

**UNIVERSITY OF NAIROBI**

**SCHOOL OF COMPUTING AND INFORMATICS**

**A RULE INDUCTION BASED CREDIT APPRAISAL SYSTEM**

By **GRACE MUGAMBI**

**REGISTRATION NUMBER: P58/70479/08**

**Supervisor**

**DR. WANJIKU NG'ANG'A**

**A RESEARCH PROJECT PRESENTED IN PARTIAL FULFILMENT OF THE  
REQUIREMENTS GORVERNING THR AWARD OF THE DEGREE OF MASTERS OF  
SCIENCE IN COMPUTER SCIENCE**

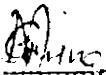
University of NAIROBI Library



0478802 2

## DECLARATION

This thesis is my original work and has not been presented for a degree or any other award in any other university or institution.

Signature.....

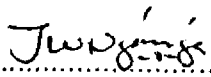
Date.....*30<sup>th</sup> May 2011*

Grace Mugambi

Registration Number: P58/70479/08

## DECLARATION BY SUPERVISOR

This thesis has been presented for examination with our approval as the university supervisors.

Signature.....

Date.....*06.06.2011*

Dr. Wanjiku Nganga

University of Nairobi, Kenya.

## ACKNOWLEDGEMENT

A research of this magnitude certainly requires much more than one person can offer. I am greatly indebted to many people that have walked with me through out this journey. It is not possible for me to make special mention of each of those that have so far assisted me in one way or the other since I set out on this task.

However, I wish to specially thank a few people for the pivotal role they have so generously played to make this work a success. I am particularly thankful to my supervisor, Dr Wanjiku Nganga for patiently leading me through the research process. Without your guidance I could not get far. I am also grateful to my lecturers at the School of Computing and Informatics of the University of Nairobi, for their support and fair criticism of this work.

Without a supportive family this effort would result to naught. I, therefore, wish to thank my entire family for their financial and psychological support through this process. This list would certainly be incomplete without a special thank you to my husband, Dr. Fred Mugambi and my daughter Angela for being so ever supportive of this and all of my other efforts. I am indeed humbled by your selflessness.

## ABSTRACT

Kenyan based institutions are facing a major problem today as they are handling a lot of data which they have collected over the decades. Kenya has in the past decade experienced growth in the industry sector. Most companies have piled up customer data especially in service-based industries. It is not possible for the human mind to be able to process this information and be able to make real-time decisions based on real-time reports. Often, the precious data is not even utilized to assist in decision making due to lack of the knowhow on handling the data.

Data mining is the process of extracting patterns from data and is becoming an increasingly important tool to transform this data into information.

This research aimed at exploring the various algorithms used in the data mining tools and use rule induction to build a model that can be used to appraise customers for a credit facility. The rules are built from data based on statistical significance calculated from the analysis of variable selection and significance. Weightings on each variable are obtained by use of statistical methods (Regression, Information value and Weights of evidence) and a lot of expertise advice.

The model has the ability to alter cell values in real time, in line with what the end user input or does on the dashboard. To be able to do a gauge so that the model gives a good or bad rating, a risk meter was designed and a minimum value was given. The risk meter uses the total weight output by the model after accepting a user's input to give a rating of bad or good.

The model was found to have an average accuracy level of close to 60% in its predictions of a good and bad credit borrower. It also gave insight into the data held that goes along way into helping the organization make decisions on areas that are business problems. For example, it was found that male borrowers have a higher defaulting rate than female have, yet less than twenty percent of borrowers were female.

For the bank to be able to predict the probable behavior of a customer's loan performance, they must have the history of the client's financial data. This may need an extension to borrow from the introduced Kenyan Banks Credit Bureau database so as to have a three-sixty degree view of a customer.

## TABLE OF CONTENTS

Declaration.....	2
Acknowledgement.....	3
Abstract.....	4
List of figures.....	7
List of Tables.....	7
<b>1.0 CHAPTER ONE: INTRODUCTION .....</b>	<b>8</b>
1.1 Overview .....	8
1.2 Outline of the report.....	9
1.3 Problem Statement .....	9
1.4 Objectives .....	11
1.5 Justification.....	12
<b>CHAPTER TWO: LITERATURE REVIEW.....</b>	<b>13</b>
2.1 Introduction.....	13
2.2 Background to Data Mining.....	13
2.3 The Foundations of Data Mining.....	14
2.5 The Scope of Data Mining.....	16
2.6 An Overview of Data Mining Techniques.....	17
2.6.1 Classical Techniques.....	17
2.6.2 Next Generation Techniques.....	21
2.7 How Data Mining Works.....	33
2.8 An Architecture for Data Mining.....	33
2.9 Knowledge Discovery.....	34
2.9.1 Pre-processing.....	35
2.9.2 Data warehouse Vs Mining.....	36
2.9.3 Results validation.....	36
2.9.4 Proposed solution to problem.....	36
2.9.5 Limitations of data mining.....	38
<b>CHAPTER THREE: RESEARCH METHODOLOGY.....</b>	<b>39</b>
3.1 Introduction.....	39
3.2 Data Mining Techniques used.....	40
3.3 Analysis and Design.....	42
3.4 Implementation.....	47
3.5 Presentation of Findings.....	50
<b>CHAPTER FOUR: RESULTS AND DISCUSSION.....</b>	<b>51</b>
4.1 Introduction.....	51

4.2 Data exploration and findings.....	52
4.3 Evaluation and Validation.....	57
4.4 Results of evaluation and discussion.....	59
<b>CHAPTER FIVE: CONCLUSIONS AND RECOMMENDATIONS.....</b>	<b>60</b>
5.1 Relevance of Study.....	60
5.2 Limitations of the model.....	61
5.3 Conclusion.....	62
5.4 Recommendations.....	62
<b>REFERENCES.....</b>	<b>63</b>

## LIST OF FIGURES

Figure 1: A decision tree is a predictive model making predictions on the basis of a series of Decision.....	22
Figure 2: Integrated Data Mining Architecture.....	34
Figure 3: The data Mining process.....	39
Figure 4: Outlier Detection.....	40
Figure 5: Typical phases of model development process.....	40
Figure 6: The regression function-Best possible line.....	41
Figure 7: Summary of the analysis, design and implementation.....	42
Figure 8: Diagram showing the final model.....	50

## LIST OF TABLES

Table 1: Steps in the Evolution of Data Mining.....	15
Table 2: Differences between the Nearest-Neighbor Data Mining Technique and Clustering...	19
Table 3: Statistical significance of variables.....	43
Table 4: Weights of Evidence and Information Value.....	46
Table 5: The weights behind the rules.....	48
Table 6: Example of a matrix based on a sample of 5.....	58

## CHAPTER ONE: INTRODUCTION

### 1.1 Overview

In most Kenyan companies, data mining is an alien word which is slowly but surely gaining familiarity in the Kenyan market. Most of these companies are experiencing an exponential growth in the data they are holding. An example is a local bank which in the period between 2000 and 2006, the customer base grew tremendously by an average of 55.3% per annum from 61,000 to 1,014,000 to make it the home of 29% of all bank account holders in Kenya. The number of staff supporting the higher business volumes also increased by 52.2% p.a. from 117 in 2000 to stand at 1394 in 2006.

One of the most challenging problems of the information society is dealing with the increasing data overload. Due to the digitalization of all sorts of content and due to the improvement and drop in cost of recording technologies, the amount of available information is enormous and is increasing... Today there's more information available to more people than ever before. But the flood of information can quickly overwhelm the human ability to process and analyze, particularly within a short time frame. That's why technology that assists with and speeds the processing, analysis, and delivery of information is in demand. Knowing that more and more companies are making information-based decisions, the thought of simply going with one's gut (particularly in a business environment as challenging as the one we're in today) should make business leaders a bit queasy. Luckily, it's no longer necessary.

Data mining is the process of extracting patterns from data. Data mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery.

This research was aimed at analyzing and comparing the different data mining algorithms available in the market, to make it easier for any company to be able to choose the best data mining algorithm that suits its business. It also aimed at solving the problem of qualifying the customers of a particular bank in Kenya for loan facilities by extracting patterns from the data they already have. It involved building a data mining model using a SAP Business agent tool and applying the induction rules, based on specific variables that will be selected. It will also look at the current processes being practiced to achieve knowledge extraction from data.

It is important to note that data mining can be used to uncover patterns in data but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data. Data mining cannot discover patterns that may be present in the larger body of data if those patterns are not present in the sample being "mined". Inability to find patterns may become a cause for some disputes between customers and service providers. Therefore data mining is not foolproof but may



be useful if sufficiently representative data samples are collected. The discovery of a particular pattern in a particular set of data does not necessarily mean that a pattern is found elsewhere in the larger data from which that sample was drawn. Hence an important part of the process is the verification and validation of patterns on other samples of data.

The related terms data dredging, data fishing and data snooping refer to the use of data mining techniques to sample sizes that are (or may be) too small for statistical inferences to be made about the validity of any patterns discovered. Data dredging may, however, be used to develop new hypotheses, which must then be validated with sufficiently large sample sets.

## **1.2 Outline of the report**

The report begins with an introduction to data mining. It also gives the problem statement, objectives and justification of the study in the first chapter. Chapter two will be the literature review, where various data mining algorithms will be analyzed and also highlights the proposed solution to the problem posed in the previous chapter. Chapter three presents the methodology that was used in the research to build the model. Chapter four gives the results and findings of the analysis done, while a conclusion and recommendations are highlighted in chapter five.

## **1.3 Problem Statement**

The technologies for generating and collecting data have been advancing rapidly. At the current stage, lack of data is no longer a problem; the inability to generate useful information from data is! The explosive growth in data and database results in the need to develop new technologies and tools to process data into useful information and knowledge intelligently and automatically. Data mining, therefore, has become a research area with increasing importance Weiss and Indurkha, (1998).

Business databases have grown tremendously in recent years, but the capabilities for analyzing such large amounts of data have not developed at the same rate as the capabilities of collecting and storing data. As a result, businesses are becoming increasingly concerned with the study of knowledge discovery methods.

Terabytes of data are generated everyday in many organizations. To extract hidden predictive information from large volumes of data, data mining techniques are needed. Organizations are starting to realize the importance of data mining in their strategic planning and successful application of data mining techniques can be an enormous payoff for the organizations. This paper discusses the current challenges faced by local banks in trying to mine data, and describes major data mining techniques.

Kenyan based institutions are facing a major problem today as they are handling a lot of data which they have collected over the decades. It is not possible for the human mind to be able to process this information

and be able to make real-time decisions based on real-time reports. Often, the precious data is not even utilized to assist in decision making due to lack of the know how on handling the data. What is even worse is the little knowledge that most organizations have on data mining tools based on their requirements, so that those that endeavor in trying to obtain one, either get stalled in the project, or end up choosing a tool that may not have been the most suitable, based on what kind of data they have.

This research aimed at exploring the various algorithms used in the data mining tools and pick one of it to demonstrate how it can be used to make decision making easier by making good use of the data a company already has and ultimately save not only on cost but also on time spent to carry out day to day activities.

Most companies are currently making decisions not based on facts or patterns from the bulk of data they have, but based on hearsay, and 'strong feelings' of the leaders. For instance, when a new product is to be introduced to the market, it is normally on a test and see the waters basis, instead of using the data collected over the years.

Rather than randomly contacting a prospect or customer through a call center or sending mail, a company can concentrate its efforts on prospects that are predicted to have a high likelihood of responding to an offer.

Catalogers have a rich history of customer transactions on millions of customers dating back several years. Data mining tools can identify patterns among customers and help identify the most likely customers to respond to upcoming mailing campaigns. More sophisticated methods may be used to optimize resources across campaigns so that one may predict which channel and which offer an individual is most likely to respond to — across all potential offers. This helps in maximizing the effort needed.

Data clustering can also be used to automatically discover the segments or groups within a customer data set. For example, in cases where many people will take an action without an offer, uplift modeling can be used to determine which people will have the greatest increase in responding if given an offer.

Businesses employing data mining may see a return on investment, but also they recognize that the number of predictive models can quickly become very large. In order to maintain this quantity of models, they need to manage model versions and move to automated data mining. Rather than one model to predict which customers will churn, a business could build a separate model for each region and customer type. Then instead of sending an offer to all people that are likely to churn, it may only want to send offers to customers that will likely take the offer. And finally, it may also want to determine which customers are

going to be profitable over a window of time and only send the offers to those that are likely to be profitable.

HR departments are not able to identify the characteristics of their most successful employees especially when their numbers are growing exponentially. Information obtained, such as universities attended by highly successful employees, can help HR focus recruiting efforts accordingly. Additionally, Strategic Enterprise Management applications help a company translate corporate-level goals, such as profit and margin share targets, into operational decisions, such as production plans and workforce levels.

Retail sellers can use their purchases data to identify customer preferences of one type of item over another through market basket analysis. If a clothing store records the purchases of customers, a data-mining system could identify those customers who favor silk shirts over cotton ones. Although some explanations of relationships may be difficult, taking advantage of it is easier. Not all data are transaction based and logical or inexact rules may also be present within a database. In a manufacturing application, an inexact rule may state that 73% of products which have a specific defect or problem will develop a secondary problem within the next six months.

Companies also need to discover their loyal and evangelistic customers. Market basket analysis has also been used to identify the purchase patterns of the Alpha consumer. Alpha Consumers are people that play a key role in connecting with the concept behind a product, then adopting that product, and finally validating it for the rest of society. Analyzing the data collected on these types of users has allowed companies to predict future buying trends and forecast supply demands.

#### **1.4 Objectives**

This study's main objective was to design and develop a credit rating system using rule induction as the underlying algorithm.

Specific objectives are:

- 1 To identify and analyze the factors that should be considered as most significant while appraising a customer for a credit facility.
2. To identify the weightings that each of the variables should be awarded and in what ratios they relate.
3. To use rule induction to build a model that demonstrates how easily data can be mined to give information.

## 1.5 Justification

Kenya has in the past decade experienced growth in the industry sector. Most companies have piled up customer data especially in service-based industries. But as technology is just setting foot in most African based companies, making information-based decisions, was not wanting until recently. Luckily, it's no longer necessary to head a great corporation based on guts only.

Data mining will help front-line managers to improve financial performance with complete, up-to-the-minute information on their departments. Hundreds of key performance indicators and a lot of reports to be produced enable financial managers to improve cash flow, lower costs, and increase profitability while maintaining more accurate, timely, and transparent financial reporting. This in turn will maximize shareholder wealth by capitalizing on strengths, harnessing opportunities, reducing weaknesses and mitigating threats. This should enable the enterprise to be; more agile, better manage costs and performance, and attract new customers e.g. by concentrating efforts on prospects that are predicted to have a high likelihood of responding to a new product;

Unfortunately, few corporations have ventured into looking for a data mining tool. The process of knowledge extraction is still very manual and painfully slow, hence wrong decisions may be made due to lack of proper knowledge, research, or hurry to conclude on the costly project. It is the purpose of this proposal to find a summary of what most of these companies may be looking for. This research proposes to do an in-depth analysis of the role played by data mining in the growth and profitability of companies which will help local decision makers appreciate it and embrace it more.

Due to hesitance in spending to invest on a not very well known technology, some corporations have taken time to acquire a data mining tool as they wait to benefit from the few successful case studies in the region. By having a good analytical summary of the capabilities and low-sides of each data mining algorithm, a decision of the most suitable tool can be reached faster and more smoothly and may be in half the time that it currently takes to implement such projects. This in turn, may encourage more institutions in the region to take the step and acquire a data mining tool, hence taking their business to the next frontier.

## CHAPTER TWO: LITERATURE REVIEW

### 2.1 Introduction

This chapter presents a summary of scholarly work on various key areas surrounding data mining. The chapter opens with an overview of the topic being studied. The chapter, also, presents relevant literature on its requirements and scope.

### 2.2 Background to Data mining

Aristotle began one of his most well known works, *The Metaphysics*, with the words "By nature, all men long to know." Humans have been "manually" extracting patterns from data for centuries, but the increasing volume of data in modern times has called for more automated approaches. Early methods of identifying patterns in data include Bayes' theorem (1700s) and regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has increased data collection and storage. As data sets have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. This has been aided by other discoveries in computer science, such as neural networks, clustering, genetic algorithms (1950s), decision trees (1960s) and support vector machines (1980s). Data mining is the process of applying these methods to data with the intention of uncovering hidden patterns. It has been used for many years by businesses, scientists and governments to sift through volumes of data such as airline passenger trip records, census data and supermarket scanner data to produce market research reports. (Note, however, that reporting is not always considered to be data mining.) As more data are gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform these data into information. It is commonly used in a wide range of profiling practices such as marketing, surveillance, fraud detection and scientific discovery.

Data mining is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. In the public sector, data mining applications initially were used as a means to detect fraud and waste, but have grown to also be used for purposes such as measuring and improving program performance.

A primary reason for using data mining is to assist in the analysis of collections of observations of behavior. Such data are vulnerable to co linearity because of unknown interrelations. An unavoidable fact of data mining is that the (sub-) set(s) of data being analyzed may not be representative of the whole domain, and therefore may not contain examples of certain critical relationships and behaviors that exist across other parts of the domain. To address this sort of issue, the analysis may be augmented using experiment-based and other approaches, such as Choice Modeling for human-generated data. In these

situations, inherent correlations can be either controlled for, or removed altogether, during the construction of the experimental design.

The key to successful applications of data mining as a Business Intelligence tool is collaboration and knowledge sharing among frontline users and technology experts in the organization (Foley, 2001; Reingruber and Knodson, 2008). Business Intelligence is a broad category of applications and technologies of gathering, accessing, and analyzing a large amount of data for the organization to make effective business decisions (Cook and Cook, 2000; Williams and Williams, 2006). Typical Business Intelligence technologies include business rule modeling, data profiling, data warehousing and online analytical processing, and Data Mining (Loshin, 2003). The central theme of Business Intelligence is to fully utilize massive data to help organizations gain competitive advantages. Knowledge Management, on the other hand, is a set of practices of the creation, development, and application of knowledge to enhance performance of the organization (Wiig, 1999; Buckman, 2004; Feng and Chen, 2007; Lee and Change, 2007; Smoliar, 2007; Wu et al., 2007; Paiva and Goncalo, 2008; Ramachandran et al., 2008). Similar to Business Intelligence, Knowledge Management improves the use of information and knowledge available to the organization (Sun and Chen, 2008). However, Knowledge Management is distinct from Business Intelligence in many aspects. Generally, Knowledge Management is concerned with human subjective knowledge, not data or objective information (Davenport and Seely, 2006).

Owing to its strength, Data Mining is known as a powerful Business Intelligence tool for knowledge discovery (Chen and Liu, 2005). The process of Data Mining is a Knowledge Management process because it involves human knowledge (Brachman *et al.*, 1996).

### **2.3 The Foundations of Data Mining**

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Data mining algorithms embody techniques that have existed for at least 20 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods.

In the evolution from business data to business information, each new step has built upon the previous one. For example, dynamic data access is critical for drill-through in data navigation applications, and the ability to store large databases is critical to data mining. From the user's point of view, the four steps listed in Table 1 were revolutionary because they allowed new business questions to be answered accurately and quickly.

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

Table 1: Steps in the Evolution of Data Mining (Thearling n. d.).

The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity

of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments.

## 2.5 The Scope of Data Mining

Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- **Automated prediction of trends and behaviors.** Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.
- **Automated discovery of previously unknown patterns.** Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

According to Thearling, databases can be larger in both depth and breadth:

- **More columns.** Analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. High performance data mining allows users to explore the full depth of a database, without preselecting a subset of variables.
- **More rows.** Larger samples yield lower estimation errors and variance, and allow users to make inferences about small but important segments of a population.



## 2.6 An Overview of Data Mining Techniques

This overview provides a description of some of the most common data mining algorithms in use today broken down into two themes:

- Classical Techniques: Statistics, Neighborhoods and Clustering
- Next Generation Techniques: Trees, Networks and Rules

These two sections have been broken up based on when the data mining technique was developed and when it became technically mature enough to be used for business, especially for aiding in the optimization of customer relationship management systems. This section should help readers to understand the rough differences in the techniques and at least enough information to be dangerous and well armed enough to not be baffled by the vendors of different data mining tools.

The main techniques that are discussed here are the ones that are used 99.9% of the time on existing business problems. There are certainly many other ones as well as proprietary techniques from particular vendors - but in general the industry is converging to those techniques that work consistently and are understandable and explainable.

### 2.6.1 Classical Techniques :

#### i) Statistical Techniques

By strict definition "statistics" or statistical techniques are not data mining. They were being used long before the term data mining was coined to apply to business applications. However, statistical techniques are driven by the data and are used to discover patterns and build predictive models. And from the users perspective you will be faced with a conscious choice when solving a "data mining" problem as to whether you wish to attack it with statistical methods or other data mining techniques. For this reason it is important to have some idea of how statistical techniques work and how they can be applied (Berson, Smith & Thearling 2005).

According to Berson, Smith and Thearling, the difference between statistics and data mining is classical data mining techniques such as CART, neural networks and nearest neighbor techniques tend to be more robust to both messier real world data and also more robust to being used by less expert users. Also, the amount of data stored in this era surpasses what statistical techniques can process.

Hence, some statistical techniques of analyzing data include

- ❖ **Regression** - Attempts to find a function which models the data with the least error. Regression is the oldest and most well-known statistical technique that the data mining community utilizes. Basically, regression takes a numerical dataset and develops a mathematical formula that fits the data. When you're ready to use the results to predict future behavior, you simply take your new data, plug it into the developed formula and you've got a prediction! The major limitation of this technique is that it only works well with continuous quantitative data (like weight, speed or age). If you're working with categorical data where order is not significant (like color, name or gender) you're better off choosing another technique.

- ❖ **Histogram representation of data.**

### ii) Nearest neighbor method

A technique that classifies each record in a dataset based on a combination of the classes of the  $k$  record(s) most similar to it in a historical dataset (where  $k \geq 1$ ), sometimes called the  $k$ -nearest neighbor technique. Clustering and the Nearest Neighbor prediction technique are among the oldest techniques used in data mining. The nearest neighbor prediction algorithm simply stated is: "Objects that are "near" to each other will have similar prediction values as well". Thus if you know the prediction value of one of the objects you can predict it for its nearest neighbors.

### iii) Clustering

Clustering is the method by which like records are grouped together. It's like classification but the groups are not predefined, so the algorithm will try to group similar items together. Usually this is done to give the end user a high level view of what is going on in the database. Clustering is sometimes used to mean segmentation - which most marketing people will tell you is useful for coming up with a bird's eye view of the business. Sometimes clustering is performed not so much to keep records together as to make it easier to see when one record sticks out from the rest- that is clustering for outliers.

The nearest neighbor algorithm is basically a refinement of clustering in the sense that they both use distance in some feature space to create either structure in the data or predictions. The nearest neighbor algorithm is a refinement since part of the algorithm usually is a way of automatically determining the weighting of the importance of the predictors and how the distance will be measured within the feature space. Clustering is one special case of this where the importance of each predictor is considered to be equivalent.

The main distinction between clustering and the nearest neighbor technique is that clustering is what is called an unsupervised learning technique and nearest neighbor is generally used for prediction or a supervised learning technique. Unsupervised learning techniques are unsupervised in the sense that when

they are run there is no particular reason for the creation of the models the way there is for supervised learning techniques that are trying to perform prediction. In prediction, the patterns that are found in the database and presented in the model are always the most important patterns in the database for performing some particular prediction. In clustering there is no particular sense of why certain records are near to each other or why they all fall into the same cluster.

Some of the differences between clustering and nearest neighbor prediction can be summarized in Table 2 (Berson, Smith & Thearling 2005).

Nearest Neighbor	Clustering
Used for prediction as well as consolidation.	Used mostly for consolidating data into a high-level view and general grouping of records into like behaviors.
Space is defined by the problem to be solved (supervised learning).	Space is defined as default n-dimensional space, or is defined by the user, or is a predefined space driven by past experience (unsupervised learning).
Generally only uses distance metrics to determine nearness.	Can use other metrics besides distance to determine nearness of two records - for example linking two points together.

Table 2: Some of the Differences between the Nearest-Neighbor Data Mining Technique and Clustering

When clustering much larger numbers of records tradeoffs are explicitly defined by the clustering algorithm. Creating homogeneous clusters where all values for the predictors are the same is difficult to do when there are many predictors and/or the predictors have many different values (high cardinality). It is possible to guarantee that homogeneous clusters are created by breaking apart any cluster that is inhomogeneous into smaller clusters that are homogeneous. In the extreme, though, this usually means creating clusters with only one record in them which usually defeats the original purpose of the clustering. The second important constraint on clustering is then that a reasonable number of clusters are formed. Where, again, reasonable is defined by the user but is difficult to quantify beyond that except to say that just one cluster is unacceptable (too much generalization) and that as many clusters as original records is also unacceptable. Many clustering algorithms either let the user choose the number of clusters that they would like to see created from the database or they provide the user a "knob" by which they can create fewer or greater numbers of clusters interactively after the clustering has been performed (Berson, Smith & Thearling 2005, p15).

For clustering the n-dimensional space is usually defined by assigning one predictor to each dimension. For the nearest neighbor algorithm predictors are also mapped to dimensions but then those dimensions are literally stretched or compressed based on how important the particular predictor is in making the prediction. The stretching of a dimension effectively makes that dimension (and hence predictor) more important than the others in calculating the distance.

There are two main types of clustering techniques, those that create a hierarchy of clusters and those that do not. The hierarchical clustering techniques create a hierarchy of clusters from small to big. The main reason for this is that, as was already stated, clustering is an unsupervised learning technique, and as such, there is no absolutely correct answer. For this reason and depending on the particular application of the clustering, fewer or greater numbers of clusters may be desired. With a hierarchy of clusters defined it is possible to choose the number of clusters that are desired.

This hierarchy of clusters is created through the algorithm that builds the clusters. Berson, Smith and Thearling suggest that there are two main types of hierarchical clustering algorithms:

- ❖ **Agglomerative** - Agglomerative clustering techniques start with as many clusters as there are records where each cluster contains just one record. The clusters that are nearest each other are merged together to form the next largest cluster. This merging is continued until a hierarchy of clusters is built with just a single cluster containing all the records at the top of the hierarchy.
- ❖ **Divisive** - Divisive clustering techniques take the opposite approach from agglomerative techniques. These techniques start with all the records in one cluster and then try to split that cluster into smaller pieces and then in turn to try to split those smaller pieces.

Of the two the agglomerative techniques are the most commonly used for clustering and have more algorithms developed for them. The non-hierarchical techniques in general are faster to create from the historical database but require that the user make some decision about the number of clusters desired or the minimum “nearness” required for two records to be within the same cluster. These non-hierarchical techniques are often run multiple times starting off with some arbitrary or even random clustering and then iteratively improving the clustering by shuffling some records around. Or these techniques sometimes create clusters that are created with only one pass through the database adding records to existing clusters when they exist and creating new clusters when no existing cluster is a good candidate for the given record. Because the definition of which clusters are formed can depend on these initial choices of which starting clusters should be chosen or even how many clusters these techniques can be less repeatable than the hierarchical techniques and can sometimes create either too many or too few clusters because the number of clusters is predetermined by the user not determined solely by the patterns inherent in the database.

There are two main non-hierarchical clustering techniques. Both of them are very fast to compute on the database but have some drawbacks. The first are the single pass methods. They derive their name from the fact that the database must only be passed through once in order to create the clusters (i.e. each record is only read from the database once). The other classes of techniques are called reallocation methods. They get their name from the movement or “reallocation” of records from one cluster to another in order to create better clusters. The reallocation techniques do use multiple passes through the database but are relatively fast in comparison to the hierarchical techniques (Berson, Smith & Thearling 2005).

Some techniques allow the user to request the number of clusters that they would like to be pulled out of the data. Predefining the number of clusters rather than having them driven by the data might seem to be a bad idea as there might be some very distinct and observable clustering of the data into a certain number of clusters which the user might not be aware of.

Hierarchical clustering has the advantage over non-hierarchical techniques in that the clusters are defined solely by the data (not by the users predetermining the number of clusters) and that the number of clusters can be increased or decreased by simple moving up and down the hierarchy.

The hierarchy is created by starting either at the top (one cluster that includes all records) and subdividing (divisive clustering) or by starting at the bottom with as many clusters as there are records and merging (agglomerative clustering). Usually the merging and subdividing are done two clusters at a time (Berson, Smith & Thearling 2005).

The main distinction between the techniques is their ability to favor long, scraggly clusters that are linked together record by record, or to favor the detection of the more classical, compact or spherical cluster. It may seem strange to want to form this long snaking chain like clusters, but in some cases they are the patterns that the user would like to have detected in the database. These are the times when the underlying space looks quite different from the spherical clusters and the clusters that should be formed are not based on the distance from the center of the cluster but instead based on the records being “linked” together.

### **2.6.2 Next Generation Techniques**

The data mining techniques in this section represent the most often used techniques that have been developed over the last two decades of research. They also represent the vast majority of the techniques that are being spoken about when data mining is mentioned in the popular press. These techniques can be used for either discovering new information within large databases or for building predictive models (Berson, Smith & Thearling 2005).

## i) Decision trees

These are tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART- Segments a dataset by creating 2-way splits. Requires less data preparation than CHAID) and Chi Square Automatic Interaction Detection (CHAID-Segments a dataset by using chi square tests to create multi-way splits). Though the older decision tree techniques such as CHAID are currently highly used the new techniques such as CART are gaining wider acceptance.

A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. For instance if we were going to classify customers who churn (say on their postpaid phone contracts) in the Mobile Telephone Industry a decision tree might look something like that found in Figure 1.

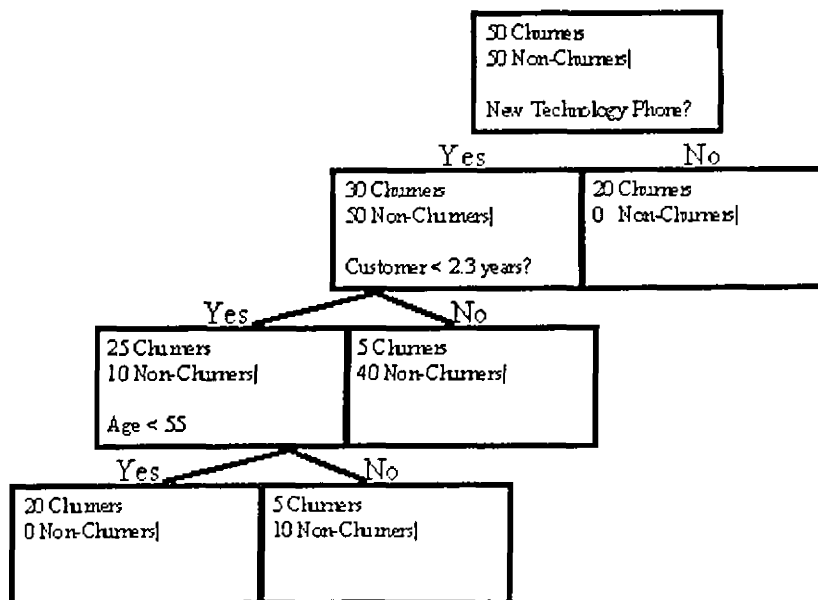


Figure 1 A decision tree is a predictive model that makes a prediction on the basis of a series of decision much like the game of 20 questions (Berson, Smith & Thearling 2005).

From a business perspective decision trees can be viewed as creating a segmentation of the original dataset (each segment would be one of the leaves of the tree). Segmentation of customers, products, and sales regions is something that marketing managers have been doing for many years. In the past this segmentation has been performed in order to get a high level view of a large amount of data - with no

particular reason for creating the segmentation except that the records within each segmentation were somewhat similar to each other. In this case the segmentation is done for a particular reason - namely for the prediction of some important piece of information. The records that fall within each segment fall there because they have similarity with respect to the information being predicted - not just that they are similar - without similarity being well defined. These predictive segments that are derived from the decision tree also come with a description of the characteristics that define the predictive segment. Thus the decision trees and the algorithms that create them may be complex; the results can be presented in an easy to understand way that can be quite useful to the business user (Berson, Smith & Thearling 2005).

Because of their tree structure and ability to easily generate rules decision trees are the favored technique for building understandable models. Because of this clarity they also allow for more complex profit and ROI models to be added easily in on top of the predictive model. For instance once a customer population is found with high predicted likelihood to attrite a variety of cost models can be used to see if an expensive marketing intervention should be used because the customers are highly valuable or a less expensive intervention should be used because the revenue from this sub-population of customers is marginal (Berson, Smith & Thearling 2005).

Due to their high level of automation and the ease of translating decision tree models into SQL for deployment in relational databases the technology has also proven to be easy to integrate with existing IT processes, requiring little preprocessing and cleansing of the data, or extraction of a special purpose file specifically for data mining.

Decision trees technology was originally developed for statisticians to automate the process of determining which fields in their database were actually useful or correlated with the particular problem that they were trying to understand. Partially because of this history, decision tree algorithms tend to automate the entire process of hypothesis generation and then validation much more completely and in a much more integrated way than any other data mining techniques. They are also particularly adept at handling raw data with little or no pre-processing. Perhaps also because they were originally developed to mimic the way an analyst interactively performs data mining they provide a simple to understand predictive model based on rules (Berson, Smith & Thearling 2005).

Because decision trees score so highly on so many of the critical features of data mining they can be used in a wide variety of business problems for both exploration and for prediction. They have been used for problems ranging from credit card attrition prediction to time series prediction of the exchange rate of different international currencies. There are also some problems where decision trees will not do as well. Some very simple problems where the prediction is just a simple multiple of the predictor can be solved much more quickly and easily by linear regression. Usually the models to be built and the interactions to

be detected are much more complex in real world problems and this is where decision trees excel (Berson, Smith, & Thearling 2005).

Berson, Smith and Thearling suggest that the first step in the process is that of growing the tree. Specifically the algorithm seeks to create a tree that works as perfectly as possible on all the data that is available, most of the time it is not possible to have the algorithm working perfectly. There is always noise in the database to some degree (there are variables that are not being collected that have an impact on the target you are trying to predict). The name of the game in growing the tree is in finding the best possible question to ask at each branch point of the tree. The process in decision tree algorithms is very similar when they build trees. These algorithms look at all possible distinguishing questions that could possibly break up the original training dataset into segments that are nearly homogeneous with respect to the different classes being predicted. Some decision tree algorithms may use heuristics in order to pick the questions or even pick them at random. CART picks the questions in a much unsophisticated way: It tries them all. After it has tried them all CART picks the best one uses it to split the data into two more organized segments and then again asks all possible questions on each of those new segments individually. If the decision tree algorithm just continued growing the tree like this it could conceivably create more and more questions and branches in the tree so that eventually there was only one record in the segment. To let the tree grow to this size is both computationally expensive but also unnecessary. Most decision tree algorithms stop growing the tree when one of three criteria is met:

The segment contains only one record. (There is no further question that you could ask which could further refine a segment of just one.)

All the records in the segment have identical characteristics. (There is no reason to continue asking further questions segmentation since all the remaining records are the same.)

The improvement is not substantial enough to warrant making the split.

After the tree has been grown to a certain size (depending on the particular stopping criteria used in the algorithm) the CART algorithm has still more work to do. One of the great advantages of CART is that the algorithm has the validation of the model and the discovery of the optimally general model built deeply into the algorithm. CART accomplishes this by building a very complex tree and then pruning it back to the optimally general tree based on the results of cross validation or test set validation. The tree is pruned back based on the performance of the various pruned version of the tree on the test set data. The most complex tree rarely fares the best on the held aside data as it has been over fitted to the training data. By using cross validation the tree that is most likely to do well on new, unseen data can be chosen (Berson, Smith & Thearling 2005).



Another equally popular decision tree technology to CART is CHAID or Chi-Square Automatic Interaction Detector. CHAID is similar to CART in that it builds a decision tree but it differs in the way that it chooses its splits. Instead of the entropy or Gini metrics for choosing optimal splits the technique relies on the chi square test used in contingency tables to determine which categorical predictor is furthest from independence with the prediction values (Berson, Smith & Thearling 2005).

Because CHAID relies on the contingency tables to form its test of significance for each predictor all predictors must either be categorical or be coerced into a categorical form via binning. Though this binning can have deleterious consequences the actual accuracy performances of CART and CHAID have been shown to be comparable in real world direct marketing response models (Berson, Smith & Thearling 2005).

## ii) Neural Networks

Neural networks are biological systems (a k a brains) that detect patterns, make predictions and learn. According to Berson, Smith and Thearling, the artificial ones are computer programs implementing sophisticated pattern detection and machine learning algorithms on a computer to build predictive models from large historical databases. Artificial neural networks derive their name from their historical development which started off with the premise that machines could be made to “think” if scientists found ways to mimic the structure and functioning of the human brain on the computer. Thus historically neural networks grew out of the community of Artificial Intelligence rather than from the discipline of statistics. More mundane applications of artificial neural networks, to the real world problems include customer response prediction or fraud detection rather than the loftier goals that were originally set out for the techniques such as overall human learning and computer speech and image understanding.

They also suggest that there are many important design decisions that need to be made in order to effectively use a neural network such as:

- i) How should the nodes in the network be connected?
- ii) How many neurons like processing units should be used?
- iii) When should “training” be stopped in order to avoid over fitting?

There are also many important steps required for preprocessing the data that goes into a neural network - most often there is a requirement to normalize numeric data between 0.0 and 1.0 and categorical predictors may need to be broken up into virtual predictors that are 0 or 1 for each value of the original categorical predictor.

Neural networks are very powerful predictive modeling techniques but some of the power comes at the expense of ease of use and ease of deployment. The model created by neural networks is represented by numeric values in a complex calculation that requires all of the predictor values to be in the form of a

number. The output of the neural network is also numeric and needs to be translated if the actual prediction value is categorical. Packaging up neural networks with expert consultants is a viable strategy that avoids many of the pitfalls of using neural networks, but it can be quite expensive because it is human intensive. One of the great promises of data mining is, after all, the automation of the predictive modeling process. These neural network consulting teams are little different from the analytical departments many companies already have in house. Since there is not a great difference in the overall predictive accuracy of neural networks over standard statistical techniques the main difference becomes the replacement of the statistical expert with the neural network expert. Either with statistics or neural network experts the value of putting easy to use tools into the hands of the business end user is still not achieved (Berson, Smith & Thearling 2005).

Neural networks are used in a wide variety of applications. They have been used in all facets of business from detecting the fraudulent use of credit cards and credit risk prediction to increasing the hit rate of targeted mailings. They also have a long history of application in other areas such as the military for the automated driving of an unmanned vehicle at 30 miles per hour on paved roads to biological simulations such as learning the correct pronunciation of English words from written text. Neural networks of various kinds can be used for clustering and prototype creation. The Kohonen network described in this section is probably the most common network used for clustering and segmentation of the database. Typically the networks are used in an unsupervised learning mode to create the clusters. The clusters are created by forcing the system to compress the data by creating prototypes or by algorithms that steer the system toward creating clusters that compete against each other for the records that they contain, thus ensuring that the clusters overlap as little as possible. Sometimes clustering is performed not so much to keep records together as to make it easier to see when one record sticks out from the rest. That is, for outlier analysis.

A neural network is loosely based on how some people believe that the human brain is organized and how it learns. Berson, Smith and Thearling give two main structures of consequence in the neural network:

- I) **The node** - This loosely corresponds to the neuron in the human brain.
- II) **The link** - which loosely corresponds to the connections between neurons (axons, dendrites and synapses) in the human brain

In order to make a prediction, they suggest that the neural network accepts the values for the predictors on what are called the input nodes. These become the values for those nodes those values are then multiplied by values that are stored in the links (sometimes called links and in some ways similar to the weights that were applied to predictors in the nearest neighbor method). These values are then added together at the node at the far right (the output node) a special threshold function is applied and the resulting number is the prediction. In this case if the resulting number is 0 the record is considered to be a good credit risk (no default) if the number is 1 the record is considered to be a bad credit risk (likely default). The neural

network model is created by presenting it with many examples of the predictor values from records in the training set and the prediction value from those same records. By comparing the correct answer obtained from the training record and the predicted answer from the neural network it is possible to slowly change the behavior of the neural network by changing the values of the link weights. It is the weights of the links that actually control the prediction value for a given record. Thus the particular model that is being found by the neural network is in fact fully determined by the weights and the architectural structure of the network. For this reason it is the link weights that are modified each time an error is made. The meaning of the input nodes and the output nodes are usually pretty well understood - and are usually defined by the end user based on the particular problem to be solved and the nature and structure of the database. The hidden nodes, however, do not have a predefined meaning and are determined by the neural network as it trains. This poses two problems:

It is difficult to trust the prediction of the neural network if the meaning of these nodes is not well understood.

Since the prediction is made at the output layer and the difference between the prediction and the actual value is calculated there, how is this error correction fed back through the hidden layers to modify the link weights that connect them (Berson, Smith & Thearling 2005)?

The learning procedure for the neural network has been defined to work for the weights in the links connecting the hidden layer. A good metaphor for how this works is to think of a military operation in some war where there are many layers of command with a general ultimately responsible for making the decisions on where to advance and where to retreat. The general probably has several lieutenant generals advising him and each lieutenant general probably has several major generals advising him. This hierarchy continues downward through colonels and privates at the bottom of the hierarchy (Berson, Smith & Thearling 2005).

This is not too far from the structure of a neural network with several hidden layers and one output node. You can think of the inputs coming from the hidden nodes as advice. The link weight corresponds to the trust that the general has in his advisors and the actual node value represents how strong an opinion this particular advisor has about this particular situation. To make a decision the general considers how trustworthy and valuable the advice is and how knowledgeable and confident each advisor is in making their suggestion and then taking all of this into account the general makes the decision to advance or retreat. In the same way the output node will make a decision (a prediction) by taking into account all of the input from its advisors (the nodes connected to it). In the case of the neural network this decision is reached by multiplying the link weight by the output value of the node and summing these values across all nodes. If the prediction is incorrect the nodes that had the most influence on making the decision have their weights.

### iii) Rule induction

This is the extraction of useful if-then rules from data based on statistical significance. Rule induction is one of the major forms of data mining and is perhaps the most common form of knowledge discovery in unsupervised learning systems. Rule induction on a data base can be a massive undertaking where all possible patterns are systematically pulled out of the data and then an accuracy and significance are added to them that tell the user how strong the pattern is and how likely it is to occur again. In general these rules are relatively simple. The rules that are pulled from the database are extracted and ordered to be presented to the user based on the percentage of times that they are correct and how often they apply.

The bane of rule induction systems is also its strength - that it retrieves all possible interesting patterns in the database. This is a strength in the sense that it leaves no stone unturned but it can also be viewed as a weakness because the user can easily become overwhelmed with such a large number of rules that it is difficult to look through all of them. The overabundance of patterns can also be problematic for the simple task of prediction because all possible patterns are culled from the database there may be conflicting predictions made by equally interesting rules. Automating the process of culling the most interesting rules and of combing the recommendations of a variety of rules as well handled by many of the commercially available rule induction systems on the market today and is also an area of active research (Berson, Smith & Thearling 2005).

Rule induction systems are highly automated and are among the best of data mining techniques for exposing all possible predictive patterns in a database. They can be modified to for use in prediction problems but the algorithms for combining evidence from a variety of rules comes more from rules of thumbs and practical experience. In comparing data mining techniques along an axis of explanation neural networks would be at one extreme of the data mining algorithms and rule induction systems at the other end. Neural networks are extremely proficient and saying exactly what must be done in a prediction task with little explanation. Rule induction systems when used for prediction on the other hand are like having a committee of trusted advisors each with a slightly different opinion as to what to do but relatively well grounded reasoning and a good explanation for why it should be done (Berson, Smith & Thearling 2005).

According to Berson, Smith and Thearling, in rule induction systems the rule itself is of a simple form of "if this and this and this then this". In order for the rules to be useful there are two pieces of information that must be supplied as well as the actual rule:

- I) **Accuracy** - How often is the rule correct?
- II) **Coverage** - How often does the rule apply?

Just because the pattern in the data base is expressed as rule does not mean that it is true all the time. Thus just like in other data mining algorithms it is important to recognize and make explicit the uncertainty in the

rule. This is what the accuracy of the rule means. The coverage of the rule has to do with how much of the database the rule “covers” or applies to. In some cases accuracy is called the confidence of the rule and coverage is called the support. The rules themselves consist of two halves. The left hand side is called the antecedent and the right hand side is called the consequent. The antecedent can consist of just one condition or multiple conditions which must all be true in order for the consequent to be true at the given accuracy. Generally the consequent is just a single condition rather than multiple conditions (Berson, Smith & Thearling 2005).

When the rules are mined out of the database the rules can be used either for understanding better the business problems that the data reflects or for performing actual predictions against some predefined prediction target. Since there is both a left side and a right side to a rule (antecedent and consequent) they can be used in several ways for your business.

**Target the antecedent.** In this case all rules that have a certain value for the antecedent are gathered and displayed to the user.

**Target the consequent.** In this case all rules that have a certain value for the consequent can be used to understand what is associated with the consequent and perhaps what affects the consequent.

**Target based on accuracy.** Some times the most important thing for a user is the accuracy of the rules that are being generated. Highly accurate rules of 80% or 90% imply strong relationships that can be exploited even if they have low coverage of the database and only occurring a limited number of times. This, for instance, is how some of the most successful data mining applications work in the financial markets - looking for that limited amount of time where a very confident prediction can be made.

**Target based on coverage.** Some times user want to know what the most ubiquitous rules are or those rules that are most readily applicable. By looking at rules ranked by coverage they can quickly get a high level view of what is happening within their database most of the time.

**Target based on “interestingness”.** Rules are interesting when they have high coverage and high accuracy and deviate from the norm. There have been many ways that rules have been ranked by some measure of interestingness so that the trade off between coverage and accuracy can be made (Berson, Smith & Thearling 2005).

It is important to recognize that even though the patterns produced from rule induction systems are delivered as if then rules they do not necessarily mean that the left hand side of the rule (the “if” part) causes the right hand side of the rule (the “then” part) to happen. Typically rule induction is used on databases with either fields of high cardinality (many different values) or many columns of binary fields. The classical case of this is the super market basket data from store scanners that contains individual

product names and quantities and may contain tens of thousands of different items with different packaging that create hundreds of thousands of Stock Keeping Units.

Sometimes in these databases the concept of a record is not easily defined within the database - consider the typical Star Schema for many data warehouses that store the supermarket transactions as separate entries in the fact table. Where the columns in the fact table have some unique identifier of the shopping basket, the quantity, the time of purchase and whether the item was purchased with a special promotion (sale or coupon). Thus each item in the shopping basket has a different row in the fact table. This layout of the data is not typically the best for most data mining algorithms which would prefer to have the data structured as one row per shopping basket and each column to represent the presence or absence of a given item. This can be an expensive way to store the data, however, since the typical grocery store contains or different items that could come across the checkout counter. This structure of the records can also create a very high dimensional space which would be unwieldy for many classical data mining algorithms like neural networks and decision trees (Berson, Smith & Thearling 2005).

Rule Induction systems provide both a very detailed view of the data where significant patterns that only occur a small portion of the time and only can be found when looking at the detail data as well as a broad overview of the data where some systems seek to deliver to the user an overall view of the patterns contained in the database. These systems thus display a nice combination of both micro and macro views:

- I) **Macro Level** - Patterns that cover many situations are provided to the user that can be used very often and with great confidence and can also be used to summarize the database.
- II) **Micro Level** - Strong rules that cover only a very few situations can still be retrieved by the system and proposed to the end user. These may be valuable if the situations that are covered are highly valuable (maybe they only apply to the most profitable customers) or represent a small but growing subpopulation which may indicate a market shift or the emergence of a new competitor (e.g. customers are only being lost in one particular area of the country where a new competitor is emerging) (Berson, Smith & Thearling 2005).

After the rules are created and their interestingness is measured there is also a call for performing prediction with the rules. Each rule by itself can perform prediction - the consequent is the target and the accuracy of the rule is the accuracy of the prediction. But because rule induction systems produce many rules for a given antecedent or consequent, there can be conflicting predictions with different accuracies. This is an opportunity for improving the overall performance of the systems by combining the rules. This can be done in a variety of ways by summing the accuracies as if they were weights or just by taking the prediction of the rule with the maximum accuracy (Berson, Smith & Thearling 2005).

The general idea given by Berson, Smith and Thearling, of a rule classification system is that rules are created that show the relationship between events captured in your database. These rules can be simple with just one element in the antecedent or they might be more complicated with many column value pairs in the antecedent all joined together by a conjunction.

The rules are used to find interesting patterns in the database but they are also used at times for prediction. There are two main things that are important to understanding a rule:

i) **Accuracy** - Accuracy refers to the probability that if the antecedent is true that the precedent will be true. High accuracy means that this is a rule that is highly dependable.

ii) **Coverage** - Coverage refers to the number of records in the database that the rule applies to. High coverage means that the rule can be used very often and also that it is less likely to be a spurious artifact of the sampling technique or idiosyncrasies of the database.

From a business perspective accurate rules are important because they imply that there is useful predictive information in the database that can be exploited - namely that there is something far from independent between the antecedent and the consequent and coverage implies how often you can use a useful rule. Having a high accuracy rule with low coverage would be like owning a race horse that always won when he raced but could only race once a year (Berson, Smith & Thearling 2005).

One of the biggest problems with rule induction systems is the sometimes overwhelming number of rules that are produced. Most of which have no practical value or interest. Some of the rules are so inaccurate that they cannot be used; some have so little coverage that though they are interesting they have little applicability, and finally many of the rules capture patterns and information that the user is already familiar with. To combat this problem researchers have sought to measure the usefulness or interestingness of rules.

Certainly any measure of interestingness would have something to do with accuracy and coverage. Berson, Smith and Thearling also expect it to have at least the following four basic behaviors:

i) Interestingness = 0 if the accuracy of the rule is equal to the background accuracy (a priori probability of the consequent). This is where a rule for attrition is no better than just guessing the overall rate of attrition.

ii) Interestingness increases as accuracy increases (or decreases with decreasing accuracy) if the coverage is fixed.

iii) Interestingness increases or decreases with coverage if accuracy stays fixed.

iv) Interestingness decreases with coverage for a fixed number of correct responses (remember accuracy equals the number of correct responses divided by the coverage).

There are a variety of measures of interestingness that are used that have these general characteristics. They are used for pruning back the total possible number of rules that might be generated and then presented to the user (Berson, Smith & Thearling 2005).

Another important measure is that of simplicity of the rule. This is an important solely for the end user. As complex rules, as powerful and as interesting as they might be, may be difficult to understand or to confirm via intuition. Thus the user has a desire to see simpler rules and consequently this desire can be manifest directly in the rules that are chosen and supplied automatically to the user (Berson, Smith & Thearling 2005).

Finally a measure of novelty is also required both during the creation of the rules - so that rules that are redundant but strong are less favored to be searched than rules that may not be as strong but cover important examples that are not covered by other strong rules (Berson, Smith & Thearling 2005).

### **Rules vs. Decision trees**

Decision trees and rules have been found to have the following similarities and differences:

i) According to Berson, Smith and Thearling, decision trees produce rules that are mutually exclusive and collectively exhaustive with respect to the training database while rule induction systems produce rules that are not mutually exclusive and might be collectively exhaustive. That is for a given record there will be a rule to cover it and there will only be one rule for rules that come from decision trees. There may be many rules that match a given record from a rule induction system and for many systems it is not guaranteed that a rule will exist for each and every possible record that might be encountered (though most systems do create very general default rules to capture these records). The reason for this difference is the way in which the two algorithms operate. Rule induction seeks to go from the bottom up and collect all possible patterns that are interesting and then later use those patterns for some prediction target. Decisions trees on the other hand work from a prediction target downward in what is known as a “greedy” search, looking for the best possible split on the next step (i.e. greedily picking the best one without looking any further than the next step). Though the greedy algorithm can make choices at the higher levels of the tree which are less than optimal at the lower levels of the tree it is very good at effectively squeezing out any correlations between predictors and the prediction. Rule induction systems on the other hand retain all possible patterns even if they are redundant or do not aid in predictive accuracy.

ii) One other thing that decision trees and rule induction systems have in common is the fact that they both need to find ways to combine and simplify rules. In a decision tree this can be as simple as recognizing



that if a lower split on a predictor is more constrained than a split on the same predictor further up in the tree that both don't need to be provided to the user but only the more restrictive one. Rules from rule induction systems are generally created by taking a simple high level rule and adding new constraints to it until the coverage gets so small as to not be meaningful. This means that the rules actually have families or what is called "cones of specialization" where one more general rule can be the parent of many more specialized rules. These cones then can be presented to the user as high level views of the families of rules and can be viewed in a hierarchical manner to aid in understanding.

## **2.7 How Data Mining Works**

How exactly is data mining able to tell you important things that you didn't know or what is going to happen next? The technique that is used to perform these feats in data mining is called modeling. Modeling is simply the act of building a model in one situation where you know the answer and then applying it to another situation that you don't (Thearling n. d).

This act of model building is thus something that people have been doing for a long time, certainly before the advent of computers or data mining technology. What happens on computers, however, is not much different than the way people build models. Computers are loaded up with lots of information about a variety of situations where an answer is known and then the data mining software on the computer must run through that data and distill the characteristics of the data that should go into the model. Once the model is built it can then be used in similar situations where you don't know the answer (Thearling n. d).

## **2.8 Architecture for Data Mining**

To best apply these advanced techniques, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on (Thearling n. d).

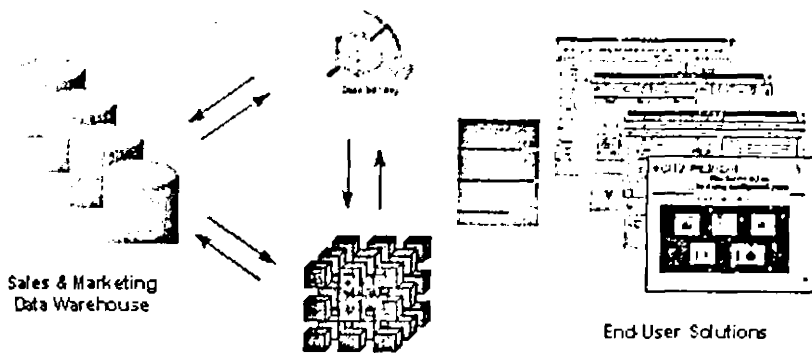


Figure 2 - Integrated Data Mining Architecture (Thearling n. d).

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. This warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access (Thearling n. d).

An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The multidimensional structures allow the user to analyze the data as they want to view their business – summarizing by product line, region, and other key perspectives of their business. The Data Mining Server must be integrated with the data warehouse and the OLAP server to embed ROI-focused business analysis directly into this infrastructure. An advanced, process-centric metadata template defines the data mining objectives for specific business issues like campaign management, prospecting, and promotion optimization. Integration with the data warehouse enables operational decisions to be directly implemented and tracked. As the warehouse grows with new decisions and results, the organization can continually mine the best practices and apply them to future decisions (Thearling n. d).

This design represents a fundamental shift from conventional decision support systems. Rather than simply delivering data to the end user through query and reporting software, the Advanced Analysis Server applies users' business models directly to the warehouse and returns a proactive analysis of the most relevant information. These results enhance the metadata in the OLAP Server by providing a dynamic metadata layer that represents a distilled view of the data. Reporting, visualization, and other analysis tools can then be applied to plan future actions and confirm the impact of those plans (Thearling n. d).

## 2.9 Knowledge Discovery

Knowledge Discovery in Databases (KDD) is the name coined by Gregory Piatetsky-Shapiro in 1989 to

describe the process of finding interesting, interpreted, useful and novel data. There are many nuances to this process, but roughly the steps are to preprocess raw data, mine the data, and interpret the results.

Recently, efforts have been made to develop new research frameworks for Data Mining (Pechenizkiy *et al.*, 2005). However, there still is a lack of attention on theories and models of Data Mining for knowledge development in business.

In the Data Mining community there have been "step-by-step data mining guides" (Lavrac *et al.*, 2004) that best describe how analytical work is done by data miners. Generally, the first step of a data miner in a Data Mining project is to understand the problem owner's concerns. Next, the data miner must prepare data in a systematic way to make data adequate and clean. Once data are prepared, Data Mining techniques and tools are applied to the data. Ideally, mining results that is interesting to the data miner would be obtained. To make the Data Mining results actionable, the data miner must explain them to the business insider. The interaction process between the business insiders and data miners is actually a knowledge-sharing process.

### **2.9.1 Pre-processing**

Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns already present in the data, the target dataset must be large enough to contain these patterns while remaining concise enough to be mined in an acceptable timeframe. A common source for data is a data mart or data warehouse.

The target set is then cleaned. Cleaning removes the observations with noise and missing data.

The clean data are reduced into feature vectors, one vector per observation. A feature vector is a summarized version of the raw data observation. For example, a black and white image of a face which is 100px by 100px would contain 10,000 bits of raw data. This might be turned into a feature vector by locating the eyes and mouth in the image. Doing so would reduce the data for each vector from 10,000 bits to three codes for the locations, dramatically reducing the size of the dataset to be mined, and hence reducing the processing effort. The feature(s) selected will depend on what the objective(s) is/are; obviously, selecting the "right" feature(s) is fundamental to successful data mining.

The feature vectors are divided into two sets, the "training set" and the "test set". The training set is used to "train" the data mining algorithm(s), while the test set is used to verify the accuracy of any patterns found.

### **2.9.2 Data warehouse Vs Mining**

It is important to clarify the difference between data warehousing and data mining. A data warehouse consists of a set of programs that extract data from the operational environment; it is a term for what you do when you bring together all the data you have collected in a useful form. Through data mining, health care providers (and other businesses) can use a warehouse to distill the often-valuable information buried within. Data mining is a term used to describe analysis of warehoused data to generate new insights. It is a much more undirected kind of analysis. Data mining begins with trend analysis and the search for patterns in the underlying data. Once a pattern of interest is identified, statistical analysis is applied to determine whether the pattern is significant. If it is found to be of significance, root cause analysis is applied to determine the cause of the trend. Interviews, telephone surveys and further data/statistical analysis are techniques that are applied during the root cause process. This is the point at which service quality improvement and marketing/communication begins (Edward Rafalski).

### **2.9.3 Results validation**

The final step of knowledge discovery from data is to verify the patterns produced by the data mining algorithms occur in the wider data set. Not all patterns found by the data mining algorithms are necessarily valid. It is common for the data mining algorithms to find patterns in the training set which are not present in the general data set, this is called over fitting. To overcome this, the evaluation uses a test set of data which the data mining algorithm was not trained on. The learnt patterns are applied to this test set and the resulting output is compared to the desired output. For example, a data mining algorithm trying to distinguish spam from legitimate emails would be trained on a training set of sample emails. Once trained, the learnt patterns would be applied to the test set of emails which it had not been trained on; the accuracy of these patterns can then be measured from how many emails they correctly classify. A number of statistical methods may be used to evaluate the algorithm such as ROC curves.

If the learnt patterns do not meet the desired standards, then it is necessary to reevaluate and change the preprocessing and data mining. If the learnt patterns do meet the desired standards then the final step is to interpret the learnt patterns and turn them into knowledge.

### **2.9.4 Proposed solution to problem**

Having gone through several common data mining algorithms, this paper intended to use the rule induction algorithms to solve the problem at hand. A local bank had so much data accumulated over the years. It was the intent of this research to help it and any other local bank that may want to automate their customer loan application appraisals by demonstrating how they can use the accumulated data to identify patterns that give information on the type of customers who are likely to pay back well, based on their credit worthiness

arrived at by use of several variables like age, gender, income bracket, employment status, education level, number of dependants, etc.

The researcher did analytics that determined the statistical significance of the variables before applying useful if-then rules from data based on the statistical significance. This was done after deciding the variables to use to appraise a customer for a credit facility qualification and hence build a model that can be used for appraisal purposes.

Two common data mining techniques for finding hidden patterns in data are clustering and classification analyses. Although classification and clustering are often mentioned in the same breath, they are different analytical approaches. Machine learning algorithms are described as either 'supervised' or 'unsupervised'. The distinction is drawn from how the learner classifies data

Clustering is an automated process to group related records together. Related records are grouped together on the basis of having similar values for attributes. This approach of segmenting the database via clustering analysis is often used as an exploratory technique because it is not necessary for the end-user or analyst to specify ahead of time how records should be related together. In fact, the objective of the analysis is often to discover segments or clusters, and then examine the attributes and values that define the clusters or segments. There are a variety of algorithms used for clustering, but they all share the property of iteratively assigning records to a cluster, calculating a measure (usually similarity, and/or distinctiveness), and re-assigning records to clusters until the calculated measures don't change much indicating that the process has converged to stable segments.

Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis, information retrieval, and bioinformatics.

Unsupervised learners are not provided with classifications. In fact, the basic task of unsupervised learning is to develop classification labels automatically. Unsupervised algorithms seek out similarity between pieces of data in order to determine whether they can be characterized as forming a group. These groups are termed clusters. In unsupervised classification, often known as 'cluster analysis' the machine is not told how the texts are grouped. Its task is to arrive at some grouping of the data.

Classification is a different technique than clustering. Classification is similar to clustering in that it also segments customer records into distinct segments called classes. But unlike clustering, a classification analysis requires that the end-user or analyst know ahead of time how classes are defined. For example, classes can be defined to represent the likelihood that a customer defaults on a loan. It is necessary that each record in the dataset used to build the classifier already have a value for the attribute used to define classes because the end-user decides on the attribute to use, classification is much less exploratory than

clustering. The objective of a classifier is not to explore the data to discover interesting segments, but rather to decide how new records should be classified –for example, is this new customer likely to default on the loan?

Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. A supervised learning algorithm analyzes the training data and produces an inferred function which should predict the correct output value for any valid input object. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way. The parallel task in human and animal psychology is often referred to as concept learning. Hence in supervised algorithms, the classes are predetermined. These classes can be conceived of as a finite set, previously arrived at by a human. The machine learner's task is to search for patterns and construct mathematical models. These models then are evaluated on the basis of their predictive capacity.

The model used classification method by use of induction rules which are easy to generate rules with, easy to interpret and have high levels of automation. One of the important problems in all of data mining is that of determining which predictors are the most relevant and the most important in building models that are most accurate at prediction. These predictors may be used by themselves or they may be used in conjunction with other predictors to form "features". And induction rules are no exceptions. There was need to understand the data so that it was easier to pick out the relevant variables to the problem at hand.

### **2.9.5 Limitations of Data Mining**

While data mining products can be very powerful tools, they are not self-sufficient applications. To be successful, data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created. Consequently, the limitations of data mining are primarily data or personnel related, rather than technology-related. Although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user. Similarly, the validity of the patterns discovered is dependent on how they compare to "real world" circumstances. Another limitation of data mining is that while it can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship.

3.1 Introduction

This chapter presents the methods and processes that were followed by the researcher to conduct the study. After analyzing the various available data mining algorithms, this research attempts to make use of rule induction technique as the underlying data mining algorithm behind a model that will appraise customers for credit worthiness.

To be effective in data mining, the model follows a four step process:

- Defining the Business Problem
- Gathering and Preparing the Data
- Model Building and Evaluation
- Knowledge Deployment

## The Data Mining Process

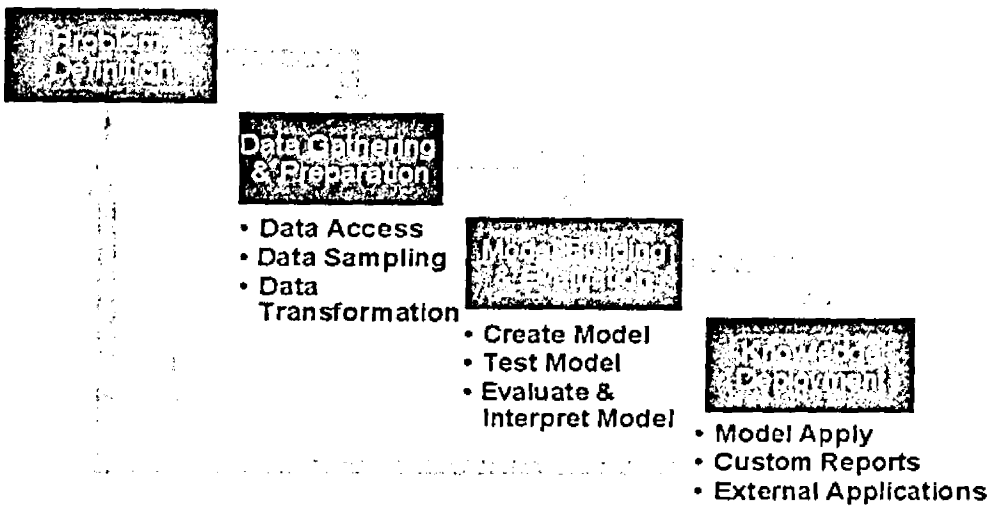


Figure 3: The data mining process

Generally, the first step was to understand the problem. Data was then prepared in a systematic way to make it adequate and clean. Once data was prepared, Data Mining techniques and tools were applied to the data. To make the Data Mining results actionable, the model was built after deliberations with the business insiders. The interaction process between the business insiders and data miners was actually a knowledge-sharing process.

### 3.2 The Data Mining Techniques used

To best apply these advanced techniques, the analyst used flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data.

#### i) Anomaly detection

Anomaly Detection identifies unusual or suspicious cases based on deviation from the norm. This was done using graphical representation of data.

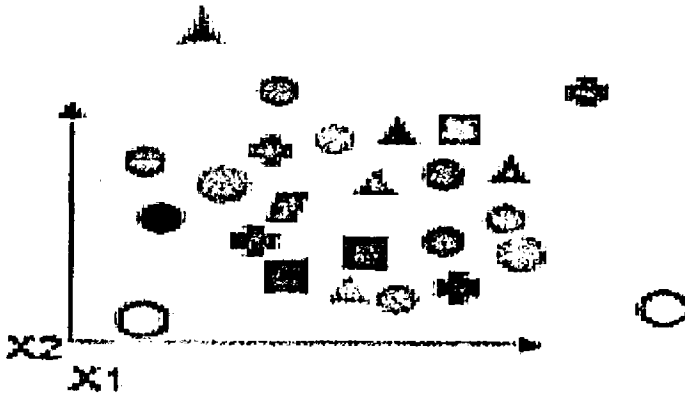


Figure 4: Outlier Detection

#### ii) Attribute Importance

Attribute Importance Ranks attributes according to strength of relationship with target attribute. It was important to develop mechanisms and processes which assist analysts and modelers to navigate through the maze of data, and identify a smaller set of variables.

Variable reduction served as the generic initial stage of variable selection process, which fuses with model-building by itself. The aim of variable reduction is to maintain a compact set of predictors that can help to accelerate model building but not losing potential predictive powers.

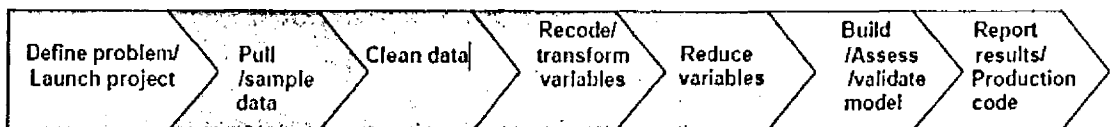


Figure 5: Typical phases of model development process (Al-Subaihi 1983)

According to Moez, Alec and Ray, data mining's Attribute Importance algorithm helps to identify the attributes that have the greatest influence on a target attribute. Often, knowing which attributes are most influential helps to better understand and manage the business and can help simplify modeling activities. Additionally, these attributes can indicate the types of data that you may wish to add to your data to augment your models.



### iii) Regression

It attempts to find a function which models the data with the least error. Regression is the oldest and most well-known statistical technique that the data mining community utilizes. Basically, regression takes a numerical dataset and develops a mathematical formula that fits the data. When you're ready to use the results to predict future behavior, you simply take your new data, plug it into the developed formula and you've got a prediction!

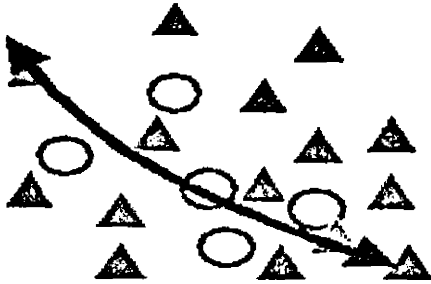


Figure 6: The regression function-Best possible line.

In the more general multiple regression model, there are  $p$  independent variables:

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \varepsilon_i$$

### iv) Rule Induction

This is the extraction of useful if-then rules from data based on statistical significance. Rule induction is one of the major forms of data mining and is perhaps the most common form of knowledge discovery. Rule induction systems are highly automated and are among the best of data mining techniques. They can be modified for use in prediction problems but the algorithms for combining evidence from a variety of rules comes more from rules of thumbs and practical experience.

In rule induction systems the rule itself is of a simple form of "if this and this and this then this". When the rules are mined out of the database the rules can be used either for understanding better the business problems that the data reflects or for performing actual predictions against some predefined prediction target.

The general idea of a rule classification system is that rules are created that show the relationship between events captured in your database. These rules can be simple with just one element in the antecedent or they might be more complicated with many column value pairs in the antecedent all joined together by a conjunction.

### 3.3 Analysis and Design

The figure below gives a summary of the analysis, the design and implementation.

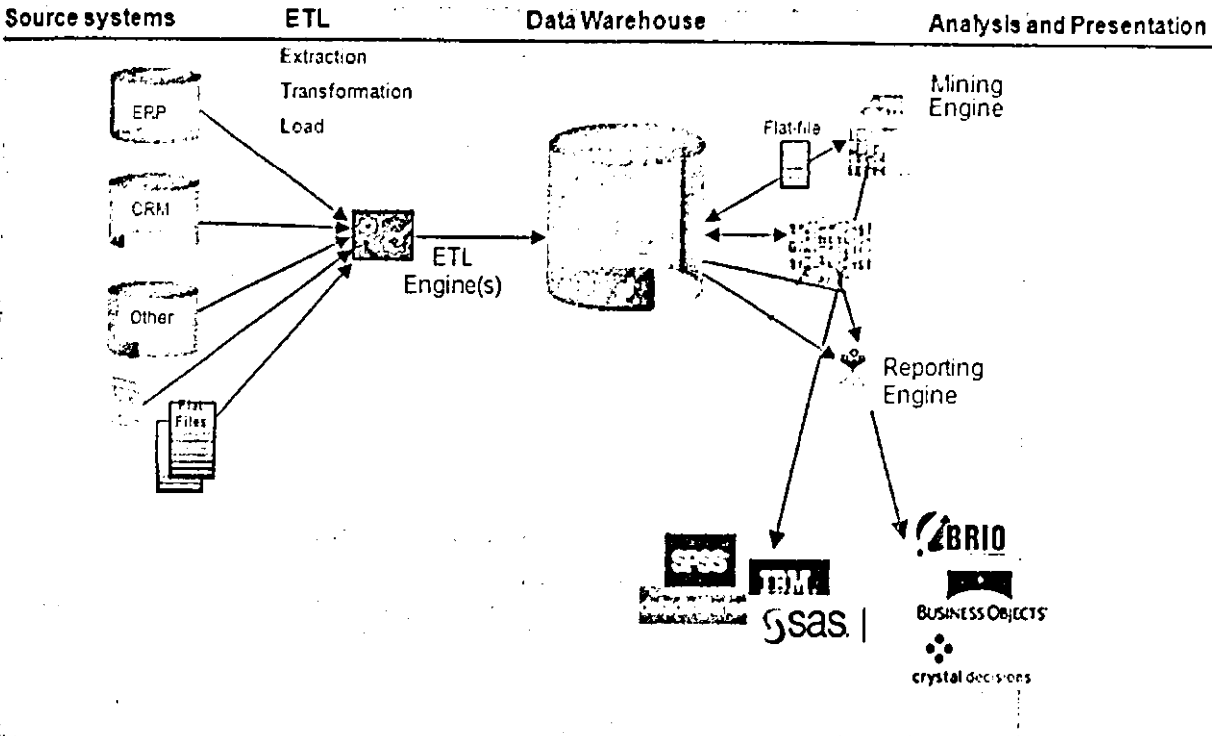


Figure 7: Summary of the analysis, design and implementation (Leopoldo 2006).

#### i) Sources of Data and Instruments of Data Collection

The researcher gathered data from a number of sources. These included databases on general customer information, credit specific databases, to physical customer files containing applications. The researcher also interviewed senior members of the credit appraisal committee, who are involved in appraising customers for a credit facility at the local bank and have a wealth of knowledge gathered over the number of years of experience to give expert insights. Four interviews were done. These interviews helped to understand the data, and also deciding the conditions to apply to the SQL queries pulling the data.

The researcher also studied relevant bank records, currently existing systems or procedures, other existing literature like books and relevant journals to gather information for this research.

#### ii) ETL

SQL queries were prepared with various conditions to be able to pull and extract data within the scope of research. After this, data was cleaned through a process of eliminating blank data and removing outliers. The remaining data was first taken through a process of translation. This ensured that all similar data was represented in the same format before finally loading the data for analysis.

### iii) Variable selection

First, all the data was coded to allow analysis by representing it numerically, it was then fed to a statistical tool whereby the Binary Logistic regression function was applied. Logistic regression is used for prediction of the probability of occurrence of an event by fitting data to a logical function or logistic curve. Like many forms of regression analysis, it makes use of several predictor variables that may be either numerical or categorical.

Data was loaded to SPSS for analysis by applying classification (binary Logistic Regression). To be able to tell the significance of each variable, difference methods of variable selection in multivariable regression were deployed to the data. They included:

- Forward Wald-Starts with no variables in the model, then for each independent variable it calculates the level of significance reflecting the variable's contribution to the model if it is included.
- Backward Elimination method-Begins with all variables included in the model and deletes one variable at a time using a partial constant.
- All-Possible-Regression-This procedure calls for considering all possible subsets of the pool of potential predictors and identifying for detailed examination a few "good" subsets according to some criterion. Various criteria for comparing the regression model may be used with the all-possible-regression selection procedure.

Each variable was then split into categories to allow a more detailed representation of data to be further looked at the modeling stage. Final variables were selected based on Weight of Evidence and Information Value.

The results inferred were as below.

Characteristic	Attribute	Comments
Age	<25	Not supported by data, but makes logical sense
Age	25 to 29	
Age	30 to 59	
Age	60+	
Gender	Female	Statistically significant
Gender	Male	
Number of TODS in past nine months	3-5	Statistically significant
Number of TODS in past nine months	6-8	
Number of TODS in past nine months	9-12	

Number of TODS in past nine months	>12	
Ratio of Avg. Paid on Time to Avg. TOD	< 1	Statistically significant
Ratio of Avg. Paid on Time to Avg. TOD	>= 1	
Avg. remittance	<2500	Statistically significant
Avg. remittance	2500-4999	
Avg. remittance	5000-7499	
Avg. remittance	7500-9999	
Avg. remittance	10000-12499	
Avg. remittance	12500-14999	
Avg. remittance	>=15000	
Ratio of Avg. remittance to Avg. TOD	< 2	Statistically significant
Ratio of Avg. remittance to Avg. TOD	2 to < 4	
Ratio of Avg. remittance to Avg. TOD	>= 4	
Performance on Other Facilities	Never taken	Not all of this data exists to validate, but makes logical sense
Performance on Other Facilities	Ever taken - no NPA	
Performance on Other Facilities	Ever taken - NPA within previous 12 months	
Performance on Other Facilities	Ever taken-existing NPA	
Number of Remittances in previous 9 months	< 2	Statistically significant. This is probably not the ideal variable. Perhaps a variable like ">= 1 remittance per month over previous 9 months" would be better
Number of Remittances in previous 9 months	3 to 5	
Number of Remittances in previous 9 months	>= 6	
Percent of Previous TODs paid on Time	<=25%	We cannot put this variable in a predictive model, because our data does not follow a cohort prospectively. Therefore, all customers who historically paid on time will show a bad rate of 0%
Percent of Previous TODs paid on Time	> 25% and <= 50%	
Percent of Previous TODs paid on Time	> 50% and <= 75%	
Percent of Previous TODs paid on Time	> 75% and < 100%	
Percent of Previous TODs paid on Time	100%	

Table 3: Statistical significance of variables

#### iv) Weightings

This was done as an attempt to produce an inferred function which should predict the correct output value for any valid input object. It was necessary that each record in the dataset used to build the classifier already have a value for the attribute used to define classes because the end-user decides on the attribute to use.

Weights were calculated based on statistical significance which was determined using a combination of Information Value, Weight of Evidence and induction rules. This eliminated the need of parameter tuning at the model. The weights were used to generate simple rules that that implements the final classification done by the model.

Weight of evidence (WOE) and information value are increasingly popular in the analytical and modeling community as they represent good alternatives to approximate non-linearity in the data.

#### **The weight of evidence (WoE).**

A criterion used to gauge the predictive power of each attribute; it measures the strength of each attribute, or grouped attributes, in separating the good and bad accounts i.e. the odds of a person with that attribute being good or bad. Negative numbers imply that the particular attribute is isolating a higher proportion of bads than goods and the larger the difference between subsequent groups, the higher the predictive ability of this characteristic.

#### **The information value (IV).**

Is a criterion used to gauge the predictive power (strength) of the characteristic.

They are computed with the following as:

**Weight Of Evidence** =  $\text{Log}(\text{Distribution Good}/\text{Distribution Bad})$

**Information Value** =  $\{\Sigma(\text{Dist Good} - \text{Dist Bad}) \times \text{Weight Of Evidence}\}$

The empirical rule of thumb for assessing the Information Variable is as follows:

Less than 0.02: the variable is not predictive;

0.02 to 0.1: the variable has weak predictive power;

0.1 to 0.3: the variable has medium predictive power;

0.3+ : the variable has strong predictive power.

The final design of the model was based on the findings below:

Characteristic	WEIGHT	WOE	Information Value (IV) (nearest 4 decimal places)		
No of Previous loans taken	43.27%		0.1779		433
		13.43692	3-10	433	
		-123.95081	11-18	150	
		-240.58337	19-26	80	
		-100.85703	27-34	200	
		0.00000	35-42	300	
No of Remittances	43.13%		0.1773		431
		-108.16667	0-6	50	
		16.16596	7-13	300	
		79.42406	14-20	431	
		-53.85737	21-27	150	
		0.00000	28-34	250	
Existing NPA	6.57%		0.027		66
		1.54697	N	66	
		-175.1001	Y	10	
Gender	4.38%		0.018		44
		32.4731548	F	44	
		-5.559254478	M	5	
Period with Bank	1.56%		0.0064		16
		2.747147941	12-21	8	
		14.02897086	22-31	16	
		-2.649647471	32-41	4	
		3.060473031	42-51	10	
		-8.099046109	52-61	2	
Ever taken facility	1.36%		0.0056		14
		19.60459927	N	14	
		-0.751028312	Y	3	
Age	0.19%		0.0008		2
		6.143614161	18-29	1	

		-5.254207847	30-41	0.1	
		2.638570229	42-53	0.8	
		-0.735229342	54-65	0.4	
		26.99966975	66-77	2	
Remaining period %	0.19%		0.0008		2
		-4.03742	0-19	0.3	
		2.78292	20-39	1.8	
		3.76714	40-59	2	
		-0.35563	60-79	0.8	
		1.85469	80-100	1.4	
	100.00%		0.4111		1000

Table 4: Weights of Evidence and Information Value

The weights are based on the significance of each variable to the decision of considering a client as a good candidate.

### 3.4 Implementation

SAP Business Objects were used to design the model. SAP Business Objects delivers data mining and predictive analytics. Using powerful data mining technology, Business Objects Predictive Workbench enables a researcher to pore over historical business information and create predictive models that can be utilized to solve both horizontal and industry-specific business problems. The predictive models help organizations achieve specific business goals, such as maximizing marketing efforts, designing optimal pricing plans, improving operational efficiencies and, most importantly, gaining agility and competitive differentiation.

The XCELSIUS tool was used to apply the findings of the analysis to a front end model. The business objects have the ability to alter cell values in real time, in line with what the end user input or does on the dashboard. SAP BusinessObjects Xcelsius Enterprise is a point-and-click data visualization tool designed specifically to create interactive analytics and dashboards with secure, live connections to SAP BusinessObjects Enterprise and SAP BusinessObjects Edge BI. You can share these meaningful visualizations live via Microsoft Office, Adobe PDF, the Web, Crystal Reports, or the SAP BusinessObjects business intelligence (BI) portal.

SAP BusinessObjects Xcelsius Enterprise:

- Provides access to secure and personalized dashboards from anywhere

- Offers personalized dashboards built on top of a trusted, secure BI platform
- Empowers business users to create dashboards with the power of what-if analysis
- Extends the power of business intelligence to more users by offering visual simplicity in analysis

**Rule induction**

This is the extraction of useful if-then rules from data based on statistical significance. Rule induction is one of the major forms of data mining and is perhaps the most common form of knowledge discovery.

If-then rules from data based on statistical significance were applied to give a function that is then applied to the model to give a live dash board.

The bane of rule induction systems is also its strength - that it retrieves all possible interesting patterns in the database. This is a strength in the sense that it leaves no stone unturned but it can also be viewed as a weakness because the user can easily become overwhelmed with such a large number of rules that it is difficult to look through all of them.

Rule induction systems are highly automated and are among the best of data mining techniques for exposing all possible predictive patterns in a database. They can be modified to for use in prediction problems but the algorithms for combining evidence from a variety of rules comes more from rules of thumbs and practical experience. In this case, the function was obtained by use of statistical methods (Regression, Information value and Weights of evidence) and a lot of expertise advice.

Rule induction systems when used for prediction, are like having a committee of trusted advisors each with a slightly different opinion as to what to do but relatively well grounded reasoning and a good explanation for why it should be done, hence the choice to use it due to the many number of attributes to be considered, each with different weightings. Typically rule induction is used on databases with either fields of high cardinality (many different values) or many columns.

In rule induction systems the rule itself is of a simple form of “if this and this and this then this”. The rules themselves consist of two halves. In this research, the left hand side, called the antecedent consisted of multiple conditions which must all be true in order for the consequent to be true at the given accuracy.

Since Xcelsius has excel embedded onto it, the weightings were subjected to the appropriate induction rules to generate the engine behind the final dashboard.

On applying the function, the rules behind the dashboard are as below:

Characteristic	WEIGHT	WOE	Information Value (IV) (nearest 4 decimal places)		Rating
No of Loans taken	43.27%		0.1779		432.7414
		13.43692	3-10	433	77.0307
		-123.95081	11-18	150	26.6850
		-240.58337	19-26	80	14.2320

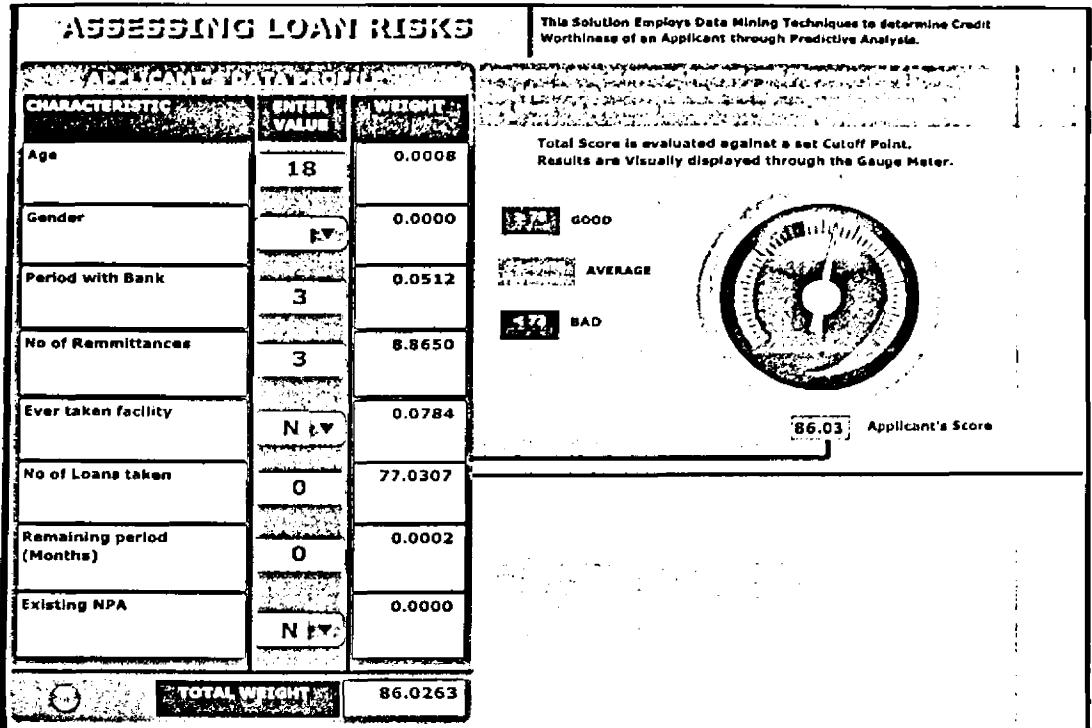


		-100.85703	27-34	200	35.5800
		0.00000	35-42	300	53.3700
<b>No of Remittances</b>	<b>43.13%</b>			<b>0.1773</b>	<b>431.2819</b>
		-108.16667	0-6	50	8.8650
		16.16596	7-13	300	53.1900
		79.42406	14-20	431	76.4163
		-53.85737	21-27	150	26.5950
		0.00000	28-34	250	44.3250
<b>Existing NPA</b>	<b>6.57%</b>			<b>0.027</b>	<b>65.6775</b>
		1.54697	N	66	1.7820
		-175.1001	Y	10	0.2700
<b>Gender</b>	<b>4.38%</b>			<b>0.018</b>	<b>43.7850</b>
		32.4731548	F	44	0.7920
		-	M	5	0.0900
		5.559254478			
<b>Period with Bank</b>	<b>1.56%</b>			<b>0.0064</b>	<b>15.5680</b>
		2.747147941	12-21	8	0.0512
		14.02897086	22-31	16	0.1024
		-			
		2.649647471	32-41	4	0.0256
		3.060473031	42-51	10	0.0640
		-			
		8.099046109	52-61	2	0.0128
<b>Ever taken facility</b>	<b>1.36%</b>			<b>0.0056</b>	<b>13.6220</b>
		19.60459927	N	14	0.0784
		-	Y	3	0.0168
		0.751028312			
<b>Age</b>	<b>0.19%</b>			<b>0.0008</b>	<b>1.9460</b>
		6.143614161	18-29	1	0.0008
		-			
		5.254207847	30-41	0.1	0.0001
		2.638570229	42-53	0.8	0.0006
		-			
		0.735229342	54-65	0.4	0.0003
		26.99966975	66-77	2	0.0016
<b>Remaining period %</b>	<b>0.19%</b>			<b>0.0008</b>	<b>1.9460</b>
		-4.03742	0-19	0.3	0.0002
		2.78292	20-39	1.8	0.0014
		3.76714	40-59	2	0.0016
		-0.35563	60-79	0.8	0.0006
		1.85469	80-100	1.4	0.0011
	<b>100.00%</b>			<b>0.4111</b>	<b>1,000.0000</b>

Table 5: The weights behind the rules

To be able to do a gauge so that the model gives a good or bad rating, a risk meter was designed and a minimum value was given through the use of significance and expert consultation. The risk meter uses the total weight output by the model after accepting a user's input to give a rating of bad or good.

Fig. 8 Diagram showing the final model



### 3.5 Presentation of Findings

After the rules are created there is also a call for performing prediction with the rules. However, it is important to note that when the rules are mined out of the database the rules can be used either for understanding better the business problems that the data reflects or for performing actual predictions against some predefined prediction target.

Results were largely presented in tables and graphs and explained thematically (continuous prose under key topics and subtopics).

A summary conclusion was also made based on the findings.

## CHAPTER FOUR: RESULTS AND DISCUSSION

### 4.1 Introduction

This chapter presents the main findings of the study. It contains analysis of results obtained after putting into practice the methodologies highlighted in chapter three.

When the rules were mined out of the database they were used first for understanding better the business problems that the data reflected before performing actual predictions against some predefined prediction target.

For a bank, there are certain factors that it needs to consider before giving a loan facility to any customer.

Some of the factors which are not part of personal details of clients' data include:

1. No of other loan facilities issued to same client in the past.
2. Average number of days he takes to make repayments
3. Average times the client has paid previous loans beyond time
4. Average times the client has paid previous loans on time
5. Number of remittances he has made to the bank so far
6. Average amount of remittances
7. The last salary date
8. The last Loan facility amount given
9. Last date of loan application
10. Last due date of payment
11. Last repaid date

Out of these, certain variable were found to be of high significance when deciding if a client is a potential defaulter. This research highlighted the following as having significance in predicting the probability of default:

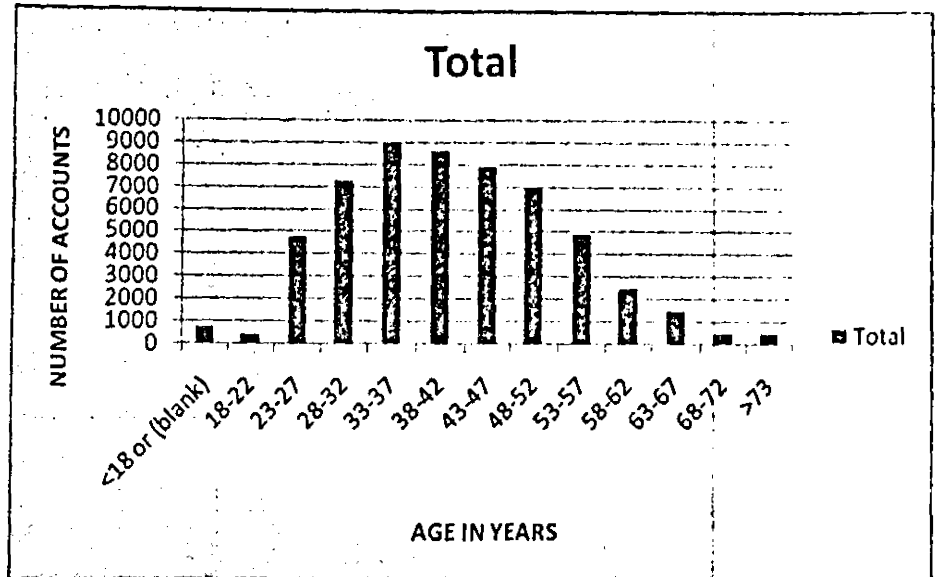
1. Number of loan facilities previously taken.
2. Number of remittances made to the bank before the application
3. Any existing Non-performing Assets.
4. Gender
5. Period of time the client has been in the bank
6. Whether the client has ever taken a loan facility with the bank.
7. Age of the client
8. Remaining period of payment for the other loan facilities.

## 4.2 Data exploration and findings

Below are the findings of how a number of variables related to the data. This helps to understand better the business problems that the data reflects.

### DISTRIBUTION ON AGE

Age	Number of accounts
<18	801
18-22	454
23-27	4806
28-32	7304
33-37	8992
38-42	8650
43-47	7940
48-52	7010
53-57	4894
58-62	2537
63-67	1481
68-72	339
>73	667
<b>Grand Total</b>	<b>55875</b>



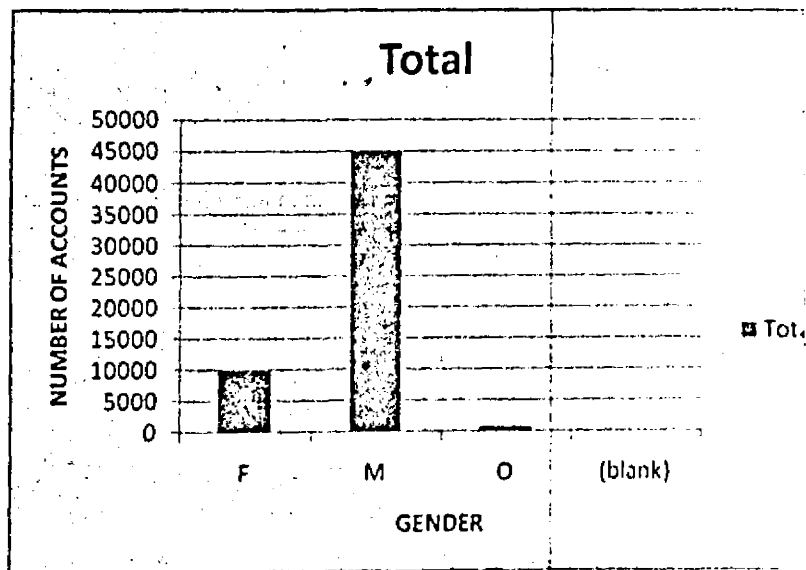
Based on the chart it is notable that the data represents a normal distribution, with most ages concentrating between 28 years and 52 years. Most borrowers were between the ages of 33-37 years. The older generation was also found to have less defaulting probability.

Among outliers was:

- <18 years
- >63 years

### DISTRIBUTION ON GENDER

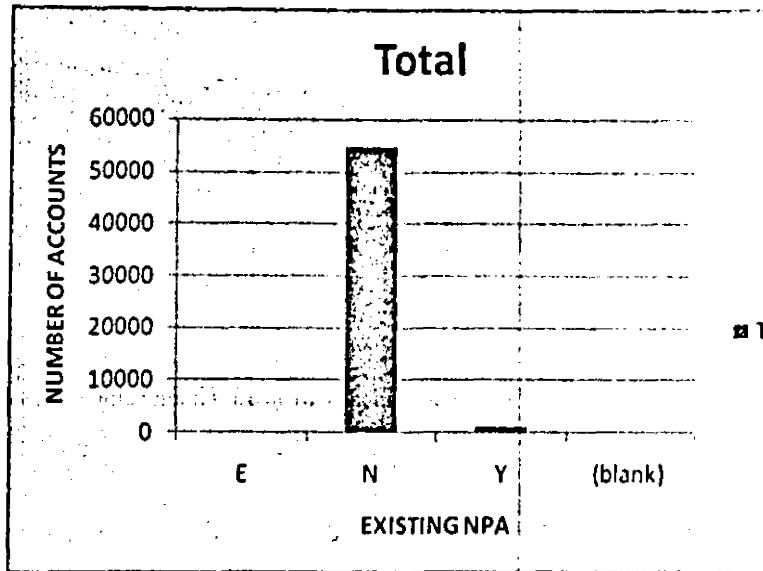
GENDER	NUMBER OF ACCOUNTS
F	9943
M	45053
O	721
(blank)	158
<b>Grand Total</b>	<b>55875</b>



Fewer females actually borrow loans. Males are at 81%, while females are at 18%, hence the reason why most defaulters are found to be male.

**EXISTING NPA**

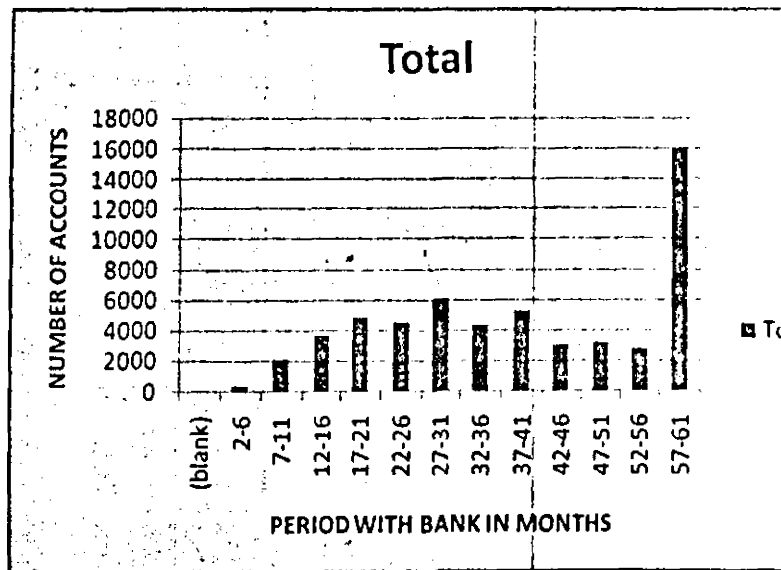
EXISTING NPA	NUMBER OF ACCOUNTS
E	29
N	54722
Y	1124
<b>Grand Total</b>	<b>55875</b>



98% do not have existing nonperforming assets. 60% of clients with existing non-performing loans, were found to definitely have a high probability of being defaulters.

**PERIOD WITH BANK**

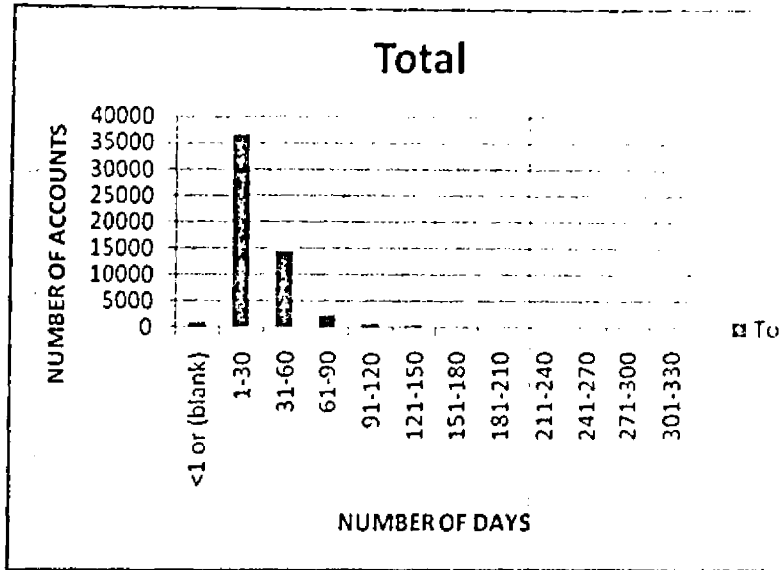
PERIOD WITH BANK	NUMBER
2-6	348
7-11	2039
12-16	3671
17-21	4787
22-26	4515
27-31	6049
32-36	4281
37-41	5222
42-46	3006
47-51	3180
52-56	2760
57-61	16017
<b>Grand Total</b>	<b>55875</b>



29% of the borrowers have been with the bank for about 5 years. The longer a customer has been with the bank, the more likely they are to default. The defaulting probability increased beyond the fourth year.

**MAXIMUM REPAYMENT DAYS**

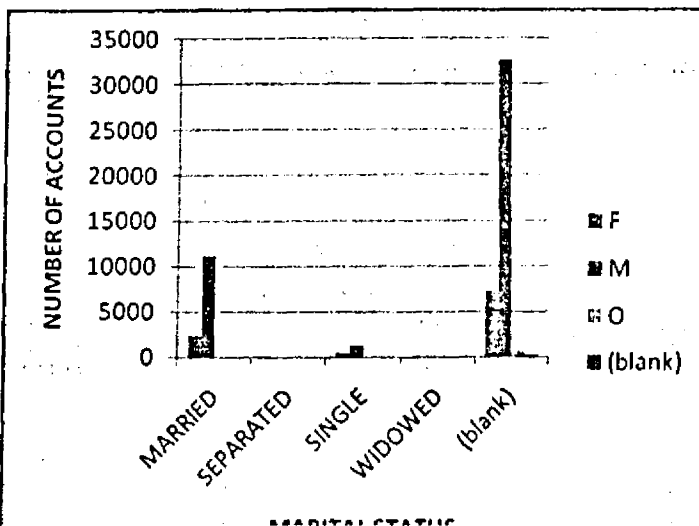
MAXIMUM DAYS	NUMBER OF ACCOUNTS
<1 or (blank)	964
1-30	36399
31-60	14382
61-90	2475
91-120	725
121-150	368
151-180	219
181-210	169
211-240	109
241-270	45
271-300	16
301-330	4
<b>Grand Total</b>	<b>55875</b>



67% of the borrowers pay within the 30 days set for repayment. Hence defaulters rate expected to be below 25%.

**MARITAL STATUS VS GENDER**

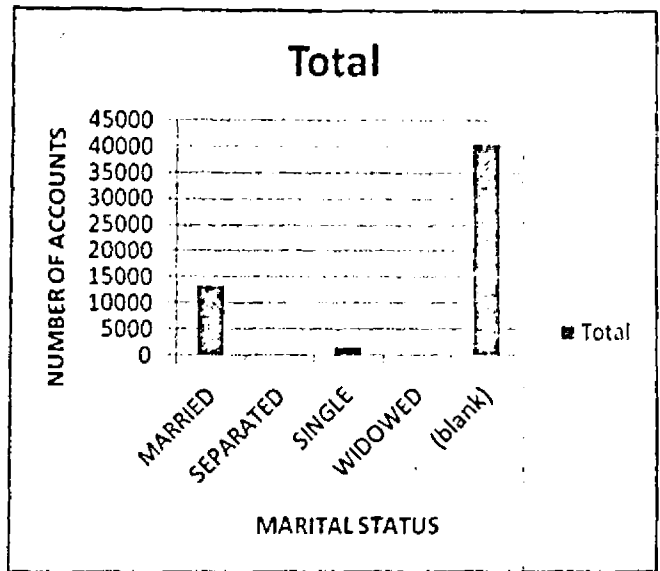
MARITAL STATUS	GENDER				Grand Total
	F	M	O	(blank)	
MARRIED	2293	11168	43		13504
SEPARATED	1				1
SINGLE	425	1264	2		1691
WIDOWED	10	3	1		14
(blank)	7214	32618	675	158	40665
<b>Grand Total</b>	<b>9943</b>	<b>45053</b>	<b>721</b>	<b>158</b>	<b>55875</b>



There's a significantly high number of married men than single men borrowing.

**MARITAL STATUS**

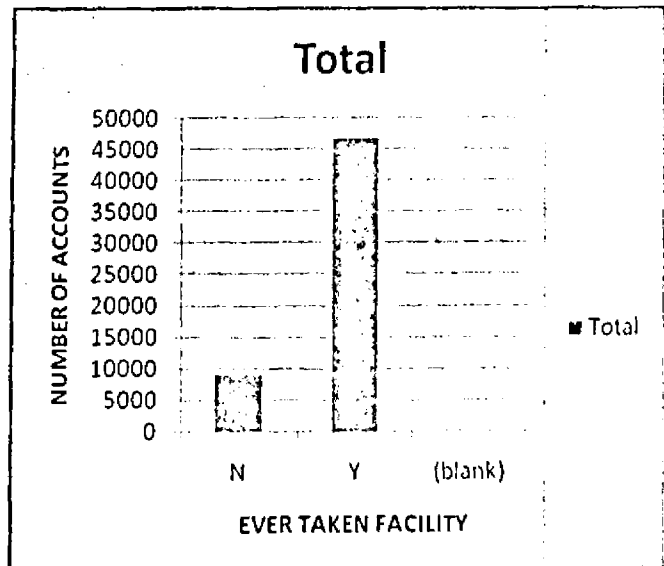
MARITAL STATUS	Count of ACCT_NUMBER
MARRIED	13504
SEPARATED	1
SINGLE	1691
WIDOWED	14
(blank)	40665
<b>Grand Total</b>	<b>55875</b>



Since 73% of this information is blank, it shows a probability of wrong data entry that is likely to affect model performance.

**EVER TAKEN FACILITY**

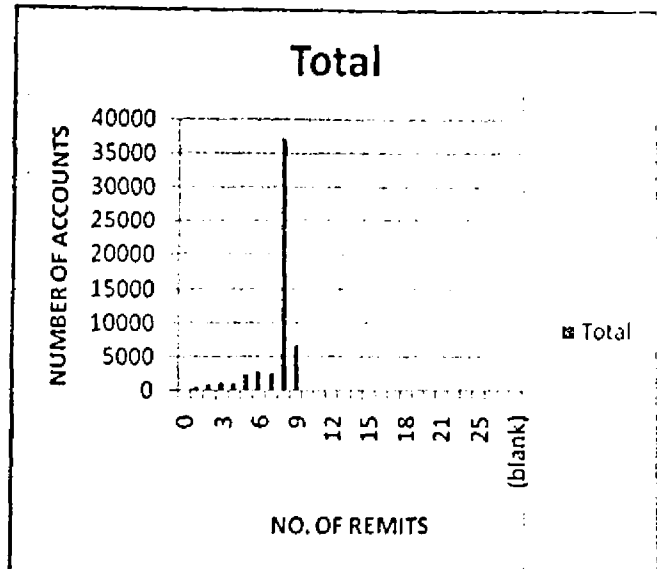
EVER TAKEN FACILITY	NUMBER OF ACCOUNTS
N	8994
Y	46881
<b>Grand Total</b>	<b>55875</b>



84% were repeat borrowers. This attribute hence should have a high significance in our findings.

## NUMBER OF REMITTANCES

NO. OF REMITS	NUMBER OF ACCOUNTS
0	136
1	556
2	894
3	1272
4	1048
5	2372
6	2677
7	2557
8	37206
9	6767
10	39
11	18
12	13
13	15
14	16
15	17
16	162
17	53
18	34
19	2
20	2
21	1
22	1
24	9
25	5
26	1
32	2
<b>Grand Total</b>	<b>55875</b>



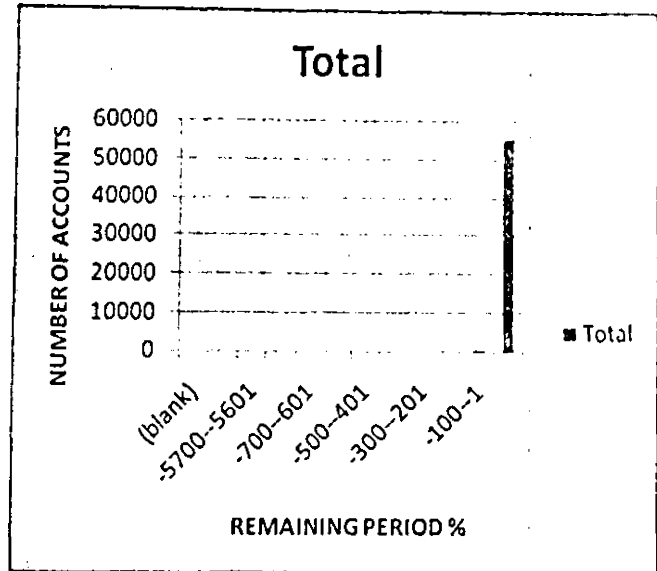
67% of borrowers had an average of 8 remittances.

The less the number of remittances you have made to the bank, the higher your chances of defaulting are.



## REMAINING PERIOD PERCENTAGE

REMAINING PERIOD	NUMBER OF ACCOUNTS.
-6000--5901	1
-5700--5601	1
-800--701	3
-700--601	1
-600--501	2
-500--401	5
-400--301	7
-300--201	9
-200--101	9
-100--1	353
0-99	55484
<b>Grand Total</b>	<b>55875</b>



•Generally most people with facilities have a remaining period of between 0-99%.

•With investigation the data with negatives were found to have been wrongly input. Some are loan accounts that were not closed even after the facility was fully paid.

### 4.3 Evaluation and Validation

Model validation should be employed at the time of model development. It is a process that is comprised of three general types of activities:

- (i) The collection of evidence in support of the model's design, estimation and evaluation at the time of development;
- (ii) The establishment of on-going monitoring and benchmarking methods by which to evaluate model performance during implementation and use;
- (iii) The evaluation of a model's performance utilizing outcomes-based measures and the establishment of feedback processes which ensure that unexpected performance is acted upon.

The first activity was taken care of at the design stage. Statistical measures were obtained before implementation. Rule induction was also the data mining algorithm chosen for the development of the model because of the following advantages:

- i) Rule Induction leaves no stone unturned. In that it retrieves all possible interesting patterns in the database.
- ii) Rule induction systems when used for prediction, are like having a committee of trusted advisors each with a slightly different opinion as to what to do but relatively well grounded reasoning and a good explanation for why it should be done. This was fitting as our data consisted of different attributes that have influence on the final target performance with difference ratios of influence.

According to Berson, Smith and Thearling, rule-based systems are evaluated based upon two primary metrics of interest: Coverage and predictive accuracy.

i) **Accuracy** - How often is the rule correct?

ii) **Coverage** - How often does the rule apply?

Just because the pattern in the data base is expressed as rule does not mean that it is true all the time. Thus just like in other data mining algorithms it is important to recognize and make explicit the uncertainty in the rule. This is what the accuracy of the rule means. The coverage of the rule has to do with how much of the database the rule “covers” or applies to. In some cases accuracy is called the confidence of the rule and coverage is called the support.

**Steps taken to measure model accuracy** (Engelmann, Hayden & Tasche 2003).

- i) Samples of data were obtained. The data was run through the model to obtain the ratings or score given by the model.
- ii) A matrix was obtained of defaulters versus non-defaulters and their score highlighted. (They were arranged from highest to lowest for ease of analysis).
- iii) Within the matrix, if defaulter and non-defaulter had the same rating, an accuracy of 0.5 was awarded. If a defaulter was found to have a lower score than the non-defaulter, an accuracy of 1 was awarded. If the non-defaulter was found to have a lower score than a defaulter, an accuracy of 0 was awarded.

		Non-Defaulters			
		b	c	d	
<b>Imposed Rating</b>		76	78	67	
<b>Defaulters</b>					
a	68	1	1	0	
e	78	0	0.5	0	
<b>Totals:</b>		1	1.5	0	

Table 6 Example of a matrix based on a sample of 5

- iv) A summation of the totals on each column was obtained. (For instance, in the example above, the total was 2.5.)
- v) Accuracy was the obtained by dividing the total obtained above by the number of comparisons done within the matrix. For example, in the above example:
 
$$\frac{\text{Total ratings}}{\text{Total Comparisons}} = \frac{2.5}{(2*3)} * 100 = 41.7\%$$
- vi) Different samples of different data combinations of data were used and a range of accuracies were obtained.

- vii) A Confidence Interval of 98% was assumed and the accuracy interval was seen to range between 48% and 71%.

#### **4.4 Results of evaluation and discussion**

The model was hence found to have an average accuracy of 59.5%. This was obtained after testing with 10 samples of data. This meant that about 60% of the time, the model rating was correct. Several factors were seen to have affected the model performance. These are:

- i) Wrong data entry- Some data could have been misrepresented due to inaccuracy during data capture. For example, no under-aged customer should obtain a loan, yet there were data samples that had ages below 18. Wrong data entry also affected the final size of data set used to develop the model as all data with blanks was eliminated. Only 55 000 records were used to build the model.
- ii) Few numbers of variables- The number of variables picked as the most statistically significant were too few for them to be able to give the true picture on the ground.
- iii) Weights of variables – The weights awarded to variables were found to be not fully accurate as they could have been lower than the reality.
- iv) One variable was found to be too strong so that the effect of the other variables was not easily evaluated. The number of previous loans taken variable carried a lot of weight as it seemed to have been found to be very significant in deciding the credit worthiness of customers.

## CHAPTER FIVE: CONCLUSIONS AND RECOMMENDATIONS

### 5.1 Relevance of study

The Model is able to give a response of whether a client whose data is run through is a good credit facility candidate or not. It is also able to give the response within a significantly short period of time, as compared to the time period taken by the bank to evaluate a customer and appraise them via the current committee meetings.

It also takes care of any partiality that may be introduced by a human evaluation.

The study has proven that an organization can use information hidden within its data banks to ease decision making and increase efficiencies within its processes.

Below are the factors that were found to be most important to be considered when appraising a salaried customer for a credit facility, in order of their statistical significance:

- i) Number of Loans previously taken
- ii) Number of Remittances made to the bank
- iii) Number of existing Non-Performing loans
- iv) Gender
- v) The period the customer has been with the Bank
- vi) If the customer has ever taken a credit facility with the bank before
- vii) Age of customer
- viii) Remaining period for the outstanding loans.

On using rule induction to build the model, it was used to gain a better understanding of the business problems that the data reflected before performing actual predictions against some predefined prediction target.

Some of the findings discovered from the data showed that the defaulting status was more common between the ages of 28 and 47. The older generation was also found to have less defaulting probability. It was also found that, the more the number of loan facility the client has, the higher the chances of defaulting. The longer a customer has been with the bank, the more likely they are to default. The defaulting probability increased beyond the fourth year. This is a trend that the institution needed to look at closely as it goes against logic. About 20% of the married and the single people are defaulters. The rate is slightly higher for singles. The less the number of remittances you have made to the bank, the higher your chances of defaulting are. 60% of clients with existing non-performing loans were found to definitely have a high probability of being defaulters. Male have a higher defaulting rate than female have. It was also noted that a very small percentage of borrowers were female. Probability of default is significantly low if the average remittances of a client are between 20 000 and 40 000. However, customers found with remittances ranging between 60 000 and 70 000 have a significantly high probability of defaulting. An average remittance of below 10 000 is also an indicator of high risk of defaulting. It was also found that if

there's a tendency to pay late in over half of their loan facilities, the probability of being a defaulter is close to 1.

The results have shown that it is possible for the bank to use its volumes of data lying within their data centers to infer great knowledge from them. Their volumes of data have proven to be a sleeping giant that can be turned into gem stones, if well managed.

Customer data was used to deduce the computations that a human mind is forced to do as they appraise customer loan applications within committees. Due to the limiting factor of the number of computations the mind can handle, errors are bound to interfere with the chance of a corporate ripping the best from its market.

Efficiency and service level agreements can be improved greatly, not to mention accuracy in decision making. The results clearly show the relationships between characteristics of a customer data and their probability to default a credit facility.

Data mining hence can be used to extract hidden patterns within data that can be used as pointers to fasten the day to day activities within an organization.

## **5.2 Limitations of the Model and further work**

There is the chance that the wrong data will be fed to the model, either maliciously or in error. This can be taken care of alongside other checks and balances that the bank has to apply in its data capturing audit processes. It can also be taken care of by introducing an online application process whereby the client does the application online, and the data required for processing the application is fed to the model via integrations from other source systems after proper verification.

This model was based on 55,000 records only, which was found to have been clean. Based on the Original dataset of over 160,000 records from the data warehouse, this is only 34% of the population and hence there is chance that the sample did not fully represent the situation at the ground.

Further work on this research topic could involve including data on the cases that had previously been appraised negatively. This data was not used because there is no soft data bank on the rejected cases. There is need for checking the other side of the coin. Only then can it be proven that all hidden patterns have been explored and their contribution considered in knowing the most significant factors to consider when appraising customers for loan facilities, their weights and the final appraisal decision.

More work can also be done if the newly introduced Credit Bureau Reporting can be used to source data of the same customers that may have taken loans in other financial institutions within the country and the performances of their loans.

### **5.3 Conclusion**

Two main factors have been deduced as having the highest weightings in consideration to get a credit facility in a bank. The number of loans previously taken carries the highest importance in classifying a customer as a good candidate for another loan or a bad one. This is so as the nature of his previous repayment pattern can be inferred and projections of a repeat case done. The number of remittances done in the bank before application is also a key indicator of a good or bad credit facility candidate. The more remittances a customer has made, the less likely he is to default. This is probably because of the emotional connection the customer will already have formed with the longer period of interaction with their financial service provider.

Existence of non-performing loan by the candidate also has some consideration, though not as high as the first two factors. This could probably be caused by the fact that different credit facilities go to different usages. The client's financial status could be affected temporarily by factors that are short lived.

Although it is harder to predict the probability of female applicants default rate, it was deduced that male applicants are higher defaulters than female applicants. This result could have been caused by the fact that 81% of all applicants were men while female were a mere 18%.

How long a customer has been with the bank, whether they have ever taken a credit facility with the bank, the remaining repayment period if any, and their age all have significance in classifying defaulters, though with very low weightings.

The model was found to have an average of about 60% accuracy level of predicting good or bad borrowers. This proves that rule induction can indeed be used as the underlying algorithm to build a model that should automate the current credit committee meetings that congregate to appraise customers for borrowing.

### **5.4 Recommendations**

For the bank to be able to predict the probable behavior of a customer's loan performance, they must have the history of the client's financial data. This may need an extension to borrow from the introduced Kenyan Banks Credit Bureau database so as to have a three-sixty degree view of a customer.

A limit of the number of remittances done to the bank before a loan application should be set. This will allow time for bonding with the customer and getting the emotional connection that will perhaps avoid *untruthfulness by the customer*.

The bank can also encourage more female applicants to take loans, by introducing products specifically for them. This is especially so because there are far less female defaulters than there are men, yet there are far too many male applicants as compared to female.

This research can be extended by building different models based on different weightings and then testing the accuracy of each model. This will show the effect of wrong weightings being imposed on to the variables. Data mining will hence be fully exploited to be able to extract all hidden patterns.

## REFERENCES

- An Introduction to Data Mining: Discovering hidden value in your data warehouse. Available at: <Http://www.thearling.com> [Accessed on August 2010]
- Berson, A., Smith, S. & Thearling, K., 1999. *Building Data Mining Applications for CRM* : McGraw-Hill Professional
- Frawley, W., Piatetsky-Shapiro, G. & Matheus, C. 1992. *Knowledge discovery in databases: an Overview*: American Association for Artificial Intelligence
- Fayad, U. 1997. *Data mining and Knowledge Discovery*: Kluwer Academic Publishers
- Weiss, S. & Indurkha, N. 1995, *Journal of Artificial Intelligence Research*, vol. 3, no. 1.
- Forcht, K. & Cochran, K., *Using data mining and data warehousing techniques*, James Madison University, Harrisonburg, Virginia, USA, James Madison University, Harrisonburg, Virginia, USA
- Kerlinger, F.N. 1973. *Foundation of behavioral research*. New York: Holt Rinehart and Winston.
- Orodho, A. 2003, *Essentials of education and social science research methods*: Masola Publishers, Nairobi.
- Rafalski, E. 2004, Using data mining/data repository methods to identify marketing opportunities in health care, *Journal of Medical Systems*, vol. 28, no. 3.
- Moez, H., Alec, C. & Ray, F. 2006. *Variable Selection in the Credit Card Industry*, Royal Bank of Scotland: Bridgeport, CT
- Leopoldo, B. 2006. *Oracle BI solution to business pain areas*, Oracle Corporation, IBERIA
- Al-Subaihi, A. 1983, *Variable Selection in Multivariable Regression Using SAS/IML*, Institute of Public Administration Riyadh 11141, Saudi Arabia
- Engelmann, B., Hayden, E., Tasche, D. 2003, Measuring the Discriminative Power of Rating Systems, *Banking and Financial Supervision* No 01/2003