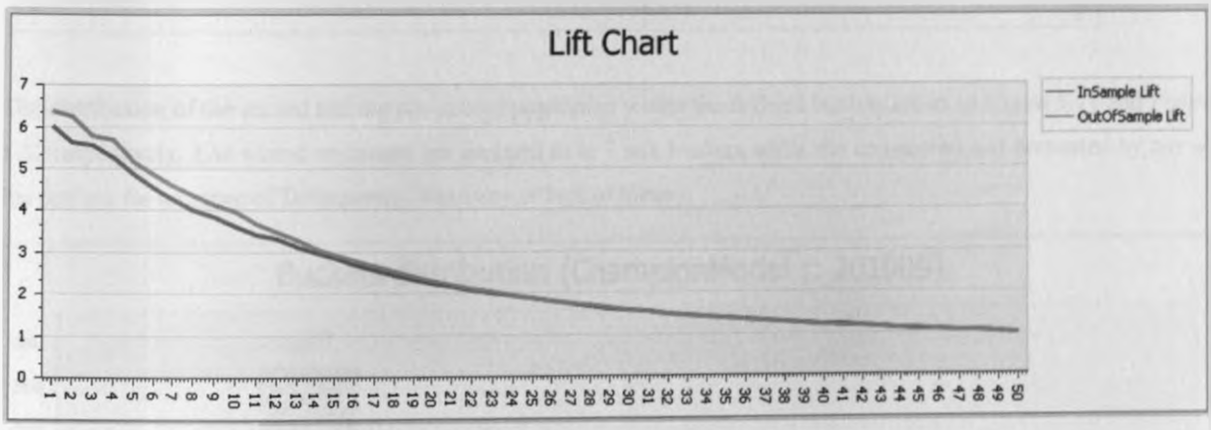


Figure 5.9 Model Lift Chart

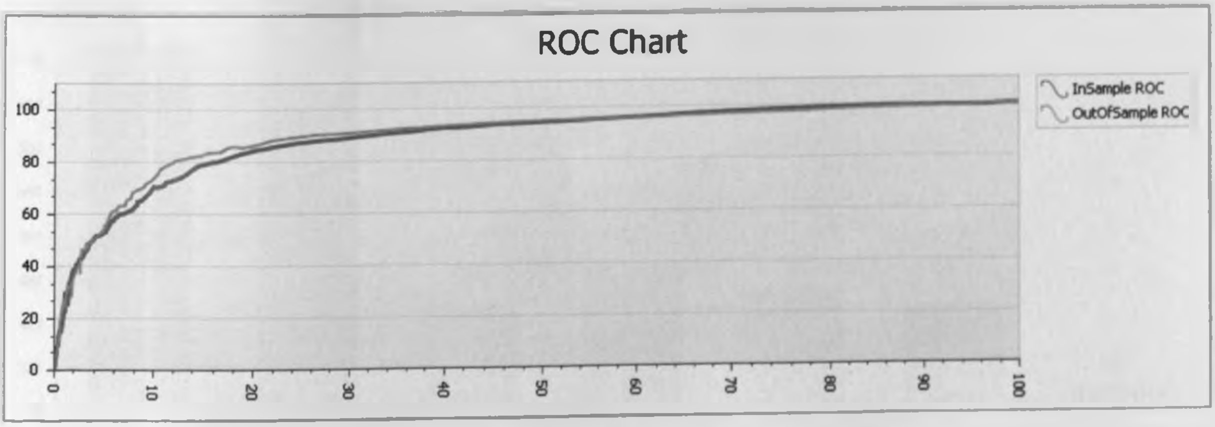


Figure 5.10 Model Receiver Operating Characteristic (ROC) Chart

Risk Bucket Results

The Risk Buckets definition for the Sample model is as shown in the Table 5.11. The bucket results include a bucket name (as descriptively provided by modeler) and an indication of the maximum Probability of Default Value of the risk bucket.

Table 5.11 Sample Model Risk Buckets Definition

BucketName	MaxPredictionValue
1	0.03062
2	0.12383
3	0.18011
4	0.35759
5	0.63722
6	0.8807
7	1

The distribution of the scored and the non-scored population within the defined buckets are as in Figure 5.11 and Figure 5.12 respectively. The scored customers are grouped in to 7 risk buckers while the non-scored and presented by any of the reasons for no score of Delinquency, inactivity or lack of history.

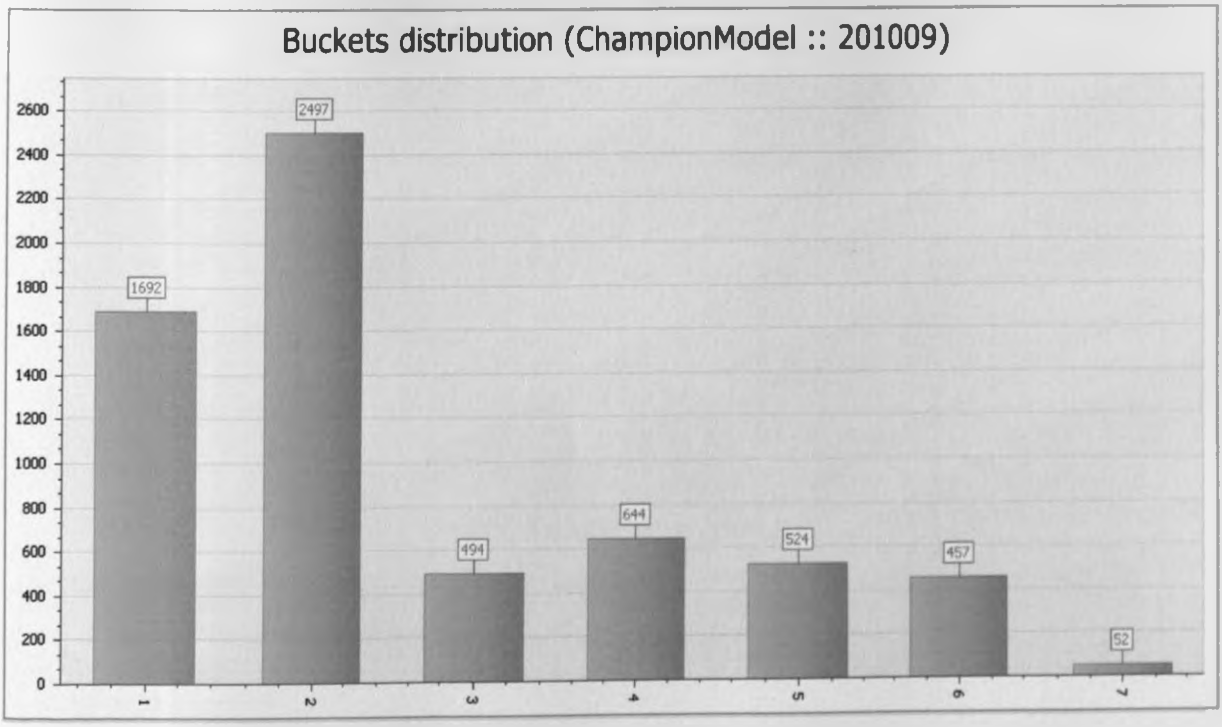


Figure 5.11 Risk Buckets Distribution by Population of scored customers

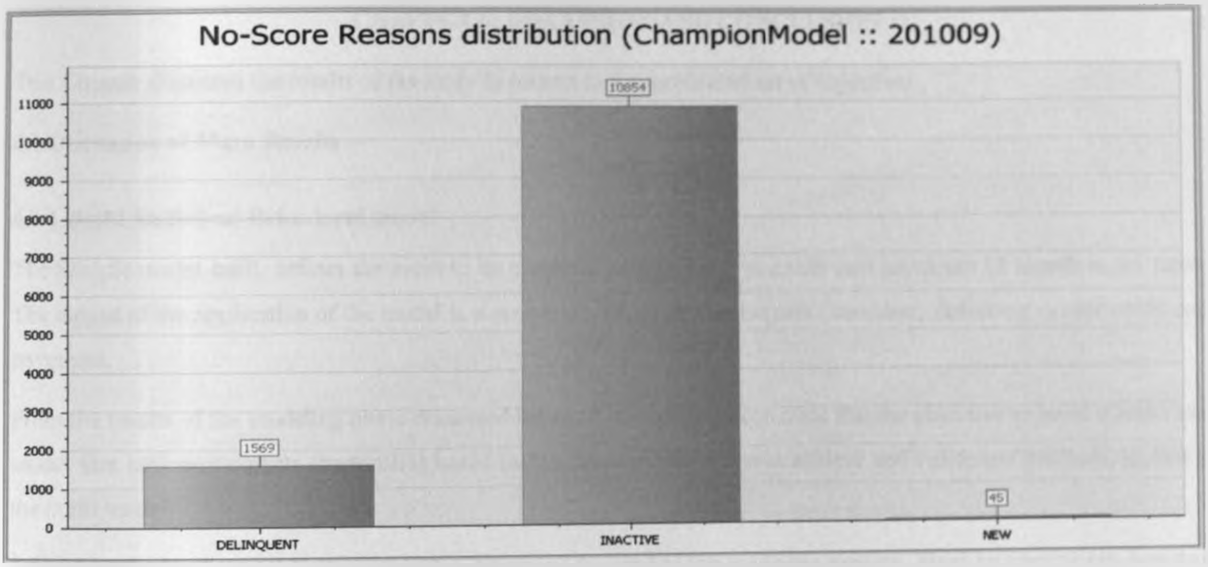


Figure 5.12 Distribution of Population by No-scored reasons

CHAPTER 6: DISCUSSION AND CONCLUSION

This Chapter discusses the results of the study in respect to the formulated set of objectives.

6.1 Discussion of Main Results

6.1.1 Build Statistical Behavioral model

The Sample model built, defines the event to be predicted as defaulting on credit card payments 12 month in the future. The output of the application of the model is a probability [0, 1] of a participant (customer) defaulting on the credit card payments.

From the results of the modeling phase discussed below, it is justified to conclude that the objective to build a behavioral model that rank participants (customers) based on the future credit risk was achieve and validation methods applied to the build model.

Several methods of model Validation and Assessment are used in the modeling process. Their interpretations based on the sample model are as follows:-

i. Gini-Index (Lorenz Curve/Capture-Response Chart)

The Capture-Response Chart in Figure 5.3.6 indicates the following:-

- ❖ The Random model would have a 50% predictive power over the entire population. This follows from the definition of random selection.
- ❖ The Ideal model would get require only 13.7% of the population to list all the defaulting members of the population.
- ❖ The In-sample model requires 20% of the population to list about 73% of the defaulting members of the entire population.
- ❖ The Out-Of sample model requires 20% of the population to list about 80% of the defaulting members of the entire population

The interpretation of the Gini Index is guided by the Table 6.1.

Table 6.1 Gini Index Range Interpretation

Meaning	Index Range
Model ranking is as good as a random ranking	0
Mild Improvement	0-0.4
Moderate Improvement	0.4-0.6
Good Improvement	0.6-0.7
Very Good Improvement	0.7-0.8
Excellent Improvement	>0.8

It is worth the note that a good Gini does not imply a good model and vice versa. The Gini for the sample model therefore implies a Very Good Improvement over use of a random model.

The Gini-Index and indication of Over-fitting

The results of the modeling process as shown in Table 5.9 indicate that it is possible to indicate the possibility of occurrence of over-fitting in the modeling process.

An indication of the occurrence of over-fitting based on the Gini-Index measure for the Learning set versus the Validation set is given by the implementation of the modeling process in the prototype.

The Sample model is shown to perform better on the Validation set of unseen example (79.85% to 77.56) over the Learning set. This model is thus better predictive in the real world and is thus, in essence, under-fitted.

A difference of more than two percentage points between In-Sample and Out-Sample Gini (in favor of In-Sample Gini) is considered over-fitting.

ii. Variable Selection, P-Value and Odds Ratio

The process of variables is guided by the statistical measures of P-Value and Odds Ratio. The Null Hypothesis (Ho) is stated as: All variable coefficients are not significant. Variables are accepted when:-

i) the P-Value of their coefficients is less than 0.05 and

ii) The Lower and Upper boundaries of the Odds ratio do not include 1. Since Exp(Coeff)=1. This implies a zero coefficient and thus the variable is not significant.

It is clear from the results in Table 5.10 that all the chosen variables meet these conditions. The interpretation of P-Values and Odds ratio have equally been simplified by the use of the traffic light metaphor and thus taking away the need for advance statistical interpretations that go with statistical modeling techniques.

iii. Gini, ROC and AUC

When changing the cut-off values for a model, to convert the output from real to Boolean, the values of the measure changes.

There then has to be a trade-off between the sensitivity and specificity. A higher cut-off means lower sensitivity and a higher specificity.

Area under the Curve (AUC), is the area under the ROC curve and by a rule of thumb, 2AUC-1=Gini.

6.1.2 Prototyping and Application of data mining concepts

The second objective of this study was to build prototype software that uses the concepts of data mining to be used in the building of models using the statistical method of Logistic Regression (LR).

This objective was achieved by the building of the prototype software Logistica®. The prototype automates the model building process using the ETSRMA methodology. The Exploration of data implements the ability to do exploratory data analysis (EDA). The prototype has the support for statistical data mining techniques and can be used to determination of probability for the occurrence of events using LR. The definition of the event to be predicted is delineable within the prototype.

6.1.3 Calibration of Probabilities to Score Cards

The third objective of this study was the conversion of the generated probabilities of defaulting of customers so as to formulate score cards.

Score cards mathematically model future display of certain behavior by a customer. In this study, score cards mathematically model the probability of defaulting on credit card loan repayments, one year in the future.

The prototype has an inbuilt statistical model for the development of score cards as shown in Table 5.8 and Figures 5.11 and 5.12.

The generation of score cards helps in the implementation of models in business environments as many customers with similar future risk profiles can be grouped together and accorded uniform consequence management.

Possible ways of implementing the results of the score cards based on the generated risk buckets in Figure 5.11 includes:-

- Award of higher Credit Limits to customers in Risk buckets 1 and 2
- Lower interest rates for customers in Bucket 1
- Debt restricting for customers on Risk Buckets 6 and 7 including conversion from Credit Card loans to personal credit since credit card facilities are normally very highly priced.

6.1.4 Automation of model building process

The prototype is an illustration of the possibility of automation of the modeling process, the selection of variables and integration of the model building environment. The integration enables non statistical users to successfully build models without the encumbrances of statistical knowledge.

6.2. Contributions of Research

This study achieves some level of automation of the modeling process for the prediction of probability of the occurrence of events using Logistic regression (LR). This automation reduces on the statistical knowledge requirements that limit the building and use of statistical models.

The usage of design metaphors like the traffic light metaphor make the modeling process easy to understand thereby improving on comprehensibility, a limitation that exists in most modeling processes.

A common limitation of most modeling process is lack of transparency in the modeling process. This is to a large extent solved by the prototype implementation adopted since the creation of data files that are used for model validation ensures ease of verification.

The research illustrates the possibly benefits of monitoring existing credit holders using behavioral credit scores as opposed to the use of application score cards that are at credit initiation. The use of behavioral score cards also eliminates the use of customer subjective information.

6.3 Limitations of the research

The major drawback of the research and prototype developed is on the over-reliance on evidence as provided in the data. This implies that statistically significant variables that are not reflected in the data do not make it to the models. The reliance on data to build model is solved by building business rules around implementation of models in the real world. Equally, the reliance of one method of modeling (LR in this case) could be misleading since better results could lie in the deployment of competing modeling methods.

The other limitation of the modeling methodology adopted is the non-responsiveness to changes in external factor such as inflation and otherwise, that can affect customer behavior. Some of the behavioral measures adopted would not take into account such external factors and thereby end of giving not real world measure of customer initiated behavior changes. The best way in practice is refreshing and rebuilding of models.

The design of the prototype provides for only a set of few transformations that could be done. This might not be as is desired in the real world where such measures as moving averages and vector transformations (that are provided for in the prototype) are the norm.

Finally, the experimental design of this study is Correlational and therefore does not provide for an explanation on the relationships that are realized between the dependent variable and independent variables. Causality cannot be inferred in this kind of experimental design.

6.4 Conclusion

The key findings of this study form the basis of the following conclusions.

i. **Use of validation methods from other fields**

   Credit scoring for prediction of defaulting on credit payments has the same ultimate goal as credit scoring in other domains. The goal is the discrimination between good and bad obligors based on the weight of evidence provided as a set of selected indicators/variables/features. The use of model validation measures such as the Receiver Operating Characteristic (ROC) curve, the Gini-Index and Percentage capture from the related domains can therefore be used as measures of weight of evidence.

ii. **Non-existence of superior credit scoring methods**

There is not an agreed method that fits all credit scoring requirements. The variance in requirements (in terms of outputs desired) based on the data present in terms demands the deployment of diverse methods. A ranking of the credit scoring techniques in any order of superiority is thus not feasible.

iii. **Trade-off between Interpretability and Predictive power**

Score cards are akin to decision rules. Their interpretation is easiest when they are as few as is feasible. On the other hand, discrimination of models mostly improves with the addition of many variables into the model. The twin factors of number of variable in a model (and their powers of discrimination) versus the interpretability of the generated score cards is thus a trade-off to be decided on by the modeler.

iv. **Multiple Validation methods**

The aim of credit scoring models is discrimination between bad and good credit holders. There however does not exist a perfect ranking system that offers perfect discrimination between defaulters and non-defaulters. The aim is to increase on the odds of correct predictions based on reducing false alarm and increasing the hit rate (from the ROC curve), increasing predictive power by increasing the Gini-Index to excellent levels of improvement (>80%) and ensuring that the difference between in-sample and out-sample Gini indexes does not make for over-fitted models.

v. **Waning or model predictive power**

The validity of a prediction model is for only a period of time. Our study therefore notes the need for periodic rebuild of the models and provides for a flexible methodology that accommodates multiple models being built from the same data set.

vi. **Deployment and Transformations for better predictive variables**

Most models developed and studies in modeling do not have deployment strategies as part of the deliverables. This study, however, provides for a model deployment strategy that is in our opinion a contribution to both academia and to financial institutions.

Finally, though complex at times, both script and extended transformations could end up in the creation of more predictive variables.

6.5 Future Work

Possible future work should involve the incorporation of alternative statistical and even Machine Learning methods that could then be used to generate outputs that are synergies of the strengths of different modeling techniques.

It is however worth noting that the delivered prototype has an extensible architecture since the LR method is just the fully built in one even though it could be modified to support the development of other statistical modeling methods.

The integration of the modeling environment that eliminates the need for statistical competencies is beneficial in improvement of the turn-around times of the model building process.

REFERENCES

Abdou, H. El-Masry A. and Pointon, J., 2007. On the applicability of credit scoring Models in Egyptian banks. *Banks and Bank Systems*, 2(1) pp. 4-20.

Altman, E. Bharath, S.T. and Saunders, A., 2002. Credit ratings and the BIS capital adequacy reform agenda. Journal of Banking & Finance, 26(5) pp. 909-921.

Anderson, R., 2007, The Credit Scoring Toolkit; Theory and Practice for Retail Credit Risk, Management and Decision Automation. New York: Oxford University Press.

Baesens, B. Van Gestel, T. Viaene, S. Stepanova, M. Suykens, J. and Vanthienen, J., 2003. Benchmarking State of the Art Classification Algorithms for Credit Scoring. Journal of the Operational Research Society, 54(6) pp. 627-635.

Bellotti, T., Crook, J. Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications 36, 3302-3308 (2009)

Basel Committee on Banking Supervision (2004), International convergence of capital measurement and capital standards: a revised framework, BIS report, June 2004.

Basel Committee on Banking Supervision (2006). Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework – Comprehensive Version. Basel: Bank of International Settlements.

Brill, J., 1998. The importance of credit scoring models in improving cash flow and collections. Business Credit, 100(1) pp 16–17.

Caire, D. Barton, S. Zubiria, A. Alexiev, Z. Dyer,, J. Bundred, F. Brislin, N. A Handbook for Developing Credit Scoring Systems In A Microfinance Context. In: USAID (United States Agency for International Development), microREPORT #66. 2006. Washington: United States.

Desai, V.S. Crook, J.N. and Overstreet, G.A., A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment. European Journal of Operational Research, 95 (1996) 24-37

Desai, V.S. Crook, J.N. and Overstreet, G.A., Credit Scoring Models in the Credit Union Environment Using Neural Networks and Genetic Algorithms. Computer Journal of Mathematics Applied in Business and Industry, 8(4) (1997) pp. 324-346.

Dodig-Crnkovic, G., 2002. Scientific Methods in Computer Science. Available at: < http://www.mrtc.mdh.se/publications/0446.pdf> [Accessed 01 March 2012].

Domingue, B. Briggs, D.C., 2009. Using Linear Regression and Propensity Score Matching to Estimate the Effect of Coaching on the SAT. [online] Available at: < http://www.colorado.edu/education/faculty/derekbriggs/Docs/Domingue_Briggs_Using%20Propensity%20Score%20Matching.pdf> [Accessed 20 February 2012].

Gestel, T.V., Baesens, B. Garcia, J. and Dijcke, P.V., 2003. A Support Vector Machine Approach to Credit Scoring. [online] Available at: < http://scorto.ru/downloads/A%20Support%20Vector%20Machine%20Approach%20to%20Credit%20Scoring.pdf > [Accessed 10 January 2012].

Glorfeld, L.W. Hardgrave, B.C., An improved method for developing neural networks: the case of evaluating commercial loan creditworthiness. Computers Ops Res, 23(10)(1996) pp. 933-944.

Gosling, J., 1995. Introductory Statistics: A comprehensive, self paced, step-by-step statistics course for tertiary students, Sydney: Pascal Press.

Hand, D. Mannila, H. and Smyth, P., 2001. Principles of Data Mining, Cambridge: MIT Press.

Jakimova, J. Minkiewicz, K., 2009, Euromoney Training EMEA, Credit Exposures and Credit Risk Modeling, Nairobi Kenya.

Joao, B., 2008. Munich Personal RePEc Archive. [online] Available at: < http://mpra.ub.uni-muenchen.de/8156/1/MPRA_paper_8156.pdf> [Accessed 11 February 2012].

Kirori, Z.K.,2011., 'A System for Credit Appraising-An Application of LogitBoost Algorithm', Msc thesis, University of Nairobi.

Lahsasna, A. Ainon, R.N. and Wah, T.Y., 2010. Credit Scoring Models Using Soft Computing Methods: A Survey. *The International Arab Journal of Information Technology*, 7(2) pp. 116-122.

Maclennan, J. Tang, Z.H. and Crivat, B., 2009. Introduction to Data Mining in SQL Server 2008, Indianapolis, Indiana: Wiley Publishing, Inc.

Mitchell, T.M., 1997. Machine Learning, London: McGraw Hill.

Montgomery, D. C. Runger, G.C., 2003. , Applied Statistics and Probability for Engineers. New York: John Wiley & Sons.

Mugambi, G.,2011., 'Rule Based Induction Credit Appraisal', Msc thesis, University of Nairobi.

Nisbet, R. Elder, J. and Miner, G., 2009. Handbook of statistical analysis and data mining applications, London: Academic Press.

Ong, M. K., 2006. The Basel Handbook: A guide to financial practitioners.2nd ed. New York: Risk Books.

Quinlan, J. R., 1986. Induction of decision trees. Machine Learning, 1 pp. 81-106.

Quittner, J., 2003. "Credit cards: subprime's tech dilemma: with delinquencies and charge-offs on the rise, the industry examines the role of automated decisioning". Bank Technology News, 16(1) pp. 19, 23.

Rajaraman, A. and Ullman, J.D., 2010. *Mining of Massive Datasets*, California: Stanford University.

Shmueli, G. Patel, N.R. and Bruce, P.C., 2002. Data Mining for Business Intelligence, New Jersey: John Wiley & Sons

Singh, Y. Bhatia, P.K. and Sangwan, O., A review of studies on machine learning techniques. International Journal of Computer Science and Security, 1(1) pp. 71-84.

Stead, A. G., MacDonald, K.G., 2002. *Constructing ROC Curves with the SAS System*. [online] Available at: <http://www2.sas.com/proceedings/sugi22/POSTERS/PAPER219.PDF> [Accessed 05 January 2012].

Swets, J.A. Dawes, R.M. and Monahan, J. 2000. Better Decisions through Science. Available at: < http://www.uchastings.edu/faculty-administration/faculty/park/class-website/docs/ScientificMethod-SwetsDawesMonahanROCcurves-s2012.pdf > [Accessed 03 March 2012].

Thomas L.C., 2000. A survey of credit and behavioral scoring; forecasting financial risk of lending to consumers. International Journal of Forecasting, 16 pp. 149–172.

Thomas, L. C. Oliver, R.W. and Hand, D.J., 2005. , A Survey of the Issues in Consumer Credit Modeling Research. The Journal of the Operational Research Society, 56(9) pp. 1006-1015.

Vitt, E. Luckevich, M. and Misner, S., 2002. *Business Intelligence: Making Better Decisions Faster*, Washington: Microsoft Press.

West, D., Neural network credit scoring models. *Computers & Operations Research*, 27 (2000) pp. 1131-1152.

Wiginton, J.C., 1980. A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis*, 15 pp. 757-770.

Yang, Y., 2005. *Adaptive Credit Scoring with Kernel Learning Methods*. [online] Available at: < http://www.crc.man.ed.ac.uk/conference/archive/2005/papers/yang.pdf > [Accessed 20 February 2012]>.

XPgroup., 1999. Extreme Programming: An agile software development resource [online] Available at: http://www.xprogramming.com/book/whatisxp/ > [Accessed 22 April 2012]>.

APPENDICES

Appendix One (Systems Documentation)

Installation requirements and Database Setup

- **System Installation**

The system is installed via a windows installer package (Logistica) that is provided. The installation is wizard guided and thus carried out as any standard windows based installation.

- **Database Management Server Installation**

The relational database management system (RDBMS) to be used is MS SQL SERVER 2008 (Express or nay higher versions). The free Express Edition is downloadable from the Microsoft download site.

- **Database set-up**

Once the RDBMS is installed, start the Microsoft SQL Server Management Studio then follow the steps below:-

Use the Database Engine to connect to the database server that was set-up

Right click on the databases node select on attach and then attach the database file provided (CreditScoring.mdf)

Add the observational data to any table in the database that is created from the above process.

- **R-installation**

The installation for the statistical software R is downloadable from the website Cran.r-project.org/bin/windows/base/old/2.6.2/. Once downloaded, the set-up is a wizard guided straight forward process. The preferred version of R is R 2.6.2. (R-2.6.2-win32).

- **RSrv250 Installation**

The RCOM server facilitates COM based connection to R Download this from any valid internet source and install it following the guidance from the wizard provided.

Appendix Two (User Documentation)

Once the system is all setup, the following procedure is used to create a model.

- **Authentication**

Start the program from the list of programs. The program has the name *Logistica.*

The simulated login is through the provision of a Username, Password and the optional selection of a model to be used.

- **Create a new Project**
  - ✓ Click on the settings menu
  - ✓ Fill in the model details in the crate project window that appears.
  - ✓ Log-out and re-login to the newly created project so as to create a model within the project.

- **Model creation Process**

The process of creating the model follows the modeling methodology (ETSRMA). The menus and thus sub-processes are:-

Explore

- ✓ From the list of columns provided, select the variable to explore
- ✓ Optionally set the analysis type
- ✓ In the statistic category, select the appropriate analysis to be done.
- ✓ Click on the analyze button

Transformation

- ✓ The transformation assumes fixed transformations (AsIs, Avg6, Avg3, AvgWoext etc) and both scalar and vector transformation.
- ✓ A sample script definition of transformations is provided with the sample database
- ✓ Not that the set of possible transformations is large and modeler creativity on the possible transformations of predictive power is key.

Flattening and Sampling

- ✓ The flattening and Sampling process of the data are a one step process done by clicking the Load Flattened Data and Load Sample buttons respectively and responding appropriately to the confirmation information messages.
- ✓ The Flattening/Sampling statistics are the provided and gives both counts and percentages out of the results of the flattening process.

Variable Re-Categorization

To Re-Categorize a variable, follow the process below:-

- ✓ Select the variable to Re-Categorize
- ✓ Select the number of groups to generate (for non-Categorical variables values)
- ✓ Click on the Generate Groups button to generate the desired number of groups
- ✓ Type in New Group number to re-assign new group numbers to the generated groups.
- ✓ Click on the Update categories button to save the new categories created.
- ✓ Once all variables are made categorical, Click on Generate Variables list to create the model project set of variables

Variable Ranking

Click on the variable ranking tab and/or the Load correlations button to view the Crammer's V ranking of the project variables.

Model Building and Assessment

To create a new model:-

- ✓ Click on the button New Model under the Model Building and Assessment Tab
- ✓ Type in a model name
- ✓ Select (by checking) the initial set of variable to include in the model
- ✓ Click on Save Model button to save the model
- ✓ Review on the model statistics and validations appropriately

To edit an existing model

- ✓ Select the model to edit
- ✓ Add or Remove (by checking/Un-checking) the variables to be included
- ✓ Click on save model to save the edited model
- ✓ Review the model statistics and validations

Generate Score Card

- ✓ Select the model for which a score card is to be generated
- ✓ Click on the Load button so as to load the selected models score-card

Group to Buckets

- ✓ Select the model to have its risk buckets generated
- ✓ Click on the Load Details button to generate or load the list of 50 ranked groups
- ✓ Click on the Graph button to view the graphical bucketing tool provided.
- ✓ Create maximum of 6 Buckets since the assumed maximum is 7
- ✓ The boundaries are ascertained by pointing along the desired border point along the graph.

Production/ Score

- ✓ Select the model and feed in period of data to be scored
- ✓ Click on the score button
- ✓ A list of scores is generated against the list of customers
- ✓ Click on the Graph button to appropriately view the graphical representation of the model score statistics