ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

POSSIBLE APPLICATION OF DATA MINING TECHNOLOGY IN SUPPORTING CREDIT RISK ASSESSMENT: THE CASE OF NIB INTERNATIONAL BANK S.C.

By
MERETEWORK SHAWUL

A thesis submitted to the School of Graduate Studies of Addis Ababa University in partial fulfillment of the requirements for the Degree of Master of science in Information Science

JULY 2004

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS

POSSIBLE APPLICATION OF DATA MINING TECHNOLOGY IN SUPPORTING CREDIT RISK ASSESSMENT: THE CASE OF NIB INTERNATIONAL BANK S.C.

By

Meretework Shawul

Name and Signature of Members of the Examining Board

Dr. Gashaw Kebede, Chairman, Examining Board _____

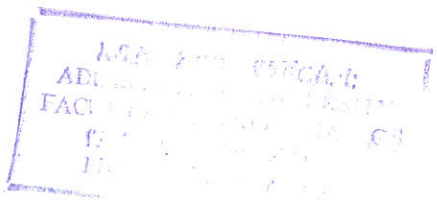
Prof. Yohannes Abate, Advisor _____

Ato Nigussie Tadesse, Advisor _____

Dr. Lishan Adam, External Examiner _____

DEDICATION

I dedicate this paper to my parents, Ato Shawul Areda and W/ro Tsehaitu Tesgera
I owe everything I have achieved and everything I am to you



ACKNOWLEDGEMENT

I would like to thank many people for their kindness and assistance during the preparation of this thesis. I sincerely thank my thesis advisors, Ato Nigussie Tadesse and Prof. Yohannes Abate for their guidance and encouragement. Special thanks goes to my sisters Meskerem Shawul and Kidist Shawul without whose constant follow up and guidance, this paper would not have seen light.

I truly appreciate the support of the people at Nib International Bank S.C. Special thanks is due to Ato Abinet Takele for his help throughout the process of collecting information. I also want to thank my friends and classmates who in one way or another contributed to the success of my thesis.

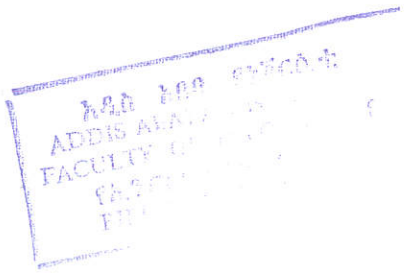
Last but by no means least, I would like to thank my father, mother and my brother Dr.Biniam Shawul, who have given me great support and encouraged me throughout this difficult but exciting journey.

Above all, I thank God who surrounded me with so many wonderful people.

Meretework Shawul

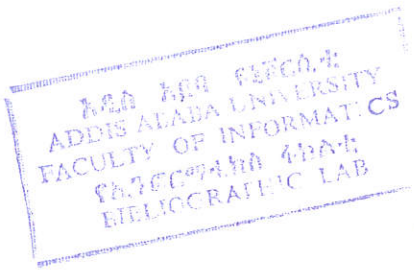
TABLE OF CONTENTS

LIST OF FIGURES



LIST OF TABLES

LIST OF APPENDICES



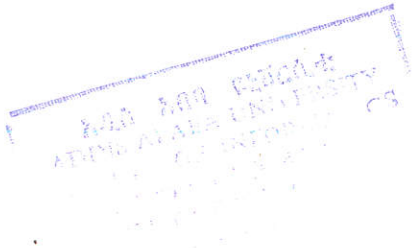
ABSTRACT

Financial institutions in a nation play a crucial role in the development of its economy. The banking sector as one type of financial institution is indisputably the new frontier of economic development in a country. In this respect, banking has to be sound and safe for its customers as well as for the stability of the currency and economy of a country. One factor that affects the well functioning of the banking sector is credit risk. This factor is also a general problem among commercial banks in Ethiopia.

In order to deal with high default rates banks in other countries are making use of data mining. The possible application of data mining in the commercial banking sector of Ethiopia has also been tested by the use of neural network technique. As credit risk is a risk type that bank managers give more emphasis in the loan disbursement process because it is one of the major reasons that cause a bank to fail, the study of the possible application of data mining needed further investigation. To this end, the present study focuses on the application of data mining to support credit risk assessment taking as a case study Nib International Bank S.C.(NIB). In doing so the aim of this research was to assess the potential applicability of decision tree technique to help in the loan disbursement decision-making process of banks.

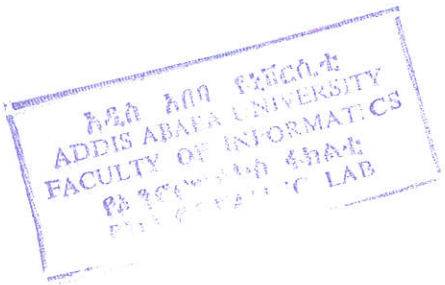
The methodology used for this research had three basic steps. These were collecting of data, data preparation, and model building and testing. The required data was selected and extracted from Nib International Bank records. Then, data preparation tasks (such as data transformation, deriving of new fields, and handling of missing variables) were undertaken. Decision tree data mining technique was employed to build and test models.

Several decision tree models were built and tested for their classification accuracy and the model with encouraging results was taken to generate rules to support credit decision makers and the procedures adopted are described in this document.



The performance of the developed model is validated using new datasets and its predictive accuracy is also tested. The result shows that the use of decision tree technique produces rules for justifiable credit decision-making and that it is the best technique that needs to be adopted for NIB bank as it presents a means of providing explanation for proposed decisions as compared to neural network techniques.

All things considered, the existence of an electronic system to support the credit risk assessment of NIB bank will promote the services of the bank to its customers as well as minimize risk.



CHAPTER ONE

INTRODUCTION

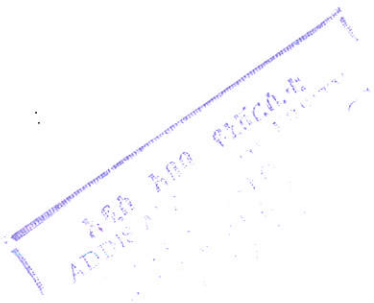
1.1 Background

Of the many sectors that are said to have a great contribution to the development of a country, financial institution is one of the leading ones. Financial institutions in a community can form the core of economic development. The banking sector as one type of financial institution is indisputably the new frontier of economic development in a country. In this respect, banking has to be sound and safe for its customers as well as for the stability of the currency and economy of a country.

Among the many factors that affect the well functioning of the banking sector, one is the risk associated with a trading partner not fulfilling his obligations in full on due date or at any time thereafter [IFCI, 2000].

Risks are classified into various categories. The most prominent financial risks, which banks are exposed to are:

- Credit risk;
- Interest rate risks;
- Foreign exchange / currency risk;
- Liquidity risk; and
- Contingency risks [Kannan, 2003].



Among the above risks that financial institutions face, credit risk is the one to which decision makers of financial institutions pay the closest attention because it has been the risk most likely to cause a bank to fail.

In this respect, credit risk assessment is a key component in the process of commercial lending. A potential borrower's credit worthiness assessment determines whether the borrower will ultimately be granted credit, and if so at what cost should be the terms of underwriting fees and interest levels.

As Banks move into a new high-powered world of financial operations and trading, with new risks, the need is felt for more sophisticated and versatile instruments for risk assessment, monitoring and controlling risk exposures [Bhatnagar, 2001]. It is, therefore, time that bank management equip themselves with systems capable of assessing, monitoring and controlling risk exposures in a more scientific manner.

Banks have historically gathered huge amounts of information on their customers. But until recently, that data was rarely seen as more than a single purpose static resource. That's because much of the data was segregated in scores of separate database in different bank departments. Information that consumers provided when they applied for automobile loans went into one group of database, mortgages into another, and checking and savings transactions in others. All remained isolated.

Because the database weren't linked, there was no easy way to get an overview of each customer's finances and bank service usage. As a result, there was no simple way to identify something as fundamental as which customers were profitable to the bank and which were not.

The new generation of computerized methods is helping the endeavor of interrogating and analyzing very large datasets automatically and efficiently, thereby extracting useful information and knowledge that are valuable for decision-making.

This method is known as "Data Mining", which stands for the extraction of previously unknown, yet valid, and actionable information from large databases and then making use of this information to make critical business decisions [Mitchell, 1997].

Data mining allows for coping with today's large business and planning problems by taking advantage of new hardware and software technologies, and using scalable algorithms to sift through a large amount of data and extract useful and valid information relatively efficiently and inexpensively. It can discover relation and can also automatically find out data patterns and relationships which could not have been discovered by a human researcher, because these are beyond human experience, education and limitations of imagination [Luan, 2002]. In addition, data mining tools can automatically produce complex, database abstracts to help users forecasting and classification [Gnardellis,n.d.].

As defined by Lloyd-Williams [1997] Data mining or Knowledge Discovery in Databases (KDD) as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. This encompasses a number of different technical approaches, such as clustering, data summarization, learning classification rules, finding dependency networks, analyzing changes, and detecting anomalies [Two Crows Corporation, 1999].

Data mining that is being used both to increase revenues and to reduce costs is already reputed among financial institutions to be fundamental to business bottom lines. Banking executives can predict whether a prospective customer will default his credit or not.

According to Berson [1999], there are many data mining techniques that are being used today. These methods are divided into two categories classical techniques (which include statistics; neighborhoods and clustering) and next generation techniques (that include trees; networks and rules).

According to Thearling [2003], the most commonly used data mining techniques are: Decision tree, neural networks, genetic algorithms, nearest neighbour method and rule induction.

Decision tree: in this technique records are presented in a tree like structure that represents sets of decisions. Decision trees generate rules for classification of a dataset. A depth discussion on decision tree is presented in chapter two.

Artificial neural networks: are simulations of biological neural networks that learn through training which are more suitable to model non-linear and complex relationships.

Genetic algorithm: is an optimization data mining technique which often adopts processes such as genetic combination and natural selection in order to find out set of parameters that best explain a predictive function [Thearling, 2003]. Genetic algorithms employ the selection, crossover, and mutation operators to evolve successive generation of solutions [Berry and Linoff, 1997].

Nearest neighbour method: is a technique that classifies each instance in a database based on the classes of the k records which are most similar to it in a historical dataset [Thearling, 2003]. On the basis of the number of records (k) it is often known as k-nearest neighbour. This

technique is an instance-based learning in which a distance function is used to determine the class of a new instance. Although this technique is simple it often works very well [Witten and Frank, 2000].

Rule induction: is a technique that extracts useful sequence of 'if-then' rules from a database based on statistical significance [Thearling, 2003]. The rule induction technique concentrates on the criterion and generation of rules with optimal accuracy.

In general most of the above methods tend to be more computationally intensive than decision tree technique, which is chosen for this present research. Among the above, decision trees are powerful and widely used data mining method for classification and prediction [Berry and Linoff, 1997], they are said to be of great use in the process of credit risk assessment [Brand & Gerritsen, 1998]. The strength and popularity of decision tree is due to the fact that in contrast to neural networks, it expresses the 'if-then' rules explicitly.

Decision tree helps when a businessperson needs to make a decision based on several factors. It can help identify which factors to consider and how each factor has historically been associated with different outcomes of the decision. In applications where the accuracy of a classification or prediction of unknown instances is the only thing that matters,(i.e. in cases where how or why the model works is not important), both neural networks and decision tree can perform a good job [Witten and Frank, 2000]. But for example if a bank firm could make use of data mining application to appraise credit application it is more acceptable to both the bank official as well as the credit applicant to know that the application is rejected on the basis of a computer-generated rules than to declare that the decision has been made by a neural

networks which provides no explanation for its action. Consequently the present research makes use of decision tree techniques.

Coming to our specific situation, in Ethiopia the financial sector is slowly changing as the currently existing six private banks are aggressively grabbing the market. The six banks together had a market share of close to 20% in 2003 [Timewell, 2003]. These Banks are nowadays gaining more customers through better service and could expect to reach 30% market share by the end of 2004 and 40 % in two year's time [Lulseged, 2003].

One factor that could cut short the expectation of these banks on the return expected is risk associated with loans. In order to minimize the risk banks are making use of many traditional methods. But nowadays, there are more effective methods that help to assess, monitor and control risk exposures in a more scientific manner.

Data mining technologies and particularly decision trees have been used effectively in developed countries for credit risk assessment. In Ethiopia, Askale [2001] had conducted a research to support loan disbursement at one of the commercial banks and by doing so has tried to show the potential application of data mining and has shown its effectiveness. Taking into consideration the research done by Askale (ibid) using neural net, this research will try to emphasize on the usefulness of such a state-of-the-art technology in the Ethiopian commercial banking context.

1.2 Statement of the Problem

Time and again, the fundamental business of lending has brought trouble to individual banks and entire banking systems [Bhatnagar, 2001]. It is therefore, imperative that the banks have adequate systems for credit assessment of individual projects and for evaluating risks associated therewith as well as the industry as a whole.

Consequently, bank managers should equip themselves fully to grapple with the demands of creating tools and systems capable of assessing, monitoring and controlling risk exposures in a more reliable manner. In this regard, data mining, and particularly decision trees help identify which factors to consider and how each factor has historically been associated with different outcomes of the decision process. A decision tree creates a model as either a graphical tree or a set of text rules that can predict (classify) each applicant as a good or bad credit risk [Brand & Gerritsen, 1998].

A number of researches have been done on the possible application of data mining techniques in the Ethiopian context. The first attempt was made by Gobena [1999] that was on the application of data mining technology and techniques in the Ethiopian Airlines and this work was extended by Henock [2002] and Denekew [2003]. Moreover, Shegaw [2002] has also assessed the potential applicability of data mining technology in the Ethiopian context with particular reference to the health sector. Tesfaye [2002] and Askale [2001] also conducted other researches on the application of data mining in the financial industry specifically at the Ethiopian Insurance company and the Dashen Bank respectively.

Among those studies, one that is more related to this research is the one conducted by Askale (ibid), who had tried to investigating the potential application of data mining technique to assess credit risk at Dashen Bank (a private commercial bank), and while doing so has used neural network as her data mining technique. Askale (ibid) had concluded that the results obtained were encouraging and attested the potential of data mining application in the banking sector for credit decision-making. The researcher has made many recommendations, among which some are: making the required data available in a computerized form, devote sufficient time for building and training an appropriate model, include as many relevant variables as possible, introduce more detailed classification of borrower's category, and studying the applicability of other data mining techniques.

The present research was conducted to supplement the research done by Askale [2001]. The researcher took into consideration the recommendations given by Askale (ibid) while using a different technique (i.e. decision tree) than neural network, which had been chosen by her.

The research was conducted in a form of case study in Nib International Bank S.C. (NIB). The unavailability of the data on which Askale's experiment was based on in addition to NIB management's friendliness and responsiveness to the researcher's enquiries to get access to and collect the required data for this study encouraged the researcher to conduct this research at NIB.

In the present experimental research undertaking, decision tree technique was chosen as the instrument for the classification activity. The technique was used in assessing applicability of data mining in helping determine credit worthiness of prospective borrowers.

1.3 Objectives

1.3.1 General Objective

The general objective of this research is to investigate the application of data mining techniques particularly the application of decision tree classifiers in credit risk assessment for the purposes of supporting decisions made on loan approval.

1.3.2 Specific Objectives

The specific objectives of the research are:

- Develop an understanding of the application domain;
- Review literature on data mining at large and the application of decision trees technique in particular;
- Collect data on which the mining process will be conducted;
- Prepare the data for model building by selecting, cleaning, and integrating it;
- Select the data mining software to be used that support Decision Tree technique;
- Build and train a computer model;
- Evaluate (test) model;
- Make a comparison with the findings of that of Askale's;
- Make recommendation based on findings.

1.4 METHODOLOGY ADOPTED

In order to build a good data-mining model for credit risk assessment, there are a number of steps that one must follow. For this research the following steps are adopted:

1.4.1 Review of Related Literature

A review of relevant literatures has been conducted to assess data mining technology in general and decision tree method and its application for credit risk assessment in particular. In this, the concepts, techniques, and researches done in the field are reviewed. Diverse books, journals, magazines, articles and the Internet pertaining to the subject matter of data mining and Knowledge Discovery in Databases (KDD) were reviewed in order to have a grasp of the potential applicability of data mining in credit risk assessment, particularly in predicting loan applicants defaulting in the banking sector.

1.4.2 Study of the business problem

In order to define the research problem properly, primary data was collected by interviewing domain experts at Nib bank as well as through observation. Then based on the information obtained from these attempts, the overall credit approval process of Nib International bank was described.

A number of fact finding methods i.e. interviews and document reviews were made use for analyzing and solving the business problem. The researcher has followed the following steps in order to develop a model, employing data mining technique.

1.4.3 Development and testing of the model

1.4.3.1 Identifying Available Data Sources

The data sources utilized in the research undertaking were documents available in manual format. These documents are forms (described in chapter 3) that customers' fill out when

requesting loans and documents utilized by the bank to track the performance of previously granted loans.

1.4.3.2 Data Collection and Preparation for Analysis

Before subjecting the raw applicants' data to analysis, it needed to be converted into a form suitable for analysis. Data preparation activities, such as editing, coding, data cleaning (consistency checks and missing responses) and statistically adjusting the data were conducted on the data collected. After doing this, a thorough check of the database was made to eliminate duplicate records and cases with the help of domain experts from the Bank.

1.4.3.3 Build and Train the Computer Model

After the data has been cleaned and formatted, the data was analysed by the use of KnowledgeSTUDIO software (used for model building) selected for this research, to meet the study's objectives. In general terms, this is the step where most of the work of creating a model was done. The training set was used to generate an explanation of the dependent (target) variable in terms of the independent (input) variables. This explanation has taken the form of a decision tree and is thoroughly explained in chapter four.

1.4.3.4 Evaluating (Testing) the Model

In order to be confident in the estimation of the model's performance, the model was applied to the final collection of reserved pre-classified records: the evaluation set. The error rate on the evaluation set was a good predictor of the error rate on unseen data. Here the domain experts from the organization played a big role in evaluating the model's performance.

1.4.3.5 Prototype Development

An attempt has been made to develop a prototype to support the credit risk assessment by using the rules generated from the decision tree model selected. The prototype is built by the use of Visual Basic 6.0 programming language.

1.5 Scope and Limitation

The scope of the present experimental research undertaking is strictly limited to assessing the possible application of data mining technology at Nib International Bank S.C, it is limited to risks associated with term loans to examine the potentials of data mining techniques in developing classification model by the use of decision tree technique in support of credit risk assessment.

The following are some of the limitations noticed during this research undertaking. First the researcher had originally planned to conduct a case study at the Dashen Bank S.C. However, for reasons of confidentiality, access to the source data was found difficult. So, a different bank i.e. Nib International Bank S.C. (NIB) was chosen to conduct the present experiment. The management's friendliness and responsiveness to the researcher's enquiries to get access to and collect the required data for this study encouraged the researcher to conduct this research at NIB. In addition to the above limitation, getting access to resources such as literature related to the study, and alternative tools for decision trees modeling was another limitation mainly due to price and ease of access.

1.6 Research Contribution

In this research an attempt was made in general to find out the applicability of data mining technology in the banking industry and in particular to assess the use of decision tree techniques in credit risk assessment. The result of the study shall be used as an input for the development of full-fledged data mining application using decision trees in supporting loan disbursement activity. This new emphasis is believed by the researcher to expand the scope of credit risk assessment and is aimed at facilitating the decision making process of bank officials. Although the study was aimed at addressing banking problems in particular, the output of the study may be used as a source of methodological approach for studies dealing with the application of data mining technology on similar problem areas.

1.7 Thesis organization

This thesis report consists of five chapters. The first chapter deals with the general overview of the study including background, statement of the problem, objectives and methodology of the research. The second and third chapters are devoted to literature review of data mining technology, decision tree technique and credit risk assessment at Nib International Bank S.C. respectively.

Chapter four reports the experiment of the research. It comprises training; building and validation of the models while results of the experiment are also analyzed and interpreted. The last chapter presents conclusions and recommendations.

CHAPTER TWO

DATA MINING TECHNOLOGY

In this chapter, an effort has been made to review the literature on the concepts and techniques of data mining in discovering knowledge from large databases as well as its application in the financial sector. It is also aimed at providing a base for the experiment undertaken.

2.1 Overview

Up until recently, the ability to analyse and understand a huge volume of data lagged far behind the capability to gather, store and manipulate data. The data that are generated and stored routinely grow into large databases amounting to giga (and even tera) bytes [Deogun et.al., 1999]. In addition, the amount of information in the world doubles every 20 months at a rate of 100% [Witten & Frank, 2000].

Organizations keep record of data generated even long after their life time has expired because they believe that there is valuable information implicitly coded within it [Fayyad, 1996]. In earlier days the data that was being collected was relatively easy to analyse manually but as data collections continues to grow in size and complexity, there is a growing need for more sophisticated techniques of analysis. One such technique is data mining which is defined as being a new generation of computerized methods for "extracting previously unknown, valid, and actionable information from large databases and then using this information to make critical business decisions" [Cabena et.al., 1998] as quoted by [Levin & Zahavi, 1999].

Data mining is considered as one of the most important frontiers in databases systems and one of the most promising interdisciplinary developments in the information industry [Han &

Kamber, 2001]. This is due to the fact that interesting knowledge, regularities or high-level information can be extracted from databases and viewed or browsed from different angles. The discovered knowledge can be applied to decision-making, process control, information management, and query processing.

Data mining, which is sometimes referred to as knowledge discovery in databases, is defined as "the process of nontrivial extraction of implicit, previously unknown and potentially useful information (such as knowledge rules , constraints, regularities) from data in databases".[G.Piatetsky-Shapiro and W.J.Frawley, 1991] as quoted by [Chen, et.al, 1997].It is an interdisciplinary approach involving tools and models from statistics, artificial intelligence, pattern recognition, heuristics, data acquisition, data visualization, optimization, information retrieval, high end computing and others [Levin & Zahavi, 1999, Han & Kamber, 2001, Deogun, et.al., 2001].

But it is important to note that data mining does not replace skilled business analysts or managers, but rather gives them a powerful new tool to improve the job they are doing. Any company that knows its business and its customers is already aware of many important, high-payoff patterns that its employees have observed over the years. What data mining can do is confirm such empirical observations and find new, subtle patterns that yield steady incremental improvement (plus the occasional breakthrough insight) [Two Crows Corporation, 1999].

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions.

2.2 DATA MINING AND KNOWLEDGE DISCOVERY IN DATABASES

There is some misunderstanding among researchers and the intellectual community in the field of data mining about the terms Data Mining and Knowledge Discovery in Databases (KDD). Many people treat data mining as a synonym for KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases. According to (Fayyad, Piatetsky-Shapiro & Smyth, 1996), KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. According to these authors Data mining is the application of specific algorithms for extracting patterns from data. Nevertheless, many authors [Carbone, 1997; Piatetsky-Shapiro, 2000; Han and Kamber, 2001] believe that the term data mining has become more popular in industries, in media and the database researches as a synonym for knowledge discovery and hence the two terms are used interchangeably. This view is also applicable throughout this study.

The phrase Knowledge Discovery in Databases was coined at the first KDD workshop in 1989 to emphasize that knowledge is the end product of a data-driven discovery. It has been popularised in the Artificial Intelligence (AI) and machine learning fields [Piatetsky-Shapiro ,1991], quoted in [Fayyad, Piatetsky-Shapiro & Smyth, 1996]. KDD has evolved, and continues to evolve, from the intersection of research fields such as machine learning, pattern recognition, databases, statistics, AI, knowledge acquisition for expert systems, data visualization, and high-performance computing. The unifying goal is extracting high-level knowledge from low-level data in the context of large datasets.

Fayyad [1996] while emphasizing on the importance of the data mining step in the KDD process state that the other steps involved in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data. But that blind application of data-mining methods (rightly criticized as data dredging in the statistical literature) can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns.

KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [Fayyad, Piatetsky-Shapiro, and Smyth 1996], using KDD interesting knowledge, regularities, or high-level information can be extracted from the relevant sets of data in databases and can be investigated from different angles, and in large databases, and can also serve as rich and reliable sources for knowledge generation and verification [Chen, et.al., 1997].

KDD is an iterative sequence of steps, which are data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and Knowledge presentation [Han & Kamber, 2001]. The Figure 1 below shows the various steps involved in the KDD process as presented by [Fayyad et.al., 1996].

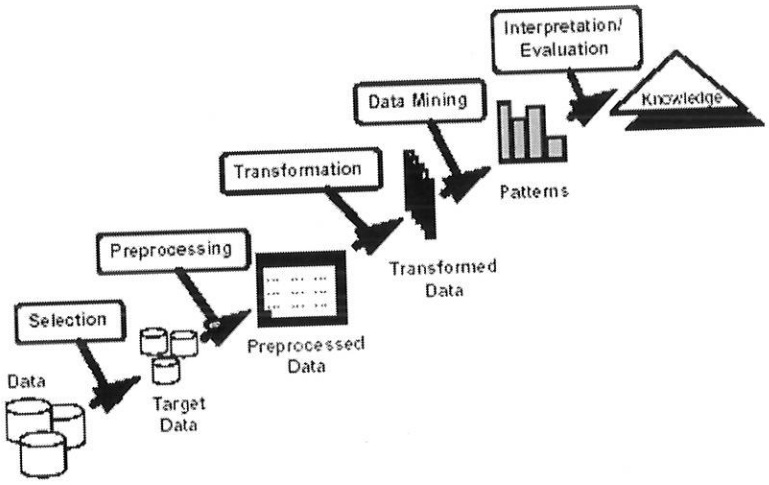


Figure 1: Steps in the Knowledge Discovery in Databases (KDD) process [Fayyad, Piatetsky-Shapiro and Smyth, 1996]

The process of building and implementing a data mining solution referred to as Knowledge discovery in databases (KDD) as described by Brand & Gerritsen [1998], Levin & Zahavi [1999], CRISP-DM [2000] and Two Crows Corporation [1999] is explained as follows.

DEFINING THE BUSINESS PROBLEM

Any data mining process should start with a clear definition of the business problem involved and the objective function, as this may direct not only the KDD process but also the data mining modeling involved. As stated by Two Crows Corporation [1999] the prerequisite to knowledge discovery is understanding ones data and the function one wants it to serve. Without this understanding, no algorithm, regardless of sophistication, is going to provide one with a result in which they should have confidence. So, to make the best use of data mining one must make a clear statement of the objective. This statement of the problem must include a way of measuring the results of the knowledge discovery project. It may also include a cost justification.

SELECTING THE TARGET DATA SET FOR ANALYSIS

The selection process is an important step in the KDD process because databases are heterogeneous, containing a wide variety of data, not all of which may be appropriate for the analysis at hand. So, one needs to extract the target data to analyse it in a way that is consistent with the problem to be solved and the objective of the project. Levin and Zahavi [1999] state that one can use subjective judgment to extract the relevant target set, but that in many other cases, one may have to use segmentation analysis, which may require the use of a data mining model, such as clustering, to extract the target dataset to participate in the data mining process.

COLLECTING, CLEANING, AND PREPARING DATA

The second step in a KDD process is collecting, cleaning and preparing data. In this step the data obtained from various sources have to be joined to create homogeneous source by resolving representation and encoding difference. After they are joined the data obtained has to be checked to resolve data conflicts, outliers, missing data, and ambiguity. On these checked data conversions and combinations are to be used to generate new data fields such as ratios or rolled-up summaries. These steps require considerable effort, often as much as 70 percent or more of the total data mining effort [Brand & Gerritsen, 1998; Berry and Linoff, 2000].

TRANSFORMATION

The other important step in data mining process is transformation of the data. According to Brand and Gerritsen [1998] rather than in the raw data, the prediction power of data resides in transformation of the data.

Data transformations are designed to account for non-linear relationships between the dependent variable and one or more independent variables (assuming all the others are constant), identifying pair-wise interaction, perhaps even higher-order interactions, between independent variables, tracking seasonal and time related effects. Data transformation is even transforming data to make them compatible with the theoretical assumptions underlying the model involved.

DATA MINING/ MODEL BUILDING

Data mining also referred to as the model-building step involves selecting data mining tools, transforming data if the tool requires it, generating samples (as necessary) for training, testing and validating the model and, finally, using the tools to build, test and select models [Levin & Zahavi, 1999].

Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patters (or models) over the data. So, at this stage of the process models and tools are selected to interrogate the data and convert it into knowledge for decision-making. Most data mining methods are based on tried and tested techniques from machine learning, pattern recognition, and statistics which are: classification, clustering, regression, and so on [Fayyad, Piatetsky-Shapiro & smyth, 1996]. These models can be selected from a wide range of models to suit the business issue concerned.

VALIDATING THE MODELS

In the validation phase of the KDD process, the model is tested for accuracy on an independent dataset, one that has not been used to create the model. In addition assessing

the sensitivity of a model as well as pilot testing the model for usability should also be performed at this stage.

DEPLOYING THE MODEL

Deployment may require building computerized systems that capture the appropriate data and generate a prediction in real time so that a decision maker can apply the prediction.
Once a data-mining model is built and validated, it can be used in one of the two main ways. The first way is for an analyst to recommend actions based on simply viewing the model and its results while the second way is to apply the model to different data sets.

When using the KDD process to achieve objectives in service giving enterprise, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases it is the user, not the data analyst, who carries out the deployment steps. In doing so, even if the analyst will not carry out the deployment effort it is important for the user to understand up front what actions need to be carried out in order to actually make use of the created models.

MONITORING

In this step what we have to consider is the fact that there are always changes in human needs. So models that were perfect to the need of an organization yesterday may no longer be regarded as perfect today. So, monitoring models requires constant revalidation of the model on new data to assess if the model is still appropriate.

2.3 DATA MINING AND DATA WAREHOUSING

Data warehouses are used extensively in banking and financial services, consumer goods and retail distribution sectors, and controlled manufacturing, such as demand-based production [Han & Kamber, 2001].

The concept of data warehouse was motivated by the need to view the entire enterprise from a single point of view instead of a collection of narrowly defied "silos".[Berry and Linoff, 2000]. It is a decision support database that is maintained separately from the organization's operational database, and it supports information processing by providing a solid platform of consolidated, historical data for analysis.

Data warehouse is defined as "a subject-oriented, integrated, time-variant, and non-volatile collection of data, which supports decision making process in enterprise management. [W.H.Inmon,1996] as quoted in [Han & Kamber, 2001].

Frequently, the data to be mined is first extracted from an enterprise data warehouse into a data mining database or data mart (a subset of corporate-wide data that is of value to a specific groups of users whose scope is confined to specific, selected groups, such as marketing data mart). There is some real benefit if the organization's data is already part of a data warehouse [Two Crows Corporation, 1999]. But a data warehouse is not a requirement for data mining.

The data-mining database may be a logical rather than a physical subset of your data warehouse, provided that the data warehouse DBMS (Database Management System) can support the additional resource demands of data mining.

2.4 DATA MINING ACTIVITIES

The two high-level primary goals of data mining in practice tend to be prediction and description [Fayyad, Piatestsky-Shapiro & Smyth, 1996]. Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest, while description focuses on finding human-interpretable patterns describing the data.

There are many data mining activities that involve extracting meaningful new information from the data. These activities are: classification, estimation, prediction, affinity grouping or association rules, clustering and description and visualization [Berry and Linoff, 2000].

In the following part a brief explanation of the stated activities will be given, of which a more thorough description is given in the work of Berry and Linoff [1997]; Two Cross Corporation [1999]; Han and Kamber [2000]; and Levin and Zahavi [1999].

2.4.1 Classification

Classification is one activity of data mining that is being put into use in various business environments and as its name implies predicts class membership. It is in fact a process of finding a set of models that describe and distinguish data classes or concepts, so as to be able to use the model to predict the class of objects whose class label is known. In doing so, it examines the feature of a newly presented object and assigns to it a predefined class. The act of classification consists of updating each record by filling in a field with a class code.

In constructing any classification model the task is to build a model that can be applied to unclassified data in order to classify it, but it should be noted that it is important that data be available for all possible outcomes so the model can learn about all cases.

According to Han & Kamber [2001], classification is a two way process. In the first step, a model is build and in the second the model is used for classification. The model is built by describing a predetermined set of data classes or concepts and by analyzing the database tuples described by attributes. This step is known as supervised learning since the class label of each training sample is known. After it has been build the model is represented in the form of classification rules, decision trees, or mathematical formulae where the rules can be used to categorize future data samples, as well as provide a better understanding of the database content. And when the model is used for classification the predictive accuracy of the model (or classifier) can be estimated. Here, the accuracy of a model on a given test set is the percentage of test set samples that are correctly classified by the model. For each test sample, the known class label is compared with the learned models class prediction for that sample. If the accuracy of the model is considered acceptable, the model can be used to classify future data tuples or objects for which the class label is not known.

There are various data mining techniques for classification and regression. Four are most commonly used commercially. These techniques are decision trees, neural networks, naïve bayes and k-nearest neighbour [Brand & Gerritsen, 1998]. The present research is concerned with classification by the use of decision trees which will be discussed in greater detail in section 2.6 of this chapter.

2.4.2 Estimation

Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variable. It also deals with rules that explain how to estimate a value given target class; the task is often to make a numerical prediction.

2.4.3 Prediction

Any prediction can be thought of as classification or estimation. The difference is one of emphasis. In prediction, historical data is used to build a (predictive) model that explains the current observed behaviour.

2.4.4 Affinity grouping or association rules

The task of affinity grouping is to determine which things go together. A good example would be to determine what things go together in a shopping cart at the supermarket, referred to as *market basket analysis.*

2.4.5 Clustering

Clustering is the task of segmenting a diverse group into a number of more similar subgroups or clusters. Although like classification, clustering is the organization of data into classes, what distinguishes clustering from classification is that clustering does not rely on predefined classes.

There are many clustering approaches: intra class similarity, which maximizes the similarity between objects in the same class and inter-class similarity, which minimizes the similarity between objects of different classes.

2.4.6 Description and visualization

Sometimes the purpose of data mining is simply to describe what is going on in a complicated database. A good enough description of behaviours will often suggest an explanation for it as well. At the very least, a good description suggests where to start looking for an explanation. Data visualization is a powerful form of descriptive data mining. It is not always easy to come up with meaningful visualizations, but the right picture really can be worth a thousand association rules since human beings are extremely practiced at extracting meaning from visual scenes.

2.5 DATA MINING APPLICATIONS

Nowadays data mining is increasingly popular because of the substantial contribution it can make to service giving organizations. It has been used in a wide variety of business activities as it offers value across a broad spectrum of industries. Two Crows Corporation [1999] mentioned telecommunications and credit card companies as the two leaders in applying data mining to detect fraudulent use of their services. In addition, the authors state that insurance companies and stock exchanges are also interested in applying data mining to reduce fraud. With the objective of developing a predictive model in support of insurance risk assessment, Tesfaye (2002) has applied the data mining technology and developed a prototype, named MIRS (Motor Insurance Renewal System).

Furthermore, they mention that data mining can be found fruitful in medical applications where it can be used to predict the effectiveness of surgical procedures, medical tests or medications. In this regard, Shegaw [2002], having the objective of developing a model that can support in

preventing and controlling child mortality at the district of Butajira, has applied data mining technology and reported an achievement in the development of the model.

In addition to the areas mentioned above data mining can also contribute to companies active in the financial markets to determine market and industry characteristics as well as to predict individual company and stock performance. Credit card companies can leverage their vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product or fraudulent credit card use, determine credit card spending by customer groups, find hidden correlations between different financial indicators, and identify stock trading rules from historical market data. With the objective of developing a model that can support the loan decision-making process at Dashen Bank S.C., Askale [2001] has investigated the potential applicability of data mining technology in the banking sector.

On the other hand, pharmaceutical firms are mining large databases of chemical compounds and of genetic material to discover substances that might be candidates for development as agents for the treatments of disease, as well as analysing the recent sales of the company [Berry and Linoff, 2000; Two Crows Corporation, 1999]. Diversified Transportation Company with a large direct sales force can apply data mining to identify the best prospects for its services, by using data mining to analyse its own customer experience, this company can build a unique segmentation identifying the attributes of high-value prospects. A large consumer package goods company can apply data mining to improve its sales process to retailers.

All things considered, Data mining plays a leading role in every facet of customer relationship management. As stated by Berry and Linoff [2000] it is only through the application of data mining techniques that a large enterprise can hope to turn the myriad records in its customer databases into some sort of coherent picture of the potential of its customers [Berry and Linoff, 2000].

2.5.1 Data Mining Application in the banking Sector

According to Wasserman [2000], the most common application of data mining in the financial sector that has been around for a long time (since the 1950s) is credit scoring. Credit scoring is a statistical method used to predict the probability that a loan applicant or existing borrower will default or become delinquent. Wasserman (ibid) further noted that even though credit scoring is widely used for consumer lending, particularly with credit cards and mortgage loans, and is becoming more common in small business lending, for larger firms, in industries that naturally accumulate large amounts of detailed transaction data, such as firms in banking, insurance, telecommunications, catalogue retail, utilities, and supermarkets, applications of data mining are increasingly widespread.

Data mining is currently used extensively in several retail markets as it is very effective at providing information used for customer profiling and behavior. One such area is banking and finance [Ballenger et.al.,1999].

Some banks are currently testing data mining tools to manage their credit portfolios more efficiently. Fabris [1998] states that data mining holds great promise in assessing the risk of a bank's entire portfolio of loans. The authors further attested that by analyzing customer behaviours such as payment habits, data mining could provide answers to vital questions such

as: What percentage of loans will be refinanced next quarter? What percentage will go to foreclosure? And what percentage will be in serious felonious status? Accurate answers to these questions allow credit risk managers to allocate optimal loan loss reserves--funds set aside to cover bad loans--which is important to profitability.

Ballenger [1999] on his part states that some of the possible applications of data mining are: mortgage approval, loan underwriting, fraud analysis and detection.

Many banks nowadays are making more and more use of data mining in their day-to-day activities, however many companies will not admit to their techniques, due to policies set because of the competitive market. But there are few cases where banks discuss their experiences [Ballenger, 1999].

Among banks that will admit to using data mining some are: Bank of America, First USA, FCC National Bank, Federal Home Loan Mortgage, Chevy Chase Bank, U.S. Bancorp, USAA Federal Savings Bank, but many others have policies not to discuss it [Ballenger, 1999].

For example, Bank of America built a data-mining model to predict attrition of small business customers. One of the key factors was the length of time small businesses held accounts with the bank. That indicator proved to be misleading, however, because about 60 percent of small businesses go bankrupt within three years. Bank of America's model was a better indicator of companies headed for bankruptcy rather than those headed to a rival bank. The bank subsequently revised the model to eliminate that factor.

Another example, Mellon Bank in Pittsburgh, PA, tracks credit card usage and broad categories of purchases as a good indicator of who may switch from a bank credit card to another [Ballenger, 1999].

[Fabris, 1998] Bank of Montreal, yet another bank willing to share its experiences has reported having analyzed mortgage customers' transactions in checking, savings and other accounts for insight into who is at risk of defaulting. The bank was surprised to find that some customers who consistently made their mortgage payments late were not necessarily at a high risk of defaulting. The bank found that a certain type of customer is in the habit of paying bills late but has the means to fulfil his or her obligations. By further analyzing the transactional behaviour of customers across all their accounts, the bank can see which customers experience periodic cash flow crunches and which may truly be in danger of defaulting.

All things considered, in addition to helping increase the value of customer relationships, mining customer information databases aids banks in managing risk.

2.6 Overview of Data Mining Technique

According to Berry and Linoff [1997], learning and understanding of different data mining techniques is essential in order to take the advantage of specific technique, to determine the best applicable technique for the problem at hand and to know the advantages and disadvantages of a technique.

It is evident that no one technique is applicable to all data mining problems. To determine the best technique suitable to the specific problem, familiarity with the available techniques is necessary. According to Thearling [2003], the most commonly used data mining techniques are: decision tree, neural networks, genetic algorithms, nearest neighbour method and rule

induction which are described in chapter one of the present research. As the present research deals with decision tree the following section deals with issues of decision tree.

2.6.1 Decision tree

One of the most commonly used data mining technique for classification tasks is a decision tree. A decision tree is described as being a model that is both predictive and descriptive. It is a way of representing a series of rules that lead to a class or value. It is a flow-chart like tree structure, where each internal node denotes a test on attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distribution [Han & Kamber, 2001].

In any decision tree the top most node is called the root node. A decision tree grows from the root node, so you can think of the tree as growing upside down, splitting the data at each level to form new nodes. The resulting tree comprises of many nodes connected by branches. Nodes that are at the end of branches are called leaf nodes and play a special role when the tree is used for prediction.

A typical example of a decision tree is shown in figure 2.2 below.

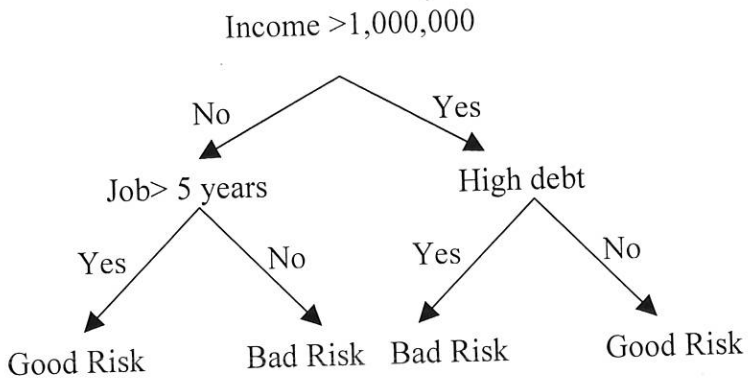


Figure 2: A Simple Decision Tree

Each node in the tree specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute. Once grown, a tree can be used for predicting a new case by starting at the root (top) of the tree and following a path down the branches until a leaf node is encountered. The path is determined by imposing the split rules on the values of the independent variables in the new instance. For example in the above decision tree an individual with "income> 1,000,000" and "High Debt" would be classified as a "Bad Risk" whereas an individual with "Income<1,000,000" and "Job>5 Years" would be classified as a "Good Risk".

Decision tree models are commonly used in Data mining to examine the data and induce the tree and its rules that will be used to make predictions. In this respect, there are many algorithms that are being used for decision tree construction like CHAID (Chi-squared Automatic Interaction Detection), C4.5/C5.0 , CART (classification and Regression) and many others which are less familiar [Berry & Linoff, 2000]. These algorithms produce trees that differ from one another in: the number of splits allowed at each level of the tree, how those splits are chosen when the tree is built, and how the tree growth is limited to prevent over fitting. But nowadays data mining software tools typically allow the user to choose among several splitting criteria and pruning rules, and to control parameters such as minimum node size and maximum tree depth allowing one to approximate any of these algorithms.

According to Two Crows Corporation [1999], there are two main types of decision trees: classification trees, and regression trees. Classification trees are used to predict categorical variables that label records and assign them to the proper class that can also provide the confidence that the classification is correct. Regression trees on the other hand are used to

predict continuous variables that estimate the value of a target variable that takes on numeric values.

Decision tree learning has therefore been applied to problems such as learning to classify medical patients by their disease, equipment malfunctions by their cause and loan applicants by their likelihood of defaulting on payments [Brand and Gerritson, 1998]. They are also said to be a good choice when the data–mining task is classification of records or prediction of outcomes.

The training process that creates the decision tree is usually called induction. In the next section we will try to look into the process of decision tree induction so as to have a good view of the process.

2.6.1.1 Decision tree Induction

The decision tree Induction process starts with a training set consisting of known classes. And the ultimate goal of the process is to build a tree that distinguishes among the classes. But, building a decision tree is not a one time task, it is an iterative process known as recursive partitioning which implies that decision trees are build by splitting the data up into partitions and then splitting it up some more [Berson, et.al., 1999].

In constructing decision trees most algorithms go through two phases: a tree-growing (splitting) phase followed by a pruning phase [Brand & Gerritsen, 1998]. The tree-growing phase is an iterative process, which involves splitting the data into progressively smaller subsets. Each iteration considers the data in only one node. The first iteration considers the root node that

contains all the data, subsequent iterations work on derivative nodes that will contain subsets of the data.

One important characteristic of the tree splitting algorithm is that it is a greedy algorithm. Greedy algorithms make decisions locally rather than globally. When deciding on a split at a particular node, a greedy algorithm does not look forward in the tree to see if another decision would produce a better overall result. What it does is once a node is split; the same process is performed on the new nodes, each of which contains a subset of the data in the parent node. The variables are analysed and the best split is chosen. This process is repeated until only nodes where no splits should be made remain.

Initially, all of the records in the training set- the preclassified records that are used to determine the structure of the tree – are together in one big box. Then the algorithm tries breaking up the data, using every possible binary split on every field. The algorithm chooses the split that partitions the data into two parts that are purer than the original. This splitting or partitioning procedure is then applied to each of the new boxes. The process continues until no more useful splits can be found. When no split can be found that significantly decreases the diversity of a give node, then it is a leaf node. Eventually, only leaf nodes remain and the full decision tree has been grown. So, the heart of the algorithm is the rule that determines the initial split [Berson, et.al, 1999].

The information gain measure is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. The attribute with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node. Such an information-theoretic approach minimizes the

expected number of tests needed to classify an object and guarantees that a simple (but not necessarily the simplest) tree is found.

Let us set an example S, which contains positive (p) and negative (n) instance of some target class S. The amount of information, needed to decide if an arbitrary example in S belongs to positive or negative instances is given as:

$$I(p,n) = -\frac{P}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

The expected information content denoted by E(A) of the set will be given by:

$$E(A) = \sum_{i=1}^{v}\frac{p_i + n_i}{p+n} I(p_i,n_i)$$

The information gain that would be obtained by branching on A denoted by Gain(A) will be given by:

$$Gain(A) = I(p,n) - E(A)$$

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data [Brand & Gerritsen, 1998]. Such methods typically use statistical measures to remove the least reliable branches, generally resulting in faster classification and an improvement in the ability of the tree to correctly classify independent test data.

2.6.1.2 Decision tree Pruning

After a data-mining product grows a tree, a business analyst must explore the model. Exploring the tree may reveal nodes or subtrees that are undesirable because of overfitting, or may contain rules that the domain expert feels are inappropriate. Pruning is a common technique used to make a tree more general. It removes splits and the subtrees created by them. In some implementations, pruning is controlled by user configurable parameters that cause splits to be pruned [Brand & Gerritsen, 1998].

There are two common approaches to tree pruning: Prepruning approach, and Postpruning approach. Prepruning is when a tree is "pruned" by halting its construction early. Upon halting the nodes become leafs. The leaf may hold the most frequent class among the subset, samples or the probability distribution of those samples. Postpruning removes branches from a "fully grown" tree. It should be noted that a tree node is pruned by removing its branches. However, prepruning and postpruning may be interleaved for a combined approach. Postpruning requires more computation than prepruning, yet generally leads to a more reliable tree.

If a tree if left to grow more and more branches till the tree reaches a very large size, it will be both computationally expensive but also unnecessary. Most decision tree algorithms stop growing the tree when one of the following three criteria's are met [Brand & Gerritsen, 1998]:

- The segment contains only one record;
- All the records in the segment have identical characteristics, and
- The improvement is not substantial enough to warrant making the split.

2.6.1.3 Trees and Rules

The knowledge represented in decision trees can be extracted and represented in the form of classification IF-THEN rules. One rule is created for each path from the root to a leaf node. Each attribute-value pair along a given path forms a conjunction in the rule antecedent ("IF" part). The leaf node holds the class prediction, forming the rule consequent ("THEN" part). The IF-THEN rules may be easier for humans to understand, particularly if the given tree is very large [Han & Kamber, 2001].

Decision tree methods are often chosen for their ability to generate understandable rules. By using decision trees it is possible to simply trace the path from the root to the leaf where that record landed in order to generate the rule that led to the classification. Many software products can output a tree as a list of rules in SQL, pseudocode, or pseudo-English. Nevertheless, when dealing with a large complex decision tree containing hundreds or thousands of leaves, the tree constructed is hardly more likely to communicate anything intelligible about the problem as a whole than a neural network technique.

2.6.1.4 Decision tree and attribute selection

Using a decision tree, it is possible to pick the most important variable for predicting a particular outcome because these variables are chosen for splitting high in the tree. Another useful consequence of the way that important variables float to the top is that it becomes very easy to spot input variables that are doing too good a job of prediction because they encode knowledge of the outcome that is available in the training data, but would not be available in the filed.

One of the problems with choosing the attributes for trees, especially near the root node, is that once the choice is made and partitioning is done, a permanent decomposition of the problem has been made. Therefore, it is of the outmost importance that precaution is made in choosing the attributes for trees.

Attributes may be relevant or irrelevant for the task at hand. When there are a large number of attributes, even some relevant attributes may be redundant in the presence of other attributes. Relevant attributes may contain useful information directly applicable to the given task by itself, or the information may be (partially) hidden among a subset of attributes. An important problem is the selection of a reasonable subset of the available attributes so that the selected subset can adequately explain (model) the target [Kononenko & Hong, 1997].

Modelling a target attribute by other attributes in the data is perhaps the most traditional data-mining task. When there are many attributes in the data, one need to know which of the attributes are relevant for modeling the target, either as a group or the one feature that is most appropriate to select within the model construction process in progress.

2.6.1.5 Advantages of Decision trees

Decision trees have many advantages, among which some that are mentioned by Berry and Linoff [2000], Berson [1999] and Two Crows Corporation [1999] are:

- Decision trees have the ability to easily generate rules. Thus, decision trees are the favoured technique for building understandable models in addition they also allow for more complex profit and ROI (return on investment) models to be added easily in on top of the predictive model.

- Decision trees have also proven to be easy to integrate with existing IT processes, requiring little pre-processing and cleansing of the data, or extraction of a special purpose file specifically for data mining.

- Decision trees can handle raw data with little or no pre-processing; they provide a simple to understand predictive model based on rules.

- Decision trees handle non-numeric data very well. This ability to accept categorical data minimizes the amount of data transformation and the explosion of predictor variables inherent in neural nets.

- Because decision tree algorithm is fairly robust with respect to a variety of predictor types (e.g. number, categorical etc.) and because it can be run relatively quickly decision trees can be used on the first pass of a data mining run to create a subset of possibly useful predictors that can then be fed into neural networks, nearest neighbor and normal statistical routines – which can take a considerable amount of time to run if there are large number of possible predictors to be used in the model.

- In decision trees numeric inputs are not sensitive to differences of scale between the inputs, and are not sensitive to outliers and skewed distributions. This means that data preparation is less of a burden with techniques such as decision trees than it is with neural networks and k-means clustering.

- Using decision tree, it is possible to pick the most important variables for predicting a particular outcome because these variables are chosen for splitting high in the tree.

- Decision trees make few passes through the data (no more than one pass for each level of the tree) and they work well with many predictor variables. As a consequence, models can be built very quickly, making them suitable for large data sets.

2.6.1.6 Disadvantages of Decision trees

Like all things, decision trees have advantages as well as disadvantages, some of the disadvantages that are stated by Berry and Linoff, [2000]; Two Crows Corporation, [1999],and Han & Kamber, [2001] are :

- Since every split in a decision tree is a test on a single variable, decision trees can never discover rules that involve a relationship between variables. This puts a responsibility on the miner to add derived variables to express relationships that are likely to be important.
- Decision trees are error prone when the number of training examples per class gets small.
- A common criticism of decision trees is that they choose a split using a "greedy" algorithm in which the decision on which variable to split doesn't take into account any effect the split might have on future splits. In other words, the split decision is made at the node "in the moment" and it is never revisited. In addition, all splits are made sequentially, so each split is dependent on its predecessor. Thus all future splits are dependent on the first split, which means the final solution could be very different if a different first split is made.
- Furthermore, algorithms used for splitting are generally univariate; that is, they consider only one predictor variable at a time. And while this approach is one of the reasons the model builds quickly —it limits the number of possible splitting rules to test — it also makes relationships between predictor variables harder to detect.

There have been numerous comparisons on the different classification methods, no single method has been found to be superior over all others for all datasets. Empirical studies show that the accuracies of many algorithms are sufficiently similar that their differences are statistically insignificant, while training times may differ substantially. In general, most neural network and statistical classification methods involving splits tend to be more computationally intensive than most decision tree methods [Han & Kamber, 2001].

CHAPTER THREE

EXISTING CREDIT APPROVAL PROCEDURE AT NIB INTERNATIONAL BANK

3.1 General

Credit risk is risk due to uncertainty in a counterparty ability to meet its obligations in accordance with agreed terms. That is why, prior to extending credit, banks will try to obtain information about the party requesting a loan and credit analysts will review the information about the counterparty. This might include the balance sheet, income statement, recent trends in the potential borrowers industry, the general current economic environment, etc. The banks may also assess the exact nature of an obligation.

Since exposure to credit risk continues to be the leading source of problems in banks worldwide, banks and their supervisors try as much as possible to draw useful lessons from past experiences. Banks today have a keen awareness of the need to be able to identify, measure, monitor and control credit risk as well as to determine that they are able to hold adequate capital against these risks and that they are adequately compensated for risks incurred.

It is in this line that, the present survey intends to assess the application of data mining technology to support credit risk assessment at Nib international bank. The purpose of this section is to conduct an analysis of the current credit approval process by pointing out the activities involved and by identifying possible data sources that can be used to this end.

According to CRISP-DM [2000] the initial step in a data mining process is understanding the business. Accordingly, this chapter is devoted to a description of the credit approval process at Nib. It starts by giving a brief overview of Nib international bank, followed by a detailed description of the credit approval process, and on the procedure of follow-up. At the end of the chapter, the findings of the survey are included.

In the process of conducting this survey a number of interviews were made with the bank senior credit analyst. In addition, the researcher made use of secondary sources like documents, memorandums, circulars and publications to support/supplement the information obtained.

3.2 Nib International Bank

Nib international bank is one of the 6 private commercial banks in Ethiopia. Nib bank was established on May 26th 1999 under license no. LBB/007/99 in accordance with the commercial code of Ethiopia, and the proclamation for licensing and supervision of banking business No. 84/1994 [NIB, 2003].

As of June 30, 2003 Nib international bank has a total of 1934 shareholders, and operates with a paid-up capital, reserve and unappropriated profit of birr 124.9 million. During the same period, it had a total of 364 staff members and operated through its 11 branches and one Agency bureau at Bole International Airport passengers' terminal.

Nib bank was established with the following main objectives:

1. Provide efficient commercial banking services to the business community and the population at large;
2. Help spread and deepen banking habit among the population;
3. Introduce and popularize modern banking practices in the country;
4. Maximize operational results and thereby enhance shareholder's benefits.

Following these objectives the bank offers many services to its customers among which; mobilizing deposits in various forms such as savings, time, demand/checking and special deposits as well as the extension of credit in the form of overdraft facilities, short term and medium term loans to different users in the various sectors of the economy are the major ones. Providing complete foreign banking services by strengthening and establishing correspondent banking relationship with a number of major global banks are also principal areas of focus of the bank.

The major sectors for which the bank grants loans are agriculture, domestic trade and services, manufacturing, exports, imports, hotel and tourism, building and construction, transportation and consumer and personal. As of June 2003 loans disbursed for loans for manufacturing, domestic trade and services, exports and imports together account for 82.5% of total loan outstanding. Loans and advances for imports (overdraft and term loan) alone had the higher share of over 27.8% of the total loans followed by loans for manufacturing 25.7%, domestic trade of over 21% and building construction of over 7.9% of the total loans and advances portfolio.

The bank over the years achieved a reasonable profit both in domestic and international banking activities and had recorded a profit before tax of birr 19.3 million and after tax of birr 13.5 million, despite these benefits the banks management major concern is that non-performing loans have increased significantly during the years, partly as a result of the weak economic condition facing borrowers and also due to mismanagement by some debtors. They also state that the major share of the responsibility still remains with them (i.e. the bank) by not making adequate credit appraisal, analysis and follow-up.

3.3 Credit Approval Process

Creditors follow a credit approval process to determine the soundness of the borrower's business and make an assessment of whether the prospective borrower agreement to pay the loans due to the bank will actually be met. That is why; each credit approval should be subject to careful analysis by a credit analyst.

Nib international bank follows specific steps in the process of credit approval. Generally stating the credit approval process at Nib bank can be represented in the following format.

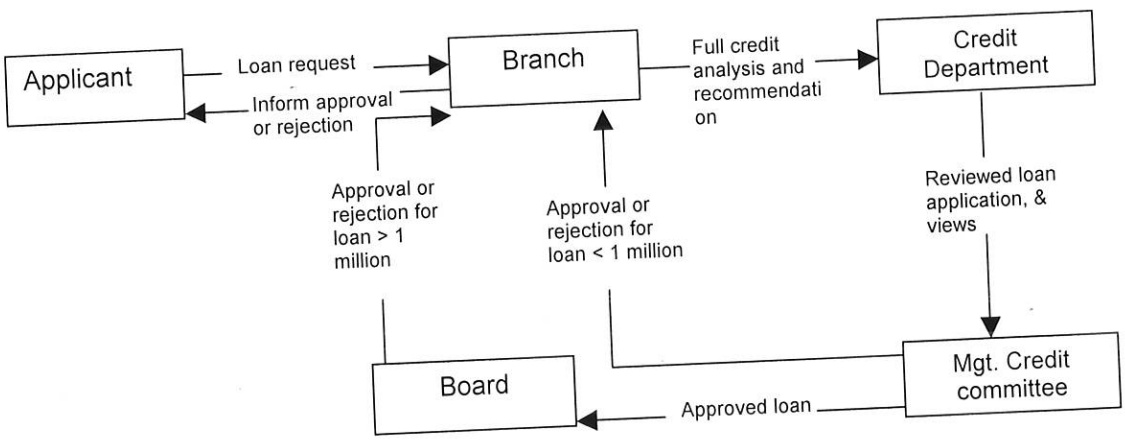


Figure 3: Credit approval process at Nib International Bank

3.1.1 Requirements for loan application

Banks must receive sufficient information to enable a comprehensive assessment of the risk profile of the prospective borrower. Accordingly, Nib bank makes use of a checklist when inspecting for completeness of the document to be submitted by a prospective borrower. In addition, the same document can be used by the applicant in preparing loan request (form attached as Annex 1).

As a minimum, according to the checklist a prospective borrower's application letter should contain the following:

1. Full name and address of the applicant (borrower), name and address of the business;
2. Owner of the business and amount of investment in the business;
3. Purpose of the credit and source of repayment;
4. The amount and type of finance requested, (e.g. overdraft, term loan, merchandise);
5. Proposed terms and conditions of the credit.

In addition to the application letter the request for loan should be supported by the following documents:

1. A trade license to operate the business renewed for current year;
2. Business profile and description of the business;
3. Financial statements where applicable preferably audited (financial credit report);
4. Inclusion of guarantees.

3.1.2 Analysis of documents

The first and initial step for banks is to ensure that the information received is sufficient to make proper credit-granting decisions. Thus, whenever a prospective borrower comes into the bank, the bank follows a number of steps in order to approve or reject the request. As described in figure 3 each party involved has specific duties and responsibilities in the approval process.

The first step to be performed as soon as receiving a request from a prospective borrower is done by branch offices. The branch credit analyst reviews the presented documents, checks authenticity and fullness and sends the document to head office credit department. In doing so, the branch follows the checklist, and gathers information by interviewing to get all relevant information from applicant, pay business visit, undertake full credit analysis, obtain credit information and have collateral appraised etc...

The following sections describe the loan approval process.

PAYING BUSINESS VISIT

Once the branch credit analyst makes sure that all documents are properly appended to the application letter, then the next step to be performed is paying a visit to the business of the prospective borrower to ascertain the correctness of the information provided, and to have a view of the business.

FINANCIAL STATEMENT ANALYSIS

For any bank to give or grant a loan, it first has to make sure of the financial soundness of the business. So analyzing the financial statement is the next step performed by the bank. If the

financial statement is audited, the bank takes the information as is, but if it is not, the bank makes use of the financial credit report form (FCR). The FCR contains 3 categories which are application, financial statement and description (Format attached as Annex 2), this form then serves to the bank as a means for analysing the financial soundness of a business where an audited financial statement is unavailable.

COLLECTING CREDIT HISTORY

According to a recent publication of commercial bank of Ethiopia (2004) a new regulation has been put in place that states if a borrower has been registered as a defaulter in one bank, he would not be granted loan in any other bank under any circumstances. Consequently, collecting credit information on a prospective borrower is vital to any bank.

At this stage the bank collects credit information on prospective borrowers which consists of past credit history. The past credit history is collected not only from its own records but from all other commercial banks.

In collecting credit history from other commercial banks, the bank requests other banks for information concerning the prospective borrower (i.e. had he had any relation with the banks?, if so how is the performance of the applicant on the repayment? etc...).

COLLATERAL APPRAISAL

Any bank while granting a loan requires sufficient guarantee for the repayment of the loan of money. The major collaterals accepted by NIB international bank are buildings, business mortgage, leased lands, vehicles, machinery, local or foreign bank guarantees, cash, and merchandise. Consequently, the bank assesses the collateral to determine the acceptability as well as to verify its sufficiency for the requested loan.

After completing the above processes, the branch credit department submits loan application with full credit analysis and recommendation to the head office credit department staff who reviews the loan application, check the analysis and attach its views and sends it to members of management credit committee.

3.1.3 Recommendation and Approval

According domain experts of the bank, the loan decision is made based on 3 criteria's. They are; know the customer, know the business and know the collateral.

Know the customer: Having a good knowledge about the customer's past credit behaviours, his financial status, risk management ability etc...

Know the business: Having a good knowledge on the economic political and social conditions on the area of the applicant's business.

Know the collateral: Knowing the ownership of the collateral given for the money to be borrowed, assessing its acceptability, assessing whether the collateral offered is proportional to the amount requested.

All things considered, recommendation is not only done by the management credit committee nor is it done by only the board of directors. Recommendation is first obtained from branch credit department who sends the loan application with full credit analysis and recommendation filled out on a loan approval form (attached as Annex 3). Having the recommendation of the branch credit department the head office credit department reviews the loan application, checks the analysis and attaches its views and sends the document to members of management credit committee.

If the loan request is less than 1 million Birr the loan will be approved at the management credit committee level, within two days of receipt. But if the loan request is greater than 1 million, loans approved by management credit committee recommended to the board should be supported by a full write-up and analysis from the credit department. The board of directors will give the answer to the credit department of branch within a maximum of 3 days.

All things considered an applicant will receive an answer as to the acceptance or rejection of his request within 18 to 30 days of the submission of his request.

3.4 Credit Follow-Up

Credit follow-up is the major activity performed by the bank to make sure that loans are paid on time and with interest. The credit follow-up process helps recognize problems related to irregular payment of loans. In addition this process will be a means to take timely measure to solve identified problems.

In Nib International Bank each branch uses a form known as Monthly loans and Advances Return form (form attached as Annex 4) in order to perform the credit follow-up on borrowers. The return form is used in order to conduct a regular follow-up on disbursed loans. This form will be faxed to the head office credit department, which will undertake appropriate measures depending on the arrears on loans disbursed. These forms are used to identify potential problem credits and other transactions to ensure that they are subject to more frequent monitoring as well as possible corrective action, classification and/or provisioning.

CHAPTER FOUR

DATA COLLECTION, PREPARATION AND MODEL BUILDING

This chapter as the fundamental part of this research project describes the tools that are used in experimentation as well as details the major steps that were carried out in the research project. Test results are then presented and discussed.

The research mainly follows the typical stages that characterize a data mining process. Thus, the present work is organized based on the CRoss-Industry Standard Process for Data Mining (CRISP-DM) process cycle [CRISP-DM, 2000]. The major steps undertaken were: data mining goal setting, data understanding, data preparation, modelling, evaluation and deployment.

4.1 Data Mining Goal

The survey conducted in Chapter 3 revealed that the present credit approval process is mainly based on information gathered on creditors. So the first data mining goal was to identify the possible data sources.

After the data sources have been identified the next task performed is identification of the variables that determine customer's credit worthiness and using these variables to build and extract rules that can be used to assess the creditworthiness of prospective borrowers. This is believed by the researcher to facilitate the credit approval process of the bank by supporting the tasks performed by the credit analyst.

The most appropriate data mining technique for the present problem, which is tree classification, is used for the present experiment. Tree classification is said to be the best because it is easily explainable, can be build interactively and the expert can easily view the tree and see how a leaf or terminal node was reached.

In addition, so as to provide a classification mechanism that could easily be used by the credit analyst, emphasis was given to the data preparation and an exploratory data analysis. This process allowed the identification of important attributes as input for the model-building phase.

The success criterion for this data-mining project is the discovery of borrower classification rules that would find out and differentiate borrowers who are creditworthy to those who are potential defaulters. Provided that reasonable customer classification rules are discovered, the bank could device a means to reduce the number of non-performing loans and be able to recognize the creditworthiness of potential creditors when a request for loan is obtained.

4.2 DATA COLLECTION

The present research was originally aimed to investigate the application of data mining technology specially the use of decision tree technique in Dashen Bank. A previous research on Dashen Bank by Askale [2001] to support loan approval process was conducted by the use of neural network technique. The present research was aimed to be a replication of that research by the use of decision tree technique, so as to be able to recommend the best technique or to assess if a combination of techniques is best to support the process.

Unfortunately, getting access to the data was very difficult for the researcher though a great effort was made to access it.

Consequently, an alternative source of data was looked for, and Nib international Bank S.C. was chosen. The choice took into consideration the willingness of the officials of the Bank to provide the researcher with data as well as with assistance throughout the process of the research.

4.2.1 Data Mining Tool Selection

Along with understanding the data mining goals, an important task is the selection of an appropriate tool for the task at hand. The selection process in this research was based on the data mining tools' ability to support various methods for different stages of the process to be conducted. Among the criteria's used for the selection, based on the work of Ken et.al.[1999], Abbott et.al. [1999] and Han and Kamber [2001] the most important ones are:

- The data mining tasks that the tool is intended for (i.e. classification);
- The inclusion of a variety of capabilities, techniques, and methodologies for data mining;
- The algorithms supported (decision trees);
- The computer architecture and operating system on which the software runs (stand alone) and a MS Windows operating system;
- The ability to handle a variety of data sources in an efficient manner (MS Access or MS Excel);

- The ability to allow the user to perform the variety of data cleansing, manipulation, transformation, visualization and other tasks that support data mining. These tasks include data selection, cleansing, enrichment, value substitution, data filtering, binning of continuous data, generating derived variables, randomizing, deleting records, etc;
- The maximum number of records the software can comfortably handle.

Even though numerous tools are available in the market, budget constraint was a major limiting factor in the acquisition of an appropriate tool. Therefore, the identification of an appropriate data mining tool was a time consuming task. The researcher after assessing a number of tools finally approached the vendor of Knowledge Studio version 4.1.1 of Angoss Software Corporation (Angoss), and was also able to download the trial version of See5 software. Since the See5 was a trial version, could only handle more than 400 records at a time and that the cost of the tool was beyond the budget allocated for the project, the researcher did not opt to use it.

On the other hand, after subsequent contacts with Angoss (vendor of Knowledge Studio software) the researcher was able to download the software from the Internet. Unlike See5, Knowledge studio can handle a large amount of data at a time, in addition it provides a number of ways to visually explore and express patterns in a data. All in all, it can help with the whole data mining process – from preparing data through to producing final graphs and reports.

Therefore, Knowledge Studio Version 4.1.1. of Angoss Software Corporation (www.angoss.com) was the sole software used as the tool for the experiment conducted since it fulfilled most of the criteria set above and because it performed well during experimentation.

4.3 DATA UNDERSTANDING

After a good understanding of the data-mining goal was established, the next step to be performed was establishing an understanding of the data to be used.

A precondition to any data mining is data itself. A good source of data for data mining purpose is identified to be the corporate data warehouse [Berry & Linoff, 2000]. The reason for this is that the data is stored in common format with consistent definitions for fields and their values. Unfortunately, the available sources of information that held past data on people who defaulted on a loan and those who did not were stored in manual format. The three fundamental documents that held past data of borrowers were the loan approval form (attached as Annex 3), the monthly loans and advances return form (attached as Annex 4) and the financial credit report (attached as Annex 2). These documents are generated by the different area banks.

The loan approval form as described in the survey of the credit approval process contains demographic data on potential borrowers as well as information about the purpose of credit request, the security offered and the value, the financial information etc... The monthly loans and advances return form contains credit follow-up. This form contains information on the amount that has been collected and the number of days in arrears and the arrears amount due by each of the borrowers. The third and the last document utilized to collect the necessary information for this research is the financial credit report form, this form shows the financial status of the borrower where an audited financial statement is unavailable. According to the investigation conducted and based on expert opinion these forms together provided the required information essential for the experimentation.

NIB at the time of this research operated with 11 branches among which two were outside Addis Ababa i.e. Adama and Awassa and with one Agency bureau at Bole International Airport passengers' terminal. Therefore, the branches taken for sampling were nine. The nine branches were chosen based on the fact that it was impossible to access all the relevant information from Adama and Awassa branches due to their geographic location added to the number of years they have been in business (i.e. 1 to 2 years). In addition Awassa branch did not have any borrower up until the period the research was conducted. Adama branch was opened in 2003 and therefore did not have the amount of data that is required for data mining.

During the period chosen for the present experiment, only 6 branches had a good number of borrowers. The number of borrowers in the rest 3 branches was very small. Consequently, the researcher used purposive sampling method where the branches whose records were very small were excluded; these are Mamokacha, Ras and Urael Branches.

Since the other remaining branches were all located in Addis Ababa, the researcher was able to access the data required for the research undertaken. Accordingly, Abinet, Adarash, Main, Shola, Tana and Tiret were the branches whose data was collected as a sample for the present research.

From among the four types of loans, which are term loan, overdraft facility, merchandise loan and letter of credit facility that are available at NIB, term loans were chosen for this research for three reasons. Term loan is a loan type that is highly prone to irregularity due to its nature i.e. fixed instalments over a relatively extended time compared to other types of loans. In addition, term loan is a loan type that is being availed in all the branch offices. Moreover, term loans are the more common loans at the Bank and they constitute major proportion of the total outstanding loans.

The most important point that needed attention before the actual data collection is the decision of the period to be covered by the research. Nib Bank was established in 1999, and operated mostly from its branches in Addis Ababa. Its' branch offices outside Addis were opened only recently, and one of them did not even have any borrowers. In addition the researcher could not include loans after 2003 since most of the loans are still outstanding and their outcome is still not known. So the loan period from 2000 to 2002, spanning two year was taken.

The number of records collected from the six branches is summarized in table 1 given below.

No.	Branch	No. of loan Accounts
1	Abinet	152
2	Adarash	91
3	Main	252
4	Shola	70
5	Tana	231
6	Tiret	126
Total		**922**

Table 1: Distribution of collected data with respect to sample branch

The data collected from the branches was stored into a single table by making use of MS Excel because of its ease in calculation and graphical facilities. Having the data in excel table was an advantage because the software chosen for analysis, i.e Knowledge Studio has the facility to import data from dBase, Lotus, Excel, SAS, SPSS, ODBC, text files and others and can only accept dataset in a single table. Keying in the data by the use of Ms Excel consumed quite a considerable amount of time due to the fact that the researcher had to go through each monthly loans and advances return form. The variables that were collected from the sources identified (found filled in box files at the head office) were based on those identified by Askale [2001] (attached as Annex 5) and by obtaining expert opinions at NIB who have suggested the addition of some additional relevant variable (attached as Annex 5).

The fields that are identified and collected from the loan approval form, financial credit report form and the monthly loans and advances return forms are as follows:

1. Fields collected from the financial credit report (or from the audited financial statement provided by borrowers)

- Asset
- Capital
- Current Asset
- Current Liability
- Liability
- Total liability
- Current liability

2. Fields collected from the loan approval form

- Business Establishment year
- Sex
- Trade Sector
- Relationship with NIB

- Relationship with other banks/third party
- Number of prior loans (number of loans the borrower settled in the past)

3. Fields collected from the monthly loans and advances return form

- Name of customer
- Amount granted
- Date granted
- Due date
- Collateral type
- Collateral value
- Term of payment (monthly, quarterly, bimonthly)
- How early a prior loan was settled
- Performance of prior loans (Whether prior loans were settled regularly or not)
- Branch
- Yearly income

These documents from which these variables were collected do not only contain information in relation to these fields collected. Other variables contained in these documents were not incorporated based on the banks experts' opinion because they were found less relevant. The collected variables were not all used in their raw format. Some of these variables were collected so as to be able to derive other important variables (ratios), like the debt to asset ratio, the current ratio, and debt to equity ratio, which are considered by the bank experts as being very important in the process of loan assessment.

These variables collected in this manner are the independent variables. On the other hand the dependent variable, which is classification of a loan, was collected from the monthly loans and advances return form. Obtaining these variables was a very time taking task because the repayment of a particular borrower was checked by examining all the monthly reports during the life of that loan period up to its settlement. For instance, if the loan period of a certain loan is twelve months, then twelve monthly credit reports had to be reviewed to see whether the borrower was regular in repaying his loan or not.

All Commercial banks in Ethiopia report their loan to the National Bank of Ethiopia under four headings, these are: regular, substandard, doubtful, and loss. A borrower is said to be regular if the settlement of the loan is done before or on the due date, his account is said substandard if he settled his loan with 90 to 180 days arrears; doubtful if the loan had arrears of 180 to 360 days and loss if the loan has 360 days and above arrears. According to NIB experts, accounts classified as loss are not actually loss so loans, which were classified under loss, were labelled as doubtful for the purpose of this research. The researcher has adopted the following classification depicted in tables 2.

Classification	Description
Regular	Loans settled with regular repayment
Substandard	Loans settled with 90 upto 180 days arrears
Doubtful	Loans with an overdue balance. Loans with arrears of 180 days and above including those that had been rescheduled but still not settled

Table 2: Description of the 3 classification of a loan

4.4 Data Preparation

The heart of data mining is transforming data into actionable results. Hence, the main goal in this section was the production of the dataset (datasets) used for modeling. Based on the work of CRISP-DM [2000] the main activities during this phase included: data cleaning, data selection and data transformation and aggregation.

4.4.1 Data cleaning

Missing values

In this phase the specific tasks that are performed are handling missing values, smooth noisy data, and resolve inconsistencies.

The data collected on NIB bank borrowers had few records, which had no recorded value for several attributes, such as the liability, capital, and asset. The cause for missing information was that the researcher was not able to depict what exactly was written for some and for others these data did not appear in the documents consulted. Asked about these problems the bank experts explain that some borrowers are not required to provide their financial information because they are well renowned personnel in the country or they are individuals that had a close relation with the bank, like major shareholders; in addition data for some borrowers had been misplaced.

Possible ways suggested by Two Crows Corporation [1999] to handle missing values was adopted in order to deal with existing missing values.

For continuous variables, i.e. variable with numerical values Two Crows Corporation (ibid) suggests that missing values be replaced with the mean value for that field. The fields whose values were organized in this manner are 'years in business', and 'duration' (i.e the period the loan was granted for). The mean value for the two continuous variables (taken by approximating the values) is given in the table below.

Field Name	Mean Value
Years in business	8 (years)
Duration	480 (days)

Ordinal variables on the other hand can be substituted by the median of that field [Two Crows Corporation, 1999]. The field that was ordered in this manner was "Classification", i.e. variable used in the present experiment as the dependent variable. The three variables selected for this field described in table 2 are "Regular", "Substandard" and "Doubtful". Based on domain experts knowledge ordered from the lowest to the highest they are doubtful, substandard and then regular. So the median value for this field is "Substandard" which is going to be used to replace the tuples with missing value for this variable.

The third method mentioned by Two Crows Corporation (ibid) to handle missing values is replacing nominal variables with their modal values. The variables handled in this manner are 'trade sector', 'security type', 'term of payment', 'sex', and 'purpose of loan'. The modal value for these nominal variables is given in the table below.

Field Name	Modal value
Trade Sector	Manufacturing
Security type	BLD (Building)
Term of payment	Monthly
Sex	M (male)
Purpose of loan	Working capital

SUMMARIZATION

Data summarization is very important if there are only few examples at the finest level of detail. In the collected data for this study, however, the distribution of facts was found to be reasonable, except that of the field "Trade Sector" with many possible categories. In order to have sufficient categories based on expert opinion and based on the work of Askale [2001], the following grouping categories were taken.

- Building materials
- Electronics
- Sundry Goods
- Clothes
- Cosmetics and Jewelry
- Manufacturing
- Food Items
- Building and construction
- Hotel
- Furniture and other household goods
- Clinic and Pharmacy
- Coffee
- Photography
- Import
- Export
- Spare Parts
- Transport
- Agriculture

HANDLING INCONSISTENT DATA ENCODING

When collecting information from the various sources identified in the previous sections the researcher experienced problems due to data encoding. According to Two Crows Corporation [1999] if such problem is not dealt with properly they may create quality problems.

The problem faced by the researcher due to data encoding was because the different branch offices who are the originators of the various sources of information considered for the present research had different encoding mechanisms for fields. The different branches of NIB encoded 'term of payment' as "quarterly" and "Q". These encoding mechanisms as they represented the same thing, a uniform encoding mechanism had to be developed. The uniform encoding mechanism adopted for these fields are:

Field Name (term of payment)	Description
Monthly	Payment is made every month
Bimonthly	Payment is made twice a month
Quarterly	Payment is done twice a year

In addition to the above, the branches used different mechanism to encode security type. For example a loan secured against buildings is encoded as 'residential building', "residential house' and 'building'. Therefore, the following format was adopted in encoding the variable "Security type".

Field Name (Security type)	Description
BLD	Loans secured against buildings only
PG	Loans secured against personal guarantor
Vehicle	Loans secured against vehicles only
Merchandise	Loans secured against merchandise only
Machinery	Loans secured against machinery only
BV	Loans secured against both building and vehicle
BP	Loans secured against both building and personal guarantor
VP	Loans secured against both vehicle and personal guarantor
Stock	Loans secured against stocks
Land	Loans secured against land
Cash	Loans secured against cash

The other field that had data encoding problem was the sex field, where the sex either in full or where an abbreviation was used. For example a borrower who is male was represented as 'male' or 'm'. So for consistency purpose the following encoding mechanism was used in data encoding.

Field name (sex)	Description
M	Male borrower
F	Female borrower
MF	Loan in the name of both a male and female borrower
CP	Loan disbursed to a company or partnership

4.4.2 Data selection

As described in the data description phase, selection of data to be incorporated in the research work for analysis took into consideration the number of years a branch had been in business, its geographic location and the number of borrowers at that branch. Accordingly, Abinet, Adarash, Main, Shola, Tana and Tiret were the branches whose data

was collected as a sample for the present research. By consulting the bank experts, records with missing financial information were excluded in order to avoid compromising the results obtained. Records excluded in this manner were 28 (3%). Consequently, the number of records used after data selections were 894 from among the records that were initially collected.

4.4.3 Data Transformation and Aggregation

This task, according to CRISP-DM [2000], includes constructive data preparation operations such as the production of derived attributes, complete new records or transformed values for existing attributes.

It is often necessary to construct new predictors derived from the raw data. Certain variables that have little effect alone may need to be combined with others, using various arithmetic or algebraic operations (e.g., addition, ratios) [Two Crows Corporation, 1999].

According to the experts at NIB, financial management books Peterson [1994], Straub [2000] and based on the findings of Askale [2001] the ratios and values that are considered essential in determining the credit worthiness of an individual are: current ratio, debt to asset ratio, debt to equity ratio and net working capital. These four financial indicators were obtained by deriving values from the existing fields and were incorporated as new fields. Additional computed fields were also incorporated based on discussion with NIB experts.

The total fields considered for the present study are 28. The fields' names and their description are listed in the table below.

No.	Name of Variable	Data type	Description	Source
1	Branch	Text	Name of the branch from where the loan the loan was given	Loans and advances return form
2	Loan No.	Number	The number of loan for the specific borrower (1st loan, 2nd loan, 3rd loan etc.)	Computed
3	Month	Date	Month on which loan was granted	Computed
4	Duration	Number	The duration of loan in number of days	Computed
5	Yearly Payment	Number	The estimated amount to be paid in a year	Computed
6	Amount Granted	Number	Amount granted	
7	Loan/Time Ratio	Number	Loan amount divided by the loan duration	Computed
8	Asset	Number	Total asset of the borrower	Financial Credit Form
9	Capital	Number	Total capital of the borrower	Financial Credit Form
10	Current Asset	Number	Total current asset of the borrower	Financial Credit Form
11	Current Liability	Number	Total current liability of the borrower	Financial Credit Form
12	Net working capital	Number	Current asset - Current Liability	Computed
13	Liability	Number	Total liability of the borrower	Financial Credit form
14	Debt/Asset Ratio	Number	Liability value divided by asset value	Computed
15	A. Debt/Asset Ratio	Number	The anticipated debt/asset ratio after considering the new loan to be granted	Computed
16	A. Current Ratio	Number	The anticipated current ratio after	Computed

			considering the new loan to be granted	
17	Debt/ Equity ratio	Number	Liability value divided by capital value	Computed
18	Security type	Text	Type of security (e.g. building, vehicle, personal guarantee)	Loan approval form
19	Security value	Number	Estimated value of the security	Loan approval form
20	Security/Loan Ratio	Number	Security value divided by the amount granted	Computed
21	Sex	Text	Sex of the borrower	Loan approval form
22	Trade Sector	Text	The kind dof business borrower is engaged in	Loan Approval form
23	Years in business	Number	The number of years the borrower has been in business	Computed
24	Term of payment	Text	the term of payment (e.g. monthly, bimonthly or quarterly)	Loans and advances return form
25	No of prior loans	Number	The number of loans borrower has settled in the past	Loan approval form
26	Per. Of prior loans	Text	Performance of past loans (i.e. whether past loans were regular or not)	Loans and advances return form
27	Per. In other types of loans	Text	Performance of past loans (not in term loans)	Loans and advances return form
28	credit relationship with other banks	Text	Credit relationship with other banks in the country (i.e. whether they were regular or not or even if there were none)	Loans approval form

Table 3: List of Independent variables (inputs) used for model building along with their descriptions

Explanation on the how the derived variables are computed is given below:

Loan No. = (No of prior loans +1)

Month = Derived from the 'date granted' column

Duration = (Expiry Date – Date Granted)

Yearly Payment: Is an approximate figure that did not take into account the interest amount

Yearly Payment = (Amount Granted * 365)/ Duration

Loan/ Time Ratio = (Amount Granted/ Duration)

Net working Capital = (Current Asset – Current Liability)

Debt to Asset Ratio = Liability / Asset

A. Debt/Asset Ratio = (A. Liability/ A. Asset)

A. Liability = (Liability + Amount Granted)

A. Asset = (Asset + Amount Granted)

Current Ratio = (Current Asset/Current Liability)

A. Current Ratio = (A. Current Asset / A. Current Liability)

A. Current Asset = (Current Asset + Amount Granted)

A. Current Liability = (Current Liability + Yearly Payment)

Security/Loan Ratio = (Security Value/ Amount Granted)

Years in Business = (1996- Business establishment year)

Debt to Equity= Total Liability/Total Capital

4.5 MODELING

Basically data classification is a two-step process. In the first step a model is built and then represented in the form of decision tree and classification rules. In the second step, after

the predictive accuracy of the model is estimated, the model is used for classifying future data tuples for which the class label is not known [Han & Kamber, 2001].

The present research was more concerned in generating rules to explain credit risks and to come to an understanding of the most important variables affecting the loan repayment of borrowers than in simply classifying borrowers as good or bad creditors, or predicting which potential borrowers are likely to be bad or good creditors. As a result, the decision tree that made the best predictions was not the one most useful for this research, but the tree that generated the soundest rules (based on domain experts' opinion) was given priority in model selection.

With this understanding, numerous decision trees were built by changing the variables utilized so as to discover the most important ones to be used in the final model building. For each decision tree constructed the corresponding rule set was extracted. And finally, on the basis of the evaluation of the rule sets made by domain experts, a tree model whose rule set is found to be meaningful was selected as a working model for classifying future records to a specific class.

To this end, the first step undertaken in the following section is the selection of the actual modelling technique that is to be used, then generating test design, building models, assess the model based on domain experts' knowledge and finally testing the predictive accuracy of the model.

4.5.1 Selection of modeling technique

According to CRISP-DM [2001] this step is where selection of the actual modelling technique that is to be used in the model building is done. Accordingly, the first step that

was carried out in this section is the selection of the algorithm and the measure that Knowledge tree can use to compute the best split.

In KnowledgeSTUDIO there are two types of algorithms. The algorithms that were available for the construction of decision tree were KnowledgeSEEKER and HeatSEEKER. KnowledgeSEEKER is a flexible algorithm that is especially good for exploration purposes and manual tree building and it can also handle a large amount of variables with either a continuous or discrete dependent variable. Unlike KnowledgeSEEKER, HeatSEEKER is a fast algorithm that is especially good for automatically generating a tree, and even if it can handle a very large number of records, it performs better with fewer variables. Since in the present research the number of variables considered were many (i.e 28), and there was a need for manually adjusting the tree and to explore the results obtained, the algorithm selected in model building was knowledgeSEEKER.

The KnowledgeSEEKER decision tree algorithm first identifies the best relationship between each of the predictor variables in the analysis and the dependent variable. These relationships may be conceived of a table or as branches on a decision tree. The categories of the tables that are formed are used to establish the values of the branches of the decision tree that is used to display the tabular relationship. For each relationship that has been identified KnowledgeSEEKER tests whether the relationship is significant. Significant relationships are presented as alternative ways of forming the branches of a statistical decision tree, one by one. Finally relationships are presented in order of statistical significance.

In addition to the above mentioned algorithms the measure used to build the tree is Adjusted – P-value Bonferroni Adjustment measure because it is the most statistically sound approach given that it reduces the possibility of producing chance results and puts all variables on a common statistical footing. The choice is also based on the fact that unordered grouping variables and variables with many values have a biased chance of being selected as branches unless adjusted tests are used.

4.5.2 Generate test Design

After the modelling algorithm and the modelling measure have been selected and before the model was actually built, a mechanism was generated to test the model's quality and validity. In this step what was performed is the separation of the dataset into training and testing set so as to build the model on the training set and estimate its quality on the testing set.

KnowledgeSTUDIO Partitioning dialog box, shown in figure 4 was used to partition the dataset into learning and testing sets.

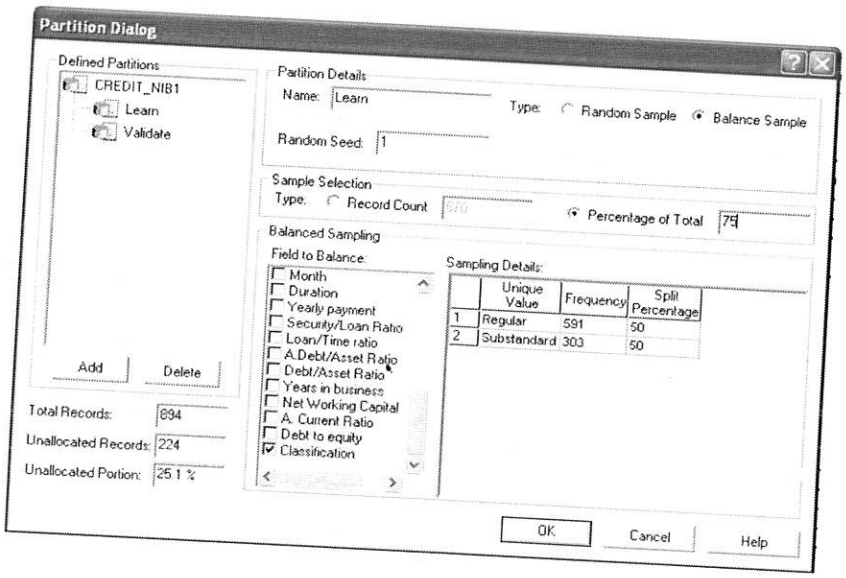


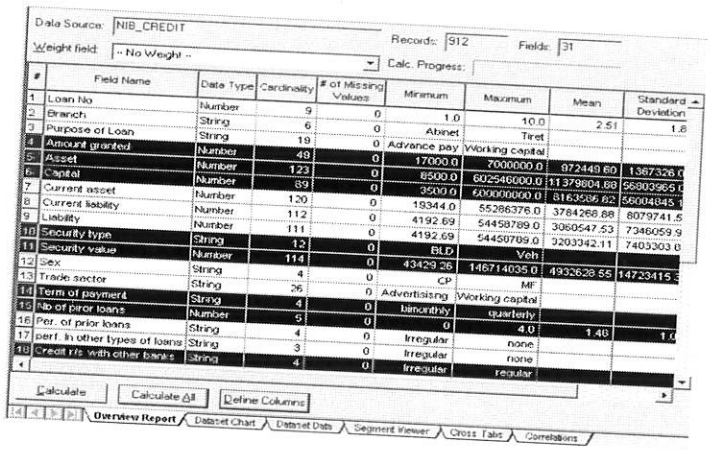
Figure 4: Partitioning the dataset for training and testing

The dataset prepared for this study, 894, is divided into two: 75% (672) for model building and the remaining 25% (222) for testing set. The present research made use of a balanced sample as the overrepresentation of one class in the sample may cause biased parameters estimations in the model. In this case, ≅75% records were used for model building and the remaining ≅25% for testing the resulting model based on balanced sample on the dependent field 'classification'.

4.5.3 Build Model

After the selection of the algorithm and the partitioning of the dataset have been done, the next step was running the modelling tool on the prepared dataset to create the best possible model. When the results obtained are not satisfactory, measures were taken like adjusting the values of the parameters used as well as excluding and including variables.

KnowledgeSTUDIO's data overview report (figure 5) is used to view variables to be used for constructing decision trees.



#	Field Name	Data Type	Cardinality	# of Missing Values	Minimum	Maximum	Mean	Standard Deviation
1	Loan No	Number	9	0	1.0	10.0	2.51	1.8
2	Branch	String	6	0	Abinet	Tiret		
3	Purpose of Loan	String	19	0	Advance pay	Working capital		
4	Amount granted	Number	49	0	17000.0	7000000.0	972449.60	1367326.0
5	Asset	Number	123	0	8500.0	602546000.0	11379804.88	56803965.0
6	Capital	Number	89	0	3500.0	600000000.0	8163596.62	56004845.1
7	Current asset	Number	120	0	19344.0	55286376.0	3784268.88	8079741.5
8	Current liability	Number	112	0	4192.69	54458789.0	3060547.53	7346059.9
9	Liability	Number	111	0	4192.69	54459709.0	3203342.11	7403303.8
10	Security type	String	12	0	BLD	Veh		
11	Security value	Number	114	0	43429.26	146714035.0	4932628.55	14723415.3
12	Sex	String	4	0	CP	MF		
13	Trade sector	String	26	0	Advertising	Working capital		
14	Term of payment	String	4	0	bimonthly	quarterly		
15	Nb of prior loans	Number	5	0	0	4.0	1.46	1.0
16	Per. of prior loans	String	4	0	Irregular	none		
17	perf. in other types of loans	String	3	0	Irregular	none		
18	Credit rlt. with other banks	String	4	0	Irregular	regular		

Figure 5: Overview report of the training dataset

Knowledge Studio data attribute editor (figure 6) and tree attribute editor which are similar give a means for adjusting parameters by selecting columns in the grid. The editors allows manipulation, among other things, of the independent variables and the dependent variable used in an analysis, as well as grouping type, order display, mapping, and missing value use.

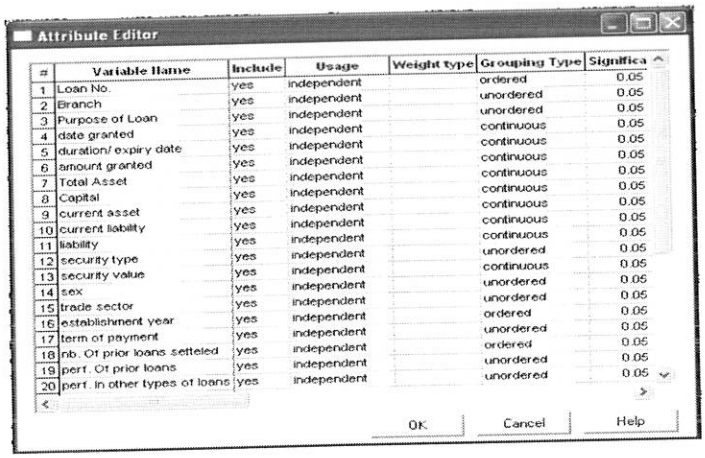


Figure 6: Overview of the tree attribute editor

4.5.3.1 Decision tree model building

PHASE 1: Building Models

Experiment 1

In this first experiment the decision tree built was based on the training set which is 75% (i.e.672) of the data that has been prepared and imported to the KnowledgeSTUDIO software. And since classification is a supervised data mining technique, the variable "classification" was set as dependent variable and all others were set as independent variables. The classification tree was built using all the default parameters suggested by the software (i.e. KnowledgeSTUDIO) using all the 28 variables.

Though each record in the dataset included 29 attributes with the inclusion of the dependent variable 'classification', the decision tree constructed had only used 9 attributes.

This could have indicated that attributes that are not considered in the construction of the decision tree are not important to discriminate the records into the predefined classes (i.e Regular, Substandard and Doubtful) provided to the model. But domain experts found the attributes identified as being neither sufficient nor all totally relevant to classify potential borrowers. Consequently, subsequent changes were made based on StatSoft [2003] suggestion that it may often be most useful to combine the automatic methods for building trees with "educated guesses" and domain-specific expertise. Therefore, some portions of subsequent trees were grown using automatic methods; and refining and modifying the tree was made based on domain expert's expertise.

Eventhough subsequent changes were made on the tree itself and on the parameters utilized; the results obtained were not very encouraging. And on closer look at the data, uneven representation was identified by the researcher as potential source for the poor decision tree that has been generated. This is due to the fact that decision trees are error prone when the number of training example per class gets small. The subdivision of the records with respect to outcome was 613 (69%) regular, 173 (19%) substandard, and 108(12%) doubtful. As explained in chapter 3 loans classified as substandard and loans classified as doubtful are both irregular loans. What differentiates these two types of loans is that substandard loans are loans that were paid with 90 to 180 days arrears and doubtful loans are loans with 180 to 360 days arrears. The researcher after careful consideration and by consulting domain experts decided to classify both substandard and doubtful loans together. The resulting records were then 613(69%) regular and 281 (31%) irregular loans classified as substandard.

EXPERIMENT 2

After the above mentioned changes have been made, the first decision tree constructed was based on all the 28 variables and on the dependent variable 'classification' which is

used to classify borrowers as regular or substandard. Figure 7 below shows the output decision tree.

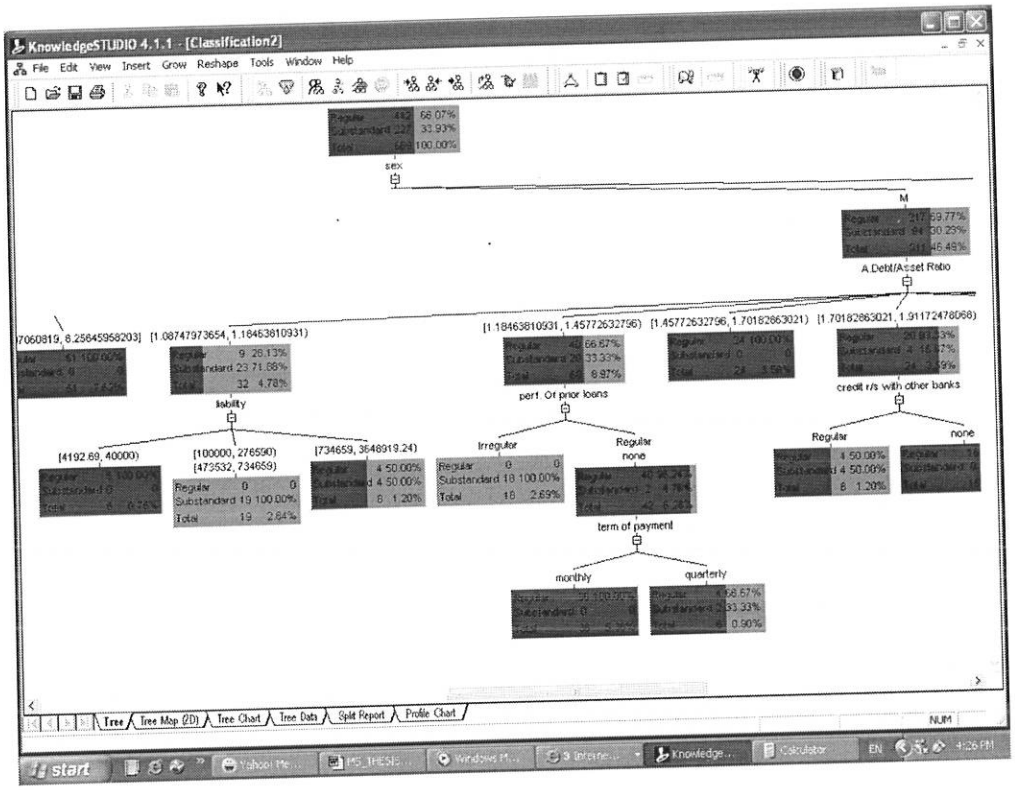


Figure 7: Output of the 1st Decision Tree

KnowledgeSTUDIO usually displays the most statistically significant variable or split at the root, but it was important to assess the variables that are used for splitting, since they might not be the most important variables to split the tree at the root. Therefore, based on expert opinion the first variable used to split the tree at the root, 'sex' was not considered as being the most important variable to consider the credit worthiness of customers, hence the next most statistically significant variable was taken.

By using the 'Next split' command that displays the next most statistically significant split, the researcher was able to obtain a tree, which according to the domain experts made use of a good variable for splitting the records at the root.

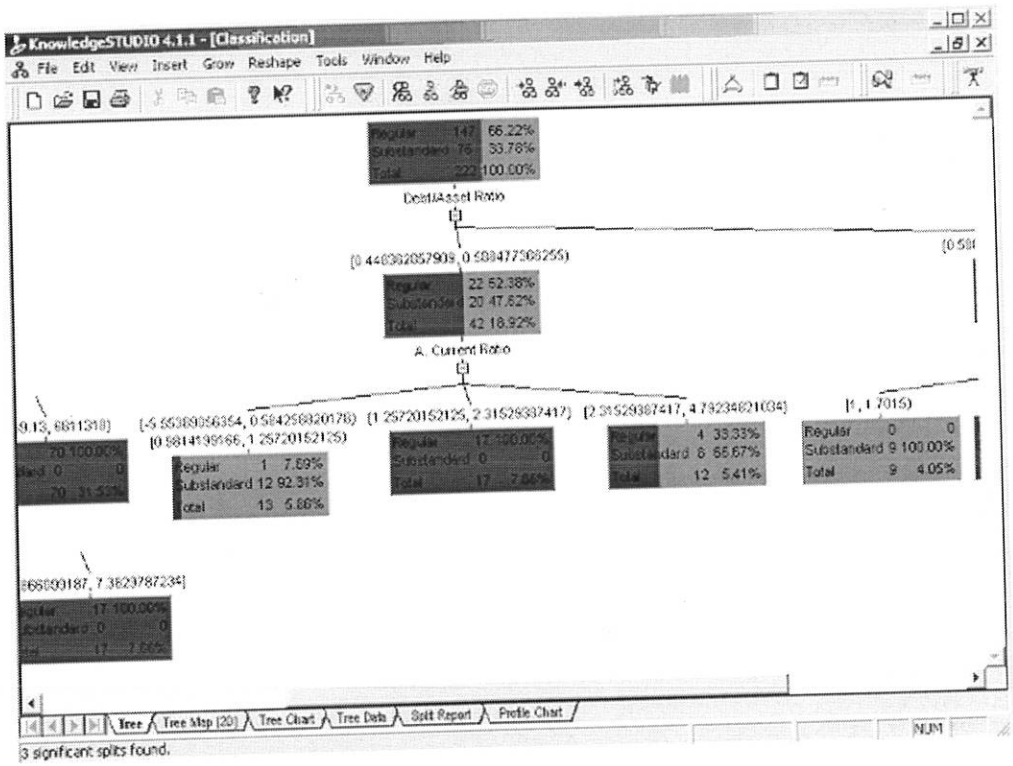


Figure 8: Output of the 2nd Decision Tree

After the tree was built, and rules are extracted and found to be mostly sound by domain experts, an evaluation of the results was made on the training set itself which showed a good accuracy, and lastly an evaluation was done on the testing set (25% of the record) that was set aside. The result of the validation of the decision tree in the form of confusion matrix is given below in the table 4.

Confusion Matrix - Classification

		Predicted	
		Regular	Substandard
Actual	Regular	140	7
	Substandard	8	67

Statistics

Total records	222
Correctly predicted	207
Percentage	93.24%

Table 4: Validation results from the second decision tree

The confusion matrix depicts that out of the total records provided to the software for testing (i.e 222), 140 and 67 records were classified correctly in the class of regular and

substandard respectively. On the other hand, 7 records were incorrectly classified as substandard while actually they were supposed to be in the regular class and 8 records were classified incorrectly as regular while actually they are in the substandard class. This result reveals that from the total records (i.e. 222), 207 were classified correctly while the remaining 15 records were classified incorrectly. Hence this indicates that records whose class is regular were classified with a minimum error as compared with the records in the class substandard, which is the risk type that the bank needs to minimize. In addition by tracing through the tree that has been constructed the researcher together with domain experts found that a variable utilized after the first split was not a good variable.

Although this training scheme has shown a good performance in terms of accuracy, the variables utilized for the decision tree construction were only 'debt/asset ratio', 'security/loan ratio', 'A. Current ratio', 'Net working capital', 'month', 'years in business', 'performance of prior loan's, and 'security value' from the total variables considered.

Even though some data mining algorithms will automatically ignore irrelevant variables and properly account for related (or covariant) columns [Edelstein, 1998], it is advised that in practice it is wise to avoid depending solely on a tool because often the knowledge of the problem domain may help in making the selection of the variables correctly. So by numerous discussions with the bank experts, eleven variables were selected from among the 28 originally considered for model building.

The variables selected were: 'Loan/time ratio', 'Amount granted', 'Asset', 'Net working capital', 'A. debt/asset ratio', 'A. current ratio', 'Security value', 'Security/loan ratio', 'Years in business', 'Debt to equity', and 'Debt/asset ratio'. But after careful consideration of the variables selected by domain experts, that were financial values, the researcher

incorporated some additional non-financial values because financial values alone should not be dependable to make decision. So the total number of variable selected from among the 28 were 16 which are: 'Loan no.', 'Month', 'Loan/time ratio', 'Amount granted', 'Asset', 'Net working capital', 'A. debt/asset ratio', 'A. current ratio', 'Debt to equity', 'Security value', 'Security /loan ratio', 'Trade sector', 'Years in business', 'Credit r/s with other banks', 'Performance of prior loans', and 'Performance in other types of loans'.

Hence, for further tests the researcher made use of the above selected variables. Several decision trees were built by varying the number and combination of the variables. Some of the results obtained from the experiments conducted are depicted in table 5 below:

Number of Variables utilized	Accuracy for the class Regular	Accuracy for the class Substandard	Total Accuracy
9	93.25%	74.25%	83.78%
7	94.13%	76.81%	85.47%
6	93.95%	91.75%	92.85%
8	90.84%	89.66%	90.25%

Table 5: Results of the Tests Conducted

In addition to the experiments conducted which are described above the researcher tried to conduct an experiment for each branch selected separately but the results obtained were not encouraging.

From additional experiments conducted, the following figure 9 shows the 'best' decision tree obtained by the use of only 7 variables out of the 16 that were considered for model building.

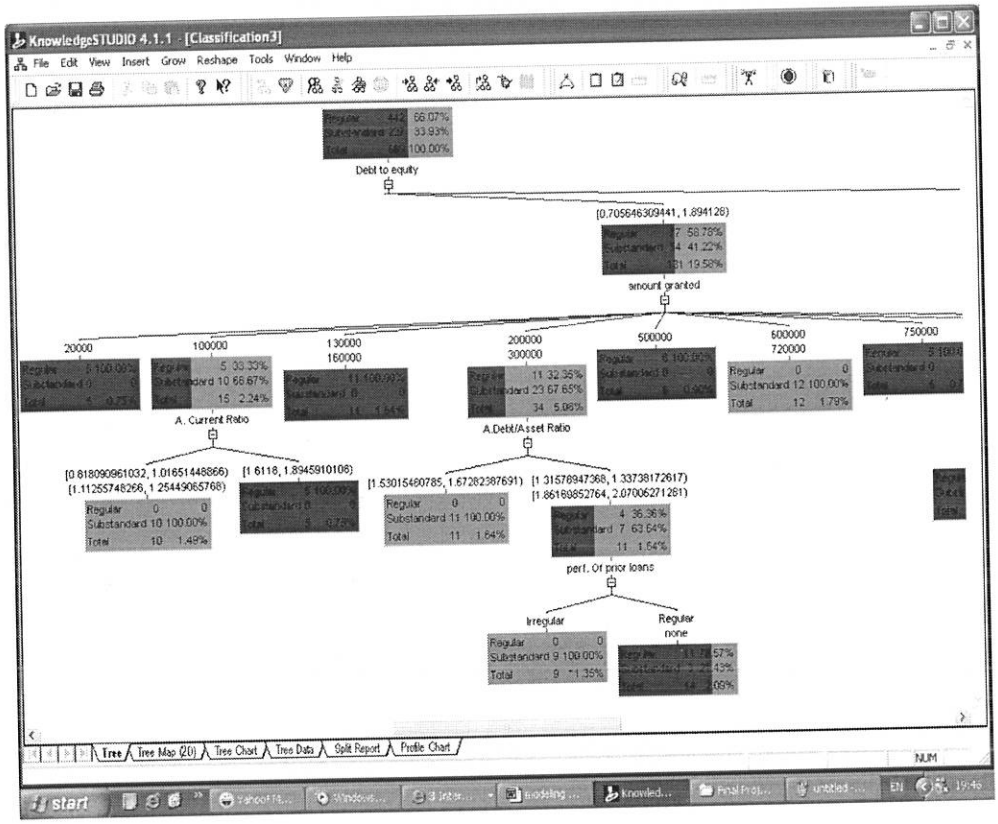


Figure 9: 'Best' decision tree with 7 variables

The decision tree constructed in this manner was validated on the training set and as it showed a good result on the dataset that it used to train it was validated on the testing set and the output confusion matrix in table 5 shows that out of the total 222, 208 (93.69%) were classified correct and (6.31%) 14 were classified incorrectly.

Validation Summary Report

Input Dataset: CREDIT_NIB1-VALIDATION

Created: Saturday, May 22, 2004 20:27:52

Confusion Matrix - Classification

		Predicted	
		Regular	Substandard
Actual	Regular	142	5
	Substandard	9	66

Statistics

Total records	222
Correctly predicted	208
Percentage	93.69%

Table 6: Validation results from the second decision tree

Even if the main aim of this section is identifying the soundest rules, consideration was also given to the misclassification costs for both credit risk[1] and commercial risk[2], as the classification of a bad debtor as a good one (i.e. credit risk) can have more severe consequences for bank than the classification of a good debtor as bad one.

Consequently from the confusion matrix depicted above we can see that 9 debtors who were considered as bad debtors were classified as correct while only 5 debtors which were good creditors were classified as bad creditors. The above decision tree had an accuracy rate of 96.5% for classifying regular creditors and only 88% for classifying substandard creditors. Thus, this indicated that records whose class is regular were classified with a minimum error as compared with the records in the class substandard. The above depicted decision tree was not considered as a good choice by the researcher because credit risk has a greater consequences for the bank than commercial risk.

Selection of a good model from subsequent experiments conducted was therefore made by taking into consideration the soundness of rules generated as well as the number of records which are substandard classified as regular (i.e. credit risk) as this type of error has to be minimized in a banking environment.

The 'best' possible model that was selected from among a number of experiments conduced is portrayed in the figure below.

[1] Credit risk: If a bad debtor is categorized as a good one
[2] Commercial risk: If a good debtor is categorized as a bad one

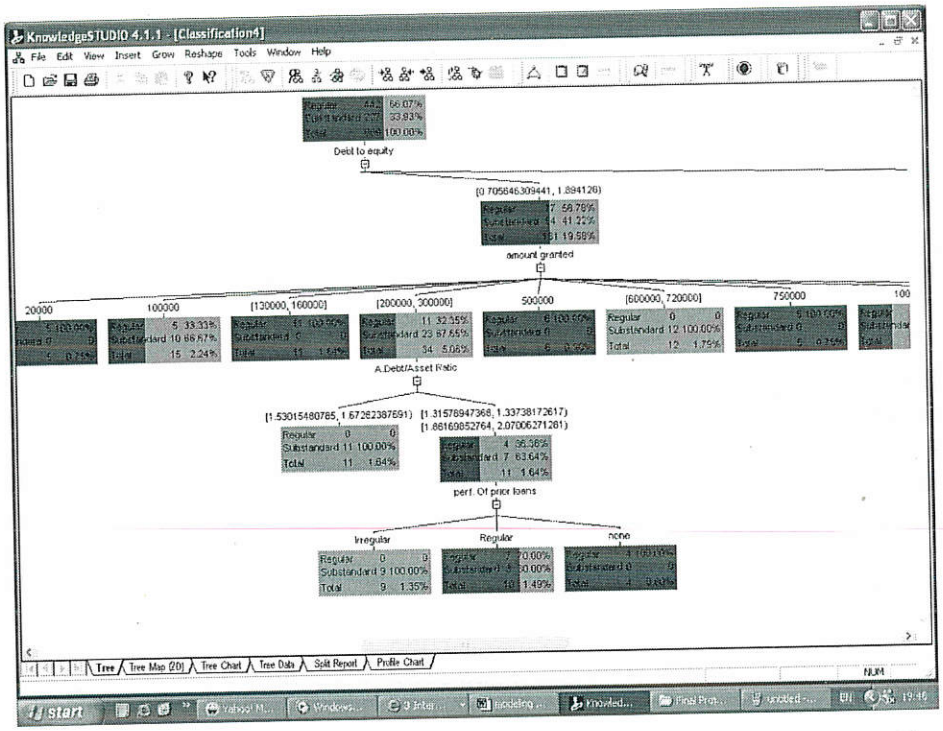


Figure 10: Final selected Decision Tree

The confusion matrix displayed in table 7 below show that 208 of the 222 records were classified correct.

Validation Summary Report

Input Dataset: CREDIT_NIB1-VALIDATION

Created: Tuesday, June 01, 2004 19:07:47

Confusion Matrix - Classification

		Predicted	
		Regular	Substandard
Actual	Regular	134	13
	Substandard	1	74

Statistics

Total records	222
Correctly predicted	208
Percentage	93.69%

Table 7: Validation of the selected decision tree

While comparing the results obtained from the decision tree in figure 9 and 10 the researcher could observe that the accuracy obtained is the same but the misclassification cost for the different classes were different.

The output of the tree in figure 10 is better explained in the confusion matrix (Table 8) below.

		Predicted		*Total*	*Accuracy Rate*
		Regular	**Substandard**		
Actual	**Regular**	134	13	147	91.16%
	Substandard	1	74	75	98.66%
Total		135	87	222	93.69%

Table 8: Output from the decision tree selected

The confusion matrix above shows that out of the total records provided to the program, 134 and 74 records were classified correctly in the class regular and substandard respectively. The accuracy rate for the prediction of substandard records is 98.66% while for the regular records it is 91.16%. This portrays that from the total records taken for testing 208 (93.69%) records were classified correctly while the remaining 14(6.31%) records were classified incorrectly. Hence, this indicated that records whose class is substandard were classified with a minimum error as compared with the records in the class regular, and because credit risk is the most costly risk for banks the result obtained from this model was considered a good one.

Although several attempts were made, it was not possible to obtain a good tree with a sound rule and an accuracy of more than 93.69%. Thus, the decision tree selected in figure 10 which is the one selected as a good one for the present research, classified 93.69% of the records correctly and 6.31% of records wrongfully.

4.5.3.2 Generating rules from Decision tree

From the decision tree developed in the aforementioned experiment, rules were developed by tracing through the branches upto leafs. KnowledgeSTUDIO (i.e. the software utilized) has the ability to generate rules from the decision tree constructed; it also allows the rules to be presented in many types. Some of the types supported are SQL, generic, Java and English. For the purpose of the present research the rules generated are in the generic type as they were found to be easily understandable. The following are some of the rules extracted from the decision tree selected and presented in figure 10, most of the rules extracted are attached as Annex 6.

RULE #1

```
if
      Debt to equity = [2.936075,7.97725844286]
then
      Classification = Regular  0.149425287356
      Classification = Substandard  0.850574712644
```

RULE #2

```
if
      Debt to equity = [0.25,0.5)
      A. Current Ratio = [1.11255748266,1.47739837398)
then
      Classification = Regular  0.666666666667
      Classification = Substandard  0.333333333333
```

RULE #3

```
if
      Debt to equity = [0.25,0.5)
      A. Current Ratio = [1.6118,8.25645958203]
then
      Classification = Regular  1
      Classification = Substandard  0
```

RULE #4

```
if
      Debt to equity = [0.25,0.5)
      A. Current Ratio = [1.11255748266,1.47739837398)
      trade sector = Clothes or Retail trade
then
      Classification = Regular  0
      Classification = Substandard  1
```

RULE #5

```
if
      Debt to equity = [0.705646309441,1.894128)
```

```
      amount granted = 20000
then
      Classification = Regular  1
      Classification = Substandard  0
```

RULE #6

```
if
      Debt to equity = [0.705646309441,1.894128)
      amount granted = 100000
then
      Classification = Regular  0.333333333333
      Classification = Substandard  0.666666666667
```

RULE #7

```
if
      Debt to equity = [0.705646309441,1.894128)
      amount granted = [200000,300000]
      A.Debt/Asset Ratio = [1.31578947368,1.33738172617) or
[1.86169852764,2.07006271281)
then
      Classification = Regular  0.363636363636
      Classification = Substandard  0.636363636364
```

RULE #8

```
if
      Debt to equity = [0.705646309441,1.894128)
      amount granted = [200000,300000]
      A.Debt/Asset Ratio = [1.31578947368,1.33738172617) or
[1.86169852764,2.07006271281)
      perf. Of prior loans = Regular
then
      Classification = Regular  0.7
      Classification = Substandard  0.3
```

The rules presented above indicate the possible conditions in which a loan record could be classified in each class.

The rules generated have indicated that attributes such as 'amount granted', 'performance of prior loans', 'credit relationship with other banks', 'trade sector', 'performance in other types of loans', 'month', 'A. Debt/Asset ratio', 'A. Current ratio', and 'debt to equity ratio' are found to be important variables for classification.

PHASE 2: Building a Predictive Model

In this phase a predictive model was built and tested based on the variables identified by the decision tree found to be relevant for the credit approval process. After assessing the

accuracy of the model, recommendation was made so as to make more experiments so as to be able to use the model developed to classify future data tuples or objects for which the class label is not known. To this end the researcher has tried to develop a prototype.

In building a predictive model the algorithm used to build the decision trees (i.e. KnowledgeSEEKER) and the measure (i.e. Adjusted P-value Bonferroni Adjustment measure) and the pruning method reduced error were used. The choice is based on the fact that reduced error method was found more appropriate for the present research.

The reduced error method start with a complete tree and run the test data through it, noting the number occurring in each value of the dependent variable at each node. For each non-leaf node, count the number of errors if the sub-tree is kept and the number if it becomes a leaf through pruning. The pruned node will often make fewer errors on the test data than the sub-tree makes. The difference between the numbers of errors (if positive) is a measure of the gain from pruning the sub-tree. From all the nodes, the pruning method chooses the one with the largest difference as the sub-tree to prune.

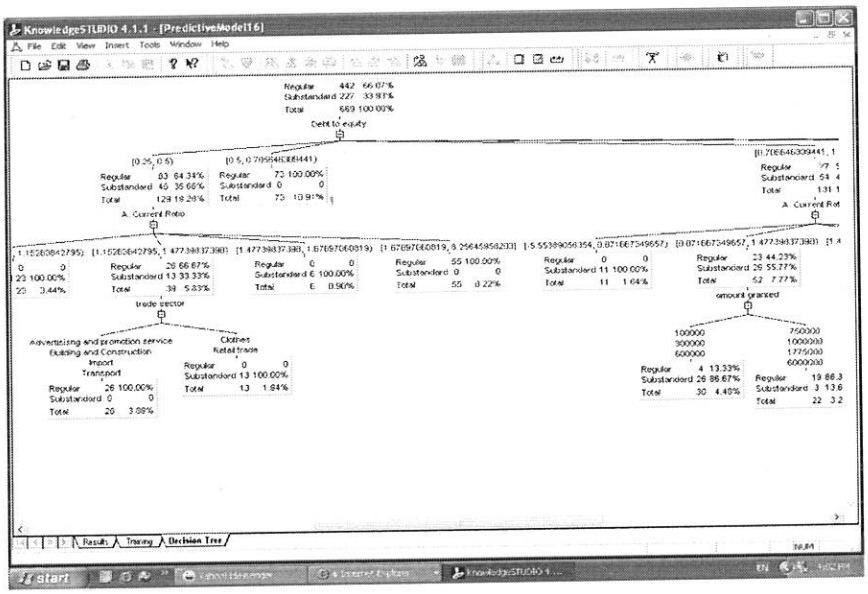


Figure 11: The predictive model developed for the 'best' decision tree

The predictive model built was evaluated, and the result obtained is given in table 9 below.

		Predicted	
		Regular	**Substandard**
Actual	**Regular**	4	1
	Substandard	0	5

Table 9: Result of the predictive model built

The predictive model built as shown in figure 11 while tested on 10 datasets (5 from each class) obtained from the bank which were not part of the training nor the testing dataset collected for the present research showed that out of the total, 1 record was wrongfully classified. One record, which was supposed to be in the regular class, was classified as substandard.

4.6 EVALUATION

During this phase, the degree to which the model meets the business objectives was assessed. Furthermore, the whole process has been reviewed and the next steps were determined.

The model-building phase was divided into two. The first phase comprised of building decision tree models using the KnowledgeSEEKER algorithm. This exercise, which was iteratively conducted, yielded various decision tree models that were used to generate rules.

The variables used for the decision trees constructed were those variables, which described a customer's credit worthiness. Each of the decision tree constructed were used

to generate rules that were analyzed in order to establish their business worth as meaningful rule sets that discovered information to help in the disbursement of loans.

The analysis, which was closely undertaken with domain experts, revealed that the variables that were used were good variables to classify borrowers into the predefined classes (i.e. regular and substandard). Each decision tree constructed was validated based on a dataset set aside for that purpose from the original dataset that was collected. This validation results were obtained in the form of confusion matrix. But decision of selecting the best decision tree was based on the soundness of the rules generated as well as the number of misclassified records of bad debtor as good ones.

The decision tree depicted in figure 10 was the 'best' from all the ones constructed in that it provided rule sets that were sounder to the problem domain. But the tree is by no means the ultimate tree, therefore it is the researcher's belief that more experimentation could yield more meaningful trees that yield more sound rules and also take into account the misclassification costs for credit risks as well as commercial risks. Time constraint was the major reason for not having built more models.

The validation test conducted on the results of the selected decision tree revealed that the accuracy of classifying debtors who are regular as regular is 91.16% while the accuracy of classifying debtors who are substandard as substandard is 98.66%, and the overall accuracy of the decision tree is 93.69%.

Once the model that generated the soundest rule to classify debtors as regular and substandard was selected, the next activity was to test the predictive accuracy of the model with the use of the independent variables identified as important and 'classification' as the

dependent variable so as to be able to use the model to classify future data tuples or objects for which the class label is not known.

Training results showed that the most important variables for the problem were, 'amount granted', 'trade sector', 'performance of prior loans', 'performance in other types of loans', 'credit relationship with other banks', 'month', 'anticipated debt asset ratio', 'anticipated current ratio' and 'debt to equity ratio'.

The variables that have been identified as being important by the researcher are more alike to the variables identified by Askale [2001], which were 'Amount Granted', 'Trade Sector','Month', 'Performance of Past loan','Anticipated Debt/Asset ratio', 'Anticipated current ratio', and the 'number of loans the borrower has settled in the past'. Even if the researcher was not able to make recommendation as to which techniques employed (i.e Neural network by Askale or decision tree in the present research) was best suited for the problem domain due to the fact that the records on which the experiments were conducted were not the same, the researcher was able to observe that some of the variable identified by the two works were the same and the accuracy that the research by the present research (93.69%) was greater than that obtained by Askale (88%). But further experiments need to be conducted to assess if the use of a combination of the techniques could yield better results or if one technique is better than the other.

The model built in the present research was intended to support the loan disbursement activity at NIB, therefore the researcher believes that the use of decision tree method is better as it can be easily explainable. The bank's credit analyst can easily explain the refusal of a request for loan with decision trees because all the reasons for the refusal can be traced back on the decision tree whereas if the model used is developed by the use of

neural networks explanation as to the reason for the refusal is not possible. All in all what should be noted is that no analysis technique can replace experience and knowledge of the business expert.

4.7 Model Deployment

The results obtained from experiments conducted were encouraging, and thus with further experiment the knowledge could be used for facilitating loan disbursement activity. According to experts' opinion, the rules generated from the decision tree that was selected as being the 'best' from those that were constructed were not all accurate. Further experimentation on a larger dataset should be made in order to derive more appropriate rules, which are good classifiers of potential borrowers as a good or bad credit risk so as to be able to predict the future outcome of records of borrowers for who's the outcome is not known. This has been included in the list of recommendations.

4.7.1 Preliminary Credit Risk Assessment: A Prototype

Data mining can be used for decision support applications [Bigus, 1996]. The decision tree generated can be used to test the probability of a borrower being a good or bad credit risk. In this study an attempt was made to develop a prototype named Nib International Bank S.C. : Preliminary credit risk assessment (figure 12) that uses the rules generated from the decision tree selected as being the 'best'. The model is deployed as part of the application where it is used to determine customers' credit risk probability. The prototype developed makes use of Visual Basic programming language.

The Visual Basic program can be seen as the main program. It functions as an interface for the user. In order to take advantage of the prototype developed the user should have all the information about his customer in a database. Then the user will calculate in order to

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 CONCLUSION

Up until recently, the ability to analyse and understand a huge volume of data lagged far behind the capability to gather, store and manipulate data. But the new generation of computerized methods is helping the endeavor of interrogating and analyzing very large datasets automatically and efficiently, thereby extracting useful information and knowledge, which are valuable for decision making.

The objective of this research undertaking was to explore the possible application of data mining technology at Nib International Bank, NIB, for developing a classification model. Such a classification model could support the bank experts at NIB in making decisions when granting loans. Such a model developed also could help in minimizing credit risks at the bank, and thereby increase the banks profitability.

Even if the methodology employed consisted of four basic steps; data collection, data preparation, model building, and evaluation and since a data-mining task is an iterative process, these steps were not followed strictly in linear order. Different problems created instance where there was a need to go back and forth between the different steps.

Since this research was intended to replicate a related research conducted by the use of neural network technique, some valuable experiences of the previous research were used. However, though the previous research's objective was to identify the possible application of neural network technique taking Dashen bank as a case study which obtained an accuracy of 88%, the possible application of other techniques like decision tree were not tested. Even though access to the same data on which the previous research was

To conclude, results from the study have shown that the problem in credit risk assessment could be supported by the use of data mining, in particular with the use of decision trees technique.

5.2 RECOMMENDATIONS

This research work has uncovered the potential applicability of data mining technology to support the loan disbursement activity at Nib International bank based on historical data accumulated on debtors. Thus, based on the findings of this research work, the researcher would like to make the following recommendations particularly in relation to the possible application of data mining technology in supporting credit disbursement activity at NIB as well as to other commercial banks.

Basically, this research was conducted for an academic purpose. However, the results of this study are found to be promising to be applied to address practical problems of credit disbursement. Hence, based on the findings of this study, the following recommendations can be forwarded:

- A data-mining task could only be efficient if the data it requires is available in electronic format. Therefore, for an efficient application of the data-mining task, accumulated records should be available in electronic format in order to facilitate the time required to key in the records, avoid errors in keying in data at the last minute as well as to facilitate the process of updating. To this end, Nib International Bank should take measures to store its records in an electronic format and to make all decisions based on collected records.
- Data mining techniques could contribute a lot in identifying potential customers that could be bad creditors. Thus, it could be more important to use data mining technique as a tool for the decision making process. In other words, Nib International Bank could optimise its credit assessment efforts by employing data mining technology.

- In this research work, an attempt has been made to assess the applicability of data mining technology to support credit risk assessment by using some set of variables that were considered important by experts. For a number of other variables, however, it remains to investigate further the effect of those variables so as to build models with better performance and accuracy than the models built in this research work.
- Further experimentation on a larger dataset should also be made in order to derive more appropriate rules, which are good classifiers of potential borrowers as a good or bad credit risk so as to be able to predict the outcome of records of borrowers.
- The present experiment provides a way to only classify borrowers as being regular or irregular potential creditors. Further experiments should be able to introduce more detailed classification of borrowers' category (i.e. Classify borrowers as regular, substandard, doubtful, and loss) so as to take measures based on the categories.
- Although in this study encouraging results were obtained, a small sample data was used for training and testing classifiers. Hence, it is appropriate to conduct the experiments with large training and testing datasets as well as making a number of trials to come up with more accurate and better performing classifiers.
- Decision tree and neural networks techniques were tested by a previous and the present research to support loan disbursement. The results of both showed promising results and hence data mining could be applied in the area of credit assessment. Therefore, it would be more optimal for commercial banks in Ethiopia to employ data mining to support loan disbursement activity.
- Further tests should also be made on the same dataset to see if neural network techniques or decision tree are the ultimate solution for credit risk assessment or if the combination of the two techniques will result in a better classification performance.

REFERECES

Angoss Software Corporation. (2001).Knowledge Studio Data Mining Software User Guide. Available URL: http://www.angoss.com

Askale. (2001). The application of Data Mining Technique in supporting loan disbursement activity at Dashen Bank S.C. A Thesis Submitted in Partial Fulfilment of the requirement for the Degree of M.Sc. I.S. Addis Ababa University: Addis Ababa.

Ballenger, Kitti, Claire de la Varre, and Sharon Yang.(1999) Data Mining. Available URL: http://www.ils.unc.edu/DataMining/OurClassPage.htm.

Basel Committee. (1999). Risk management principles for electronic banking. Available URL: http://www.bis.org/publ/bcbs98.htm

Berry, Michael J. A. and Linoff, Gordon. (1997). Data Mining Techniques: for Marketing, Sales, and Customer support. New york; John Willy& Sons, Inc.

Berson, A. & Smith, S. & Thearling K. (1999). An overview of data mining techniques.

Bhatnagar, R.G. (2001). Risk management by commercial banks- Time to hammer out the chinks. Financial Daily. 2001. Available URL: http://www.thehindubusinessline.com/businessline/2001/08/02/stories/040208ma.htm

Bigus, J.P. (1996). Data Mining With Neural networks in Solving Business Problems-From Application to Development to Decision Support. New York: McGraw-Hill.

Brand, E. & Gerritsen, R.(1998). Classification and Regression. Available at URL: http://www.dbmsmag.com.

Cabena, P., et. al. (1998). Discovering Data Mining - From concept to Implementation, printice Hall, New Jersy.

Carbone, P. (1997). Data Mining or "Knowledge Discovery In Databases": An Overview. Available URL: http://www.mitre.org/pubs/data_mgt/Papers/DMHdbk.pdf.

Chen, M.S., et al. (1997). Data Mining : An Overview from Database Perspective. National Taiwan University, Taipei, Taiwan.

CRISP-DM. (2000).CRISP-DM 1.0: Step-by-step data mining guide. Available URL: http://www.crisp-dm.org.

Denekew Abera. (2003). Application of data mining techniques to support customer relationship management at Ethiopian Airlines. A Thesis Submitted in Partial Fulfilment of the requirement for the Degree of M.Sc. I.S. Addis Ababa University: Addis Ababa.

Deogun, Jitender S. (2001). Data Mining: research Trends, Challenges, and Applications. Available URL: http://citeseer.nj.nec.com/deogun97data.html

Edelstein, Herb. (1998). Data mining-Let's Get Practical: How to identify a strategic problem statement, prepare the right data, and build and apply a robust model.Database Programming & Design Magazine. Available URL: http://www.db2mag.com/98smEdel.htm

Fabris, Peter. (1998). Data mining. CIO Magazine, 'Advanced Navigation.' Available URL: http://www.cio.com/archive/051598_mining_content.html

Fayyad, Usma, Piatetsky-shpiro, G. and Smyth, P. (1996). From Data Mining to knowledge Discovery in Databases. Available URL: http://citeseer.nj.nec.com/fayyad96from.html

Gnardellis, T. & Boutsinas, B.(n.d.). An experimenting with data mining in education. *Available URL:* http://cs.hbg.psu.edu/comp594/paper/GB.pdf

Gobena, Mikael. (2000). Flight Revenue Information Support System for Ethiopian Airlines. A Thesis Submitted in Partial Fulfilment of the requirement for the Degree of M.Sc. I.S. Addis Ababa University: Addis Ababa.

Han, Jiawei and Kamber, Micheline. (2001). Data Mining: concepts and Techniques. San Fransisco; Morgan kufman Publishers.

Henock, Woubshet. (2002). Application of data mining techniques to support customer Relationship management at Ethiopian airlines. Masters Thesis Addis Ababa University, Addis Ababa.

Ken, et.al. (1999). A methodology for evaluating and selecting data mining software. Available URL: http://csdl.computer.org/comp/proceedings/hicss/1999/0001/06/6009.pdf

NIB. (2003). 4th annual proceedings. Available URL: http://www.addischamber.com/bcontact/nib/major.htm

Peterson, Pamela P.(1994). Financial management and analysis. McGraw-Hill. America.

Shegaw Anagaw. (2002). Application of data mining technology to predict child mortality patterns: The case of (Butajira Rural Health Project) BRHP. A Thesis Submitted in Partial Fulfilment of the requirement for the Degree of M.Sc. I.S. Addis Ababa University: Addis Ababa.

Tesfaye, Hintsay. (2002). Predictive Modeling Using Data Mining Techniques In Support to Insurance Risk Assessment. A Thesis Submitted in Partial Fulfilment of the requirement for the Degree of M.Sc. I.S. Addis Ababa University: Addis Ababa.

Witten, I. and Frank, E. (2000). Data mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco: Morgan Kaufmann publishers.

Han, Jiawei and Kamber, Micheline. (2001). Data Mining: concepts and Techniques. San Fransisco; Morgan kufman Publishers.

IFCI. (2000).Source of Risk, Overview: Credit Risk. Risk Institute. *Available URL:* http://riskinstitute.ch/00013403.htm

Kannan, R.(2003). Indian banking today and tomorrow- Risk Assessment and risk management. *Available URL:* http://www.geocities.com/kstability/inbank/risk/project-map.html

Kononenko, I. and Hong, S. J. (1997). Attribute Selection for Modeling. At URL: http://www.research.ibm.com/dar/papers/pdf/gcshong_with_cover.pdf

Levin, Nissan and Zahavi,J. (1999). Data Mining. Available URL: www.urbanscience.com/Data_Mining.pdf

Lloyd - Williams, Michael. (1997). Discovering the Hidden secrets in your Data - the data Mining approach to Information. Available URL: http://informationr.net/ir/3-2/paper36.html

Luan, Jing.(2002). Data mining and knowledge management in higher education: potential applications.

Available URL: http://www.cabrillo.edu/services/pro/oir_reports/DM_KM2002AIR.pdf

Lulseged Teferi.(2003) Private banks gain ground in Ethiopia. The Banker.

Available URL: http://www.thebanker.com/news/fullstory.php/aid/337/Private_banks_gain_ground_in_Ethiopia.html

Raghavan, V. & Deogun, J. & Sever. (1997) H. Data mining: Trends and Issues

Straub, J. (2000). The Agile Manager's Guide to: Understanding financial Statements. Velocity Business Publishing, Inc.: New Delhi.

Thearling, Kurt. (1997). Understanding data mining: it's all in the interaction.

Thearling ,K. (2003). An introduction to data mining. http://www3.shore.net/~kht/text/dmwhite.htm.

Timewell, Stephen. (2003) Private banks gain ground in Ethiopia. The Banker.

Available URL: http://www.thebanker.com/news/fullstory.php/aid/337/Private_banks_gain_ground_in_Ethiopia.html

Two Crows Corporation. (1999). Introduction to Data Mining and knowledge discovery. Available URL: http://www.twocrows.com/

Wasserman, Miriam. (2000). Mining Data. Available URL: http://www.bos.frb.org/economic/nerr/rr2000/q3/mining.htm

GLOSSARY OF TERMS

Asset: Anything of value owned by an organization

Balance Sheet: A financial statement that reveals the value of assets, liabilities, and equity. The balance sheet equation is: asset are equal to the sum of liabilities plus owner's equity.

Capital: It represents the value or net worth of the organization. (Capital is equal to assets less liabilities)

Current Asset: Cash, marketable securities, accounts receivables, and inventories, which in the normal course of business will be turned into cash within a year.

Current Liability: Liability to be paid to creditors within a year.

Collateral (Security): Assets that are pledged or mortgaged to secure a loan, thereby reducing risk to the lender.

Current Ratio: This is current asset divided by current liabilities. Current ratio is a measurement of an enterprise to meet its short-term debt.

Debt to Asset Ratio: This is measured by dividing total liabilities by total asset.

Debt to Equity Ratio: This ratio is measured by dividing total liabilities by stockholder's equity.

Income Statement: A financial Statement that reveals revenues and related expenses together with the resulting income or loss. Additionally, extraordinary revenue and expenses would be shown following operating income (or loss).

Liabilities: Amounts that are owed to creditors.

Term loan: Generally, a bank loan would be considered as a term loan if it's duration life is more than one year. A term loan is usually repaid with an amortization schedule with monthly or quarterly payments.

Net Working Capital: Current assets minus current liabilities. Net working capital is a measurement of an enterprise to meet its short-term debt with its current asset.

Annex 1: Check List

NIB INTERNATIONAL BANK S.C.

CHECK LIST

CHECK LIST

1. **Full name and address of the applicant (borrower)**
- Name and address of the business
- Owner of business and amount of investment in the business
- Applicant letter stating why the loan is requested, how it is going to be paid, what collateral to be offered.
- Marital status of applicant

2. **Trade license to operate the business renewed for the current year**
- Investment license for projects
- Power to borrow: Article and memorandum of association (where applicable)
- Date of establishment of the business

3. **Business profile and description of the business**
e.g.: a) export, import, wholesale
b) Management of the business – qualitative observation such as strength or weaknesses of the management, experience, education, number of employees…

4. **Amount and type of finance requested**
- Overdraft, term loan, merchandise
- Contribution of applicant towards the business
- Intended period of loan, mix of the loan

5. **Purpose of the loan**
- Working capital, purchase of equipment machineries, inputs
- Business plan
- Competitive advantage (if any)

6. **Type and value of security offered**
- Building, merchandise, vehicles, machinery, etc.
- What safety margin is set? Compare with bank's policy
- Ownership document title to property, ownership booklet for trucks, machineries, and deposit book let or certificate, etc…

7. **Credit information**
- With other banks
- Others

8. **Financial statements where applicable, preferably audited**
- Balance sheet
- Profit and loss
- Note to the accounts and verification against the nature of business
- Cash flow for two to three years for existing business or at least for a period of
- term loan applied
- In the case of project, over the project life with feasibility study

CHECK LIST

9. Guarantees

- Bid Bond guarantee
- Advance payment guarantee
- Performance guarantee
- In all case letter of request is required
- Evidence indicating purpose such as contracts envisaged is necessary
- Collateral to be offered vis-à-vis bank policy

10. Processing imports applications

- Valid import license/ investment license
- Suppliers/manufacturers Proforma invoices
- Import clearance certificate from NBE
- Insurance coverage from the goods to be imported
- L/C application in case of import on letters of credit basis or P/O in case of import as CAD basis
- For certain goods additional approval must be obtained from pertinent government institutions
- Type of goods must be free from restriction
- Country of origin of goods must be free from embargo
- Maintenance of account with NIB

11. Margin facility

- 100% margin paid
- L/C margin facility available or to be made available

12. Internal procedure for advising of arrival of import documents

- Checking of documents for compliance with terms and conditions
- Immediate advice for clean and/or discrepant documents
- In case of clean documents pass entries whereas in case of discrepant documents advise customer and keep document in custody
- In case of clean documents, convert foreign currency to local currency, debit margin held, debit advance on import bills and credit correspondent. When payment is made, debit customers account including interest and release documents

13. Process of export for valuable goods

- Valid export license
- Sales contract, proforma invoices and export and customs declarations
- Export and customs declarations
- Export letter of credit
- Exports can be in the mode of

a) Letter of credit, advance payment and consignment basis on exceptional cases

- All commercial banks can allow exports for goods other than coffee against submission of
    a) valid foreign trade license for export
    b) Customs declaration 5 copies
    c) Banks declaration 6 copies
    d) Proforma Invoices

14. Procedures for advances on export bills

- Application letter
- All documents approved for the export of the goods for which advance is sought
- Customs exit certificate
- Certificate of cleanness and grade or quality
- Truck-way-bill from warehouse to port of loading
- All risk insurance coverage for the goods from warehouse to port of loading

Annex 2: Financial Credit Report Form

Nib International Bank S.C.

የፋይናንስ መግለጫ ቅጽ

FINANCIAL CREDIT REPORT

ቅርንጫፍ: Branch_____ ቀን Date _____

1. የአመልካች ስም/ የዋስ ስም NAME OF APPLICANT/GUARANTOR _____ ዕድሜ AGE _____

2. የሚስት/የባል ስም NAME OF WIFE/HUSBAND _____

3. አድራሻ ADDRESS: ከተማ TOWN _____ ወረዳ WREDA _____ ቀበሌ KEBELE _____ የቤት ቁ HOUSE No _____ የስልክ ቁ PHONE No. _____ ፖ.ሣ.ቁ. P.O. BOX _____

የሥራ BUSINESS _____

የመኖሪያ ቤት RESIDENCE _____

የፈቃድ ቁጥር ከ ውስጥ ንግድ ሚኒስቴር LICENCE No. MINISTRY OF DOMESTIC TRADE _____ የውጪ ንግድ ሚኒስቴር MINISTRY OF FOREIGN TRADE _____

የማዘጋጃ ቤት MUNICIPAL _____ ሌሎች (ከሆነ) OTHERS (ካለ ሆነ) _____

የኅብረት ሥራ ማኅበር የመዘገበው ድርጅት HASIDA _____ ሌሎች OTHERS _____

አመልካች የተፈቀደለት የሥራ ዓይነት(ዝርዝር) 5. THE TYPES OF BUSINESS THE APPLICANT IS LICENCED TO OPERATE

_____

_____

6. አመልካች የተቋቋመበት ዘመን DATE ESTABLISHED _____ መነሻ ካፒታል በብር ORIGINAL INVESTMENT BR _____ አሁን ያለው ካፒታል በብር PRESENT CAPITAL BIRR _____

7. የሥራተኛ ብዛት NUMBER OF EMPLOYEES: ቋሚ PERMANENT _____ ጊዜያዊ TEMPORARY _____

8. የተጠየቀው የብድር ዓይነትና የገንዘብ ልክ ብር TYPE AND AMOUNT OF FACILITY APPLIED FOR _____

9. ብድሩ የተፈለገበት ምክንያት (በዝርዝር) PURPOSE OF LOAN (Specify) _____

10. የዋስትናው ዓይነትና ግምት TYPE AND VALUE OF SECURITY OFFERED _____

11. የብድሩ አከፋፈል (ፅቅድ) PROPOSED MODE OF REPAYMENT _____

12. አመልካች ከባንኩ ጋር ያለው ግንኙነት፡ በአስቀማጭነት፣ በዋስትና as DEPOSITOR _____ as GUARANTOR _____ የቅርንጫፍ ባንክ ስምና የሂሣብ ቁጥር NAME OF BRANCH BANK A/C No. _____ ብድሩን የሰጠ የቅርንጫፍ ስም NAME OF LENDING BR _____

13. የቀድሞ ብድሮች ሁኔታ በቅርንጫፉ ኃላፊ የሚሞላ RECORD OF PREVIOUS LOANS (to be filled by branch)

Loan No.	Amount in Birr	Date Granted	Date Settled	REMARKS
1.				
2.				
3.				
4.				
5.				

የሂሣብ ሠንጠረዥ

BALANCE SHEET

AS AT ———

LINE No.	ሐ ብ ት ASSETS	የሥሚክሮት ንብረት ልክ DECLARED ቀን Date ———	ታይቶ የተረጋገጠው ንብረት ልክ CHECKED Date ———	KEY
1	ጥሬ ገንዘብ፣ በባንክ ያለ Cash a) in bank			
	በእጅ ያለ b) On hand			
2	ወደፊት የሚሰበሰብ ሀሰቃ ሽያጭ Receivables a) Accounts			
	ከተሰሩ ሠነዶች b) Notes			
3	በሱቅ ወይም በመጋዝን ያለ የሸቀጥ መጠን Goods in Stock			
4	ለዕቃ ግዥ የተደረገ የቅድሚያ ክፍያ Prepayment on Merchandise			
5	ተንቀሳቃሽ ሐብት CUURENT ASSETS			
6	የፋብሪካ ዕቃዎችና መሣሪያዎች Equipment and machinery			
7	ተሸከርካሪ Motor Vehicles			
8	የቤትና የቢሮ ዕቃዎች Furniture & Fillings			
9	ቤቶች Buildings			
10	ሌላ ተጨማሪ ሐብት Other Assets			
11	ቋሚ ሐብት (የተጣራ) FIXED ASSETS (Net)			
12	ጠቅላላ ሐብት TOTAL ASSETS BIRR			
	ዕዳ LIABILITIES			
13	የሚከፈል ዕዳ፣ ለዕቃ ግዥ Payable: a) Accounts			
	የተሰሩ ሰነዶች b) Notes			
14	ለግብር የሚከፈል Tax payable			
15	የባንክ ብድር Bank Loans			
16	ከረዥም ጊዜ ዕዳዎች በዚህ ዓመት የሚከፈል Current portion, Long Term debt			
17	ሌላ ዕዳ Others			
18	በቅርብ የሚከፈል ዕዳ (ጊዜያዊ ዕዳ) CURRENT LIABILITY			
19	በረጅም ጊዜ የሚከፈሉ ዕዳዎች Long Term Debts			
20	ካፒታልና መጠባበቂያዎች Capital & Reserves			
21	የዕዳና ካፒታል ድምር TOTAL LIABILITIES & CAPITAL BIRR			

ቦ ገ ቢ ና የ ወ ጪ ሠ ን ጠ ረ ዥ

LINE No.	ገቢ INCOME	AMOUNT	KEY
22	ሽያጭ Sales		
23	የተሸጠው ዕቃ ዋጋ Cost of Goods Sold		
	መነሻ ዕቃዎች Beginning inventory, as at Date		
	ሲደመር የተገዙ ዕቃዎች Add purchases for the period		
	ሲቀነስ መጨረሻ የተሩ ዕቃዎች Less ending inventory, as at		
24	ሌሎች Others		
25	ያልተጣራ ትርፍ Gross Profit		
	ወ ጭ ዎ ች EXPENSES		
26	ደመወዝ Wages & Salaries		
27	የንግድ ቤት ኪራይ Business Premises Rent		
28	የፍጆታ መገልገያዎች ወጭ Utilities		
29	ማደሻና ጥገና Maintenance & Repair		
30	መድን Insurance		
31	የእርጅና ተቀናሽ Depreciation		
32	የግል (የግል መኖሪያ ቤትን ከራይ ይጨምራል) Personal (including residential rent)		
33	ሌሎች Others		
34	ጠቅላላ ወጭዎች TOTAL EXPENSES		
35	ትርፍ ከግብር በፊት INCOME before tax		
36	ግብር TAXES		
37	የተጣራ ትርፍ ከግብር በኋላ INCOME after tax		
38	በዋስትና ያለበት ሰዉ Guarantee Liability (as below)		

የአመልካች ፊርማ
Applicant's Customer Signature

	ሬሾ ትንተና ናሙናዎች ANALYTICAL & COMPARATIVE RATIOS	ዘንድሮ ዓመት This Year	ባለፈው ዓመት Last Year	
39	የተጣራ መንቀሳቀሻ ካፒታል Net Working Capital			
40	ተንቀሳቃሽ ሀብት በፈሳሾ ዕዳ ማካፈያ Current Ratios			
41	የሽያጭና የልዩ ማካፈያ Sales to Receivable Ratio			
42	የሽያጭ የመንቀሳቀሻ ሀብት ማካፈያ Sales to Current Asset Ratio			
43	ትርፍ ከተንቀሳቃሽ ሀብት ማካፈያ Income Before Tax as % of Current Asset			
44	ጠቅላላ ዕዳ ከጥሬ ሀብት ማካፈያ Total Debt to Worth Ratio			

ተረጋግጧል
Processed by

Annex 3: Loan Approval Form

Nib International Bank S.C.
Loan Approval Form

CONFIDENTIAL

Code No. _____
Category _____
Trade License No. _____

BRANCH _____

LAF No. _____

Date _____

1. **Name of Applicant** _____
2. **Facility Applied For** _____
3. **Purpose of Facility** _____
4. **Applicant's Business** _____

5. Present Facilities	Limits Approved	Present Balance	Date Granted	Expiry Date	Interest Rate	Loan Status
Overdraft Facility	.....	.....	.....	.....	.....	.....
Temporary Overdraft Facility	.....	.....	.....	.....	.....	.....
Term loan at Birr.....Per month/qrt.	.....	.....	.....	.....	.....	.....
Merchandise Loan Facility at...% Adv.	.....	.....	.....	.....	.....	.....
Adv. On Imp/Exp. Bills at ...% Adv.	.....	.....	.....	.....	.....	.....
Sub Total	.....	.....	.....	.....	.....	.....
Letters of Credit at _____margin (sight)	.....	.....	.....	.....	.....	.....
Guarantees Issued by the Bank (Foreign)	.....	.....	.....	.....	.....	.....
Guarantees Issued by the Bank (Local)	.....	.....	.....	.....	.....	.....
" " " " " (Local)	.....	.....	.....	.....	.....	.....
" " " " " (Local)	.....	.....	.....	.....	.....	.....
" " " " " (Local)	.....	.....	.....	.....	.....	.....
Grand Total	.....	.....	.....	.....	.....	.....

6. Description of Security Offered	Evidenced by	Value
For Existing Facilities		
For New Request		

7a. C/A ____ Opened on _____ Balance of Birr. ____
7b. Turnover of OD or C/A This year Birr ____ Last year Birr _____

SWING			SWING
HIGHEST DEBIT	**HIGHEST CREDIT**	**HIGHEST DEBIT**	**HIGHEST CREDIT**
1)	1)	1)	1) .....
2)	2)	2)	2) .....
3)	3)	3)	3) .....
LOWEST DEBIT	**LOWEST CREDIT**	**LOWEST DEBIT**	**LOWEST CREDIT**
1)	1)	1)	1) .....
2)	2)	2)	2) .....
3)	3)	3)	3) .....

8. I) Applicant's Net **Monthly**/ Yearly Income_____II) Guarantor's Net **Monthly**/ Yearly Income Br. _____
9. Import and Export figures (Show Separately amount authorized and utilized for two years)

Birr _____ Birr _____ Birr _____ Birr _____
Birr _____ Birr _____ Birr _____ Birr _____

10. **OTHER LIABILITIES TO THE BANK AND THIRD PARTIES (INDICATE IF ANY OVERDUE)**

BORROWER	GUARANTOR

11. SUMMERY OF THE FINANCIAL STATEMENTS OF THE BORROWER AND GUARANTOR

(Borrower in 000's Birr)			(Guarantor in 000's Birr)		
Particulars:	This year	Last Year	Particulars:	This year	Last Year
Net Working Capital	Birr	Birr	Net Working Capital	Birr	Birr .....
Capital & Reserves	Birr.....	Birr.....	Capital & Reserves	Birr.....	Birr.....
Profit/ Loss	Birr.....	Birr.....	Profit/ Loss	Birr.....	Birr.....

12. **FACILITIES WITH OTHER BANKS** (Give details In case of absence of Information specify efforts made)

13. **GENERAL REMARKS AND SUMMARY OF PAST RECORDS:**

14. **RECOMMENDATION OR DECISION OF THE BRANCH CREDIT COMMITTEE** (Give reasons for declining)

15. **DECISION OF HEAD OFFICE MANAGEMENT CREDIT COMMITTEE** (Give reasons for declining or for deviation from the Recommendation)

16. **DECISION OF THE BOARD** (Give reasons for declining or for deviation from the Recommendation)

Annex 4: Monthly loans and Advances Return Form

No.	Name of Borrower	Amount Granted	Date Granted	Due date	Periodic repayment		Amount outstanding		Arrears		During the Month		Other	Collateral		Remark
						Amount	Current	Over Due	Arrears amount	No, of days in arrears (past due)	Amount Disbursed	Amount Collected		Type	Value	

Annex 5: List of Independent variables (inputs) used for model building along with their descriptions

No.	Name of Variable	Data type	Description	Source
1	Branch	Text	Name of the branch from where the loan the loan was given	Loans and advances return form
2	Loan No.	Number	The number of loan for the specific borrower (1st loan, 2nd loan, 3rd loan etc.)	Computed
3	Month	Date	Month on which loan was granted	Computed
4	Duration	Number	The duration of loan in number of days	Computed
5	Yearly Payment	Number	The estimated amount to be paid in a year	Computed
6	Amount Granted	Number	Amount granted	
7	Loan/Time Ratio	Number	Loan amount divided by the loan duration	Computed
8	Asset	Number	Total asset of the borrower	Financial Credit Form
9	Capital	Number	Total capital of the borrower	Financial Credit Form
10	Current Asset	Number	Total current asset of the borrower	Financial Credit Form
11	Current Liability	Number	Total current liability of the borrower	Financial Credit Form
12	Net working capital	Number	Current asset - Current Liability	Computed
13	Liability	Number	Total liability of the borrower	Financial Credit form
14	Debt/Asset Ratio	Number	Liability value divided by asset value	Computed
15	A. Debt/Asset Ratio	Number	The anticipated debt/asset ratio after considering the new loan to be granted	Computed
16	A. Current Ratio	Number	The anticipated current ratio after considering the new loan to be granted	Computed
17	Debt/ Equity ratio	Number	Liability value divided by capital value	Computed
18	Security type	Text	Type of security (e.g. building, vehicle, personal guarantee)	Loan approval form
19	Security value	Number	Estimated value of the security	Loan approval form
20	Security/Loan Ratio	Number	Security value divided by the amount granted	Computed
21	Sex	Text	Sex of the borrower	Loan approval form
22	Trade Sector	Text	The kind of business borrower is	Loan Approval

			engaged in	form
23	Years in business	Number	The number of years the borrower has been in business	Computed
24	Term of payment	Text	the term of payment (e.g. monthly, bimonthly or quarterly)	Loans and advances return form
25	No of prior loans	Number	The number of loans borrower has settled in the past	Loan approval form
26	Per. Of prior loans	Text	Performance of past loans (i.e. whether past loans were regular or not)	Loans and advances return form
27	Per. In other types of loans	Text	Performance of past loans (not in term loans)	Loans and advances return form
28	credit relationship with other banks	Text	Credit relationship with other banks in the country (i.e. whether they were regular or not or even if there were none)	Loans approval form

Annex 6: Format for collecting Borrower's data with hypothetical records

Loan No.	Branch	Purpose of Loan	date granted	duration/ expiry date	amount granted	Total Asset	Capital	current asset
5	Abinet	Working capital	21/06/2001	30/07/2002	3.000.000,00	13.945.689,00	3.000.000,00	2.361.667,00
2	Adarash	Working capital	7/11/2000	3/04/2001	100.000,00	700.000,00	600.000,00	335.773,23
10	Main	Working capital	13/09/2001	12/09/2002	2.000.000,00	46.439.198,00	16.310.000,00	29.891.145,00

current liability	liability	security type	security value	sex	trade sector	establishment year	term of payment	nb. Of prior loans settled
10.754.659,00	10.945.689,00	BLD	30.143.784,67	CP	Building and Construction	1992	quarterly	3
100.000,00	100.000,00	BLD	110.000,00	M	Import	1990	bimonthly	3
30.129.198,00	30.129.198,00	BM	2.684.080,00	CP	Building and Construction	1991	quarterly	2

perf. Of prior loans	perf. In other types of loans	credit r/s with other banks	Month	Duration	Yearly payment	Security/Loan Ratio	Loan/Time ratio	A. Debt/Asset Ratio
Regular	Regular	Regular	June	404	2.710.396,04	10,0	7426	0,8
Regular	Regular	none	November	148	246.621,62	1,1	676	0,1
Regular	Regular	Regular	September	365	2.000.000,00	1,3	5479	0,6

Years in business	Net Working Capital	A.Current Ratio	Debt to equity	Classification
4	-8.392.992,00	0,22	3,6	Substandard
6	235.773,23	3,36	0,2	Regular
5	-238.053,00	0,99	1,8	Substandard

Annex 7: Rules Extracted

```
Generic Rules extracted

RULE #1
 (Whole Tree)
       Classification = Regular  0.660687593423
       Classification = Substandard  0.339312406577

RULE #2
if
       Debt to equity = [0.00404761392411,0.100723194379)
then
       Classification = Regular  1
       Classification = Substandard  0

RULE #3
if
       Debt to equity = [0.15,0.25)
then
       Classification = Regular  0.926829268293
       Classification = Substandard  0.0731707317073

RULE #4
if
       Debt to equity = [0.25,0.5)
then
       Classification = Regular  0.643410852713
       Classification = Substandard  0.356589147287


RULE #5
if
       Debt to equity = [0.705646309441,1.894128)
then
       Classification = Regular  0.587786259542
       Classification = Substandard  0.412213740458

RULE #6
if
       Debt to equity = [1.894128,2.936075)
then
       Classification = Regular  0.407407407407
       Classification = Substandard  0.592592592593

RULE #7
if
       Debt to equity = [2.936075,7.97725844286]
then
       Classification = Regular  0.149425287356
       Classification = Substandard  0.850574712644

RULE #8
if
       Debt to equity = [0.100723194379,0.15)
       Month = August, December, March, May or October
then
       Classification = Regular  1
       Classification = Substandard  0

RULE #9
if
       Debt to equity = [0.100723194379,0.15)
       Month = January or September
then
```

```
        Classification = Regular  0
        Classification = Substandard  1

RULE #10
if
        Debt to equity = [0.15,0.25)
        Month = April, August, December, February, January, March,
November or September
then
        Classification = Regular  1
        Classification = Substandard  0

RULE #11
if
        Debt to equity = [0.15,0.25)
        Month = June
then
        Classification = Regular  0
        Classification = Substandard  1

RULE #12
if
        Debt to equity = [0.25,0.5)
        A. Current Ratio = [0.818090961032,1.11255748266)
then
        Classification = Regular  0
        Classification = Substandard  1

RULE #13
if
        Debt to equity = [0.25,0.5)
        A. Current Ratio = [1.11255748266,1.47739837398)
then
        Classification = Regular  0.666666666667
        Classification = Substandard  0.333333333333

RULE #14
if
        Debt to equity = [0.25,0.5)
        A. Current Ratio = [1.6118,8.25645958203]
then
        Classification = Regular  1
        Classification = Substandard  0

RULE #15
if
        Debt to equity = [0.25,0.5)
        A. Current Ratio = [1.11255748266,1.47739837398)
        trade sector = Advertising and promotion service, Building and
Construction, Import or Transport
then
        Classification = Regular  1
        Classification = Substandard  0

RULE #16
if
        Debt to equity = [0.25,0.5)
        A. Current Ratio = [1.11255748266,1.47739837398)
        trade sector = Clothes or Retail trade
then
        Classification = Regular  0
        Classification = Substandard  1

RULE #17
```

```
            Classification = Regular  0
            Classification = Substandard  1

      RULE #10
      if
            Debt to equity = [0.15,0.25)
            Month = April, August, December, February, January, March,
      November or September
      then
            Classification = Regular  1
            Classification = Substandard  0

      RULE #11
      if
            Debt to equity = [0.15,0.25)
            Month = June
      then
            Classification = Regular  0
            Classification = Substandard  1

      RULE #12
      if
            Debt to equity = [0.25,0.5)
            A. Current Ratio = [0.818090961032,1.11255748266)
      then
            Classification = Regular  0
            Classification = Substandard  1

      RULE #13
      if
            Debt to equity = [0.25,0.5)
            A. Current Ratio = [1.11255748266,1.47739837398)
      then
            Classification = Regular  0.666666666667
            Classification = Substandard  0.333333333333

      RULE #14
      if
            Debt to equity = [0.25,0.5)
            A. Current Ratio = [1.6118,8.25645958203]
      then
            Classification = Regular  1
            Classification = Substandard  0

      RULE #15
      if
            Debt to equity = [0.25,0.5)
            A. Current Ratio = [1.11255748266,1.47739837398)
            trade sector = Advertising and promotion service, Building and
      Construction, Import or Transport
      then
            Classification = Regular  1
            Classification = Substandard  0

      RULE #16
      if
            Debt to equity = [0.25,0.5)
            A. Current Ratio = [1.11255748266,1.47739837398)
            trade sector = Clothes or Retail trade
      then
            Classification = Regular  0
            Classification = Substandard  1

      RULE #17
```

```
if
        Debt to equity = [0.705646309441,1.894128)
        amount granted = 20000
then
        Classification = Regular  1
        Classification = Substandard  0

RULE #18
if
        Debt to equity = [0.705646309441,1.894128)
        amount granted = 100000
then
        Classification = Regular  0.333333333333
        Classification = Substandard  0.666666666667

RULE #19
if
        Debt to equity = [0.705646309441,1.894128)
        amount granted = [200000,300000]
then
        Classification = Regular  0.323529411765
        Classification = Substandard  0.676470588235

RULE #20
if
        Debt to equity = [0.705646309441,1.894128)
        amount granted = [600000,720000]
then
        Classification = Regular  0
        Classification = Substandard  1

RULE #21
if
        Debt to equity = [0.705646309441,1.894128)
        amount granted = [200000,300000]
        A.Debt/Asset Ratio = [1.53015480785,1.67282387691)
then
        Classification = Regular  0
        Classification = Substandard  1

RULE #22
if
        Debt to equity = [0.705646309441,1.894128)
        amount granted = [200000,300000]
        A.Debt/Asset Ratio = [1.31578947368,1.33738172617) or
[1.86169852764,2.07006271281)
then
        Classification = Regular  0.363636363636
        Classification = Substandard  0.636363636364

RULE #23
if
        Debt to equity = [0.705646309441,1.894128)
        amount granted = [200000,300000]
        A.Debt/Asset Ratio = [1.31578947368,1.33738172617) or
[1.86169852764,2.07006271281)
        perf. Of prior loans = Irregular
then
        Classification = Regular  0
        Classification = Substandard  1

RULE #24
if
        Debt to equity = [0.705646309441,1.894128)
```

```
        amount granted = [200000,300000]
        A.Debt/Asset Ratio = [1.31578947368,1.33738172617) or
[1.86169852764,2.07006271281)
        perf. Of prior loans = Regular
then
        Classification = Regular  0.7
        Classification = Substandard  0.3

RULE #25
if
        Debt to equity = [0.705646309441,1.894128)
        amount granted = [200000,300000]
        A.Debt/Asset Ratio = [1.31578947368,1.33738172617) or
[1.86169852764,2.07006271281)
        perf. Of prior loans = none
then
        Classification = Regular  1
        Classification = Substandard  0

RULE #26
if
        Debt to equity = [1.894128,2.936075)
        A.Debt/Asset Ratio = [1.33738172617,1.4)
then
        Classification = Regular  0.333333333333
        Classification = Substandard  0.666666666667

RULE #27
if
        Debt to equity = [1.894128,2.936075)
        A.Debt/Asset Ratio = [1.22449600453,1.27777691358)
then
        Classification = Regular  0
        Classification = Substandard  1

RULE #28
if
        Debt to equity = [1.894128,2.936075)
        A.Debt/Asset Ratio = [1.33738172617,1.4)
        credit r/s with other banks = Regular
then
        Classification = Regular  0.6
        Classification = Substandard  0.4

RULE #29
if
        Debt to equity = [1.894128,2.936075)
        A.Debt/Asset Ratio = [1.33738172617,1.4)
        credit r/s with other banks = none
then
        Classification = Regular  0
        Classification = Substandard  1

RULE #30
if
        Debt to equity = [2.936075,7.97725844286]
        A. Current Ratio = [1.25449065768,1.47739837398)
then
        Classification = Regular  0.529411764706
        Classification = Substandard  0.470588235294

RULE #31
if
        Debt to equity = [2.936075,7.97725844286]
```

```
       A. Current Ratio = [1.25449065768,1.47739837398)
       nb. Of prior loans settled = 1 or 2
then
       Classification = Regular  0.272727272727
       Classification = Substandard  0.727272727273

RULE #32
if
       Debt to equity = [2.936075,7.97725844286]
       A. Current Ratio = [1.25449065768,1.47739837398)
       nb. Of prior loans settled = 3
then
       Classification = Regular  1
       Classification = Substandard  0
```

DECLARATION

eclare that this thesis is my original work and has not been presented
other University, and that all sources of material used for the thesis
wledged.

Meretework Shawul

July 2004

submitted for examination with our approval as University Advisors.

Prof. Yohannes Abate

Nigussie Tadesse

July 2004