ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION STUDIES FOR AFRICA

POSSIBLE APPLICATION OF DATA MINNING TECHNOLOGY IN SUPPORTING LOAN DISBUSREMENT ACTIVITY AT DASHEN BANK S.C.

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

By

Askale Worku

July 2001

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION STUDIES FOR AFRICA

POSSIBLE APPLICATION OF DATA MINING TECNOLOGY IN SUPPORTING LOAN DISBURSEMENT ACTIVITY AT DASHEN BANK S.C.

BY

ASKALE WORKU

Name and Signature of Members of the Examining Board

Ato Getachew Jemenah, Chairman, Examining Board

Ato Tesfaye Biru, Advisor

Ato Nigussie Tadesse, Advisor

Dr. Kamal Bechkoum, External Examiner

DEDICATION

In memory of my father, Worku Wubneh

ACKNOWLEDGMENT

There are many people who supported me during the course of this research work. First and for most I want to thank my mother, Yekaba, for her full support, understanding and help. And my sister and brother, Mahtsente and Germachew, for their quick replies and assistance in obtaining necessary software.

This research work has been refined through the support of my advisors Ato Negussie Tadesse and Ato Tesfaye Biru to whom I am very grateful.

Dr. Nega Alemayehu gave me very quick and valuable replies to all the frequent communications we made through the Internet while living in the hectic life of the West we so often hear about. Thank you.

This work would have been all but impossible without the support I got from Dashen Bank management and Credit Department staff.

I am also grateful to all members of the School of Information Studies for Africa (SISA) staff, all my classmates, relatives and friends. Particularly, I want very much like to thank my friend Daniel for his encouragement and support.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

The commercial banking sector plays vital role in the economic development efforts of a country. And the viability of the sector relies on the ability of the institutions to maintain a positive inflow of resources. And one of the factors that affect the ability of the commercial banking sectors to maintain positive flow of resources is the problem of default rates. Among the Ethiopian banking sector there is a general problem of high default rate and the commercial banks have tried to tackle this problem through different ways.

One technique that has become popular in addressing problem of credit risk in other countries is data mining. Data mining technology has enabled banks in other countries to make good prediction on the probability that a certain borrower would default or not. But, so far, no commercial bank in Ethiopia has used data mining technology for such purposes i.e. assessment of credit risk. Thus, the objective of this research work was to see if application of data mining could also be beneficial in the Ethiopian banking context. For reasons of familiarity Dashen Bank S.C. was selected as a case study.

The methodology employed for the research had basically three stages. These were collection of data, preparation of data and model building and testing. Data was collected from two kinds of documents that were available at the head office of Dashen Bank. Then the data was prepared which included summarization, deriving of new fields and handling of missing data. The data mining technique employed for the model building and testing was neural network. Neural network software was thus used in building and testing a number of models.

From the numerous trials, many models with encouraging results were obtained. These models indicated that data mining application for credit decision-making is feasible at Dashen Bank. But one major limitation was unavailability of data in an electronic form. However, a survey in the IT department of Dashen Bank suggested that this problem is being duly addressed.

Chapter 1

Introduction

1.1 Background

The commercial banking sector is one of the most important financial institutions in mobilizing financial resources for an economic development of a country. An efficient contribution of the banking sector for the economic development of a country heavily relies on the capacity of the sector to maintain a positive inflow of resources. One of the factors that affect the capacity of banks to maintain a positive inflow of resources is a problem of high default rate in loan collection.

In the absence of an efficient loan management scheme, commercial banks will face a serious problem of collecting disbursed loans. A declining rate of loan collection is a threat not only to the individual banking institutions but also to the overall macro economic development of a country. A macro economic development of a country demands an efficient and stable mobilization of financial resources in the banking system of a country.

Thus, one of the fundamental challenges the commercial banking sector deals with is the capacity to design and implement an effective loan disbursement mechanism that ensures the highest possible rate of loan repayment and the minimum level of default. The challenge of developing and implementing such a system depended on the experience based knowledge of experts in the financial sector and the banking industry, in particular.

The practical challenge for the credit experts in a commercial bank is to be able to distinguish a high-risk borrower from a credit worthy borrower who meets the credit obligation properly and timely. A profitable operation of a banking institute and its capacity to maintain a positive inflow of resources in the banking system of the national economy is significantly affected by the task of properly differentiating a credit worthy borrower from a delinquent one.

Credit experts and other professionals in commercial banks make a decision on the basis of knowledge extracted from past experiences and certain rules by which the acceptability of a loan request is evaluated. The loan evaluation process of banks takes into account a set of factors that are believed to affect the capacity of a borrower to repay the loan amount in accordance with the loan agreement. On the basis of such an evaluation by credit experts, banks make a decision whether or not to entertain a given loan request.

The essence of credit evaluation process is a prediction of credit experts whether or not a given loan applicant will meet the obligation to repay the credit. And such a prediction is made by following certain guidelines and more importantly on the basis of an experience-based knowledge of experts in the field.

Since recently, advances in science and technology has responded in a creative technology that can capture the learning's of experts to develop a model and make a prediction as close to the truth as possible on the basis of past data on the subject. Such is the application of data mining technology in the banking industry to predict whether or not a particular loan applicant is credit worthy or not.

Data mining is defined as 'the process of extracting valid, previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions' (Connolly et. al., 1999). According to the writings of Sun Microsystems Computer Company (1997), '...data mining is a process of extracting meaningful and unobvious information from databases that enables companies to efficiently solve business problems and give them a competitive edge.'

What is central to the idea of data mining was the desire to extract a hidden knowledge from a huge amount of data and make use of the discovered knowledge for future decisions. For Benning et. al. (1999) what led to the development of the data mining technology in the 1990s is the need for companies to extract information from huge databases.

The application of data mining technology has increasingly become very popular and proved to be relevant for many sectors such as retail trade, health care, telecommunications, and banking (Ballenger et al., 1999).

In the banking sector data mining has been applied for a variety of purposes. One of these applications is in assessing credit worthiness of borrowers. 'Perhaps the most common application of data mining - and one of the ones that has been around longest (since the 1950s) - is credit scoring, a statistical method used to predict the probability that a loan applicant or existing borrower will default or become delinquent' (Wasserman, 2000). Literature on data mining applications assert that banks can use data mining techniques to assess which customers present the highest risk for defaulting on a loan. (Oracle Corporation, 1999; Benning et.al., 1999.; Ballenger et. al., 1999 and Sun Microsystems Computer Company, 1997)

Data mining extracts the useful information or pattern by building models. For instance, a predictive model can be developed by using the payment history of loan recipients and then the model would be used to identify people who are likely to default on loans.

While the technology of data mining is steadily growing fast in the developed parts of the world, it remains to be unknown in the Ethiopian banking industry. Given experiences elsewhere in terms of the benefits acquired in applying data mining technology in the banking sector, it is only proper to explore the relevance and significance of such a state-of-the-art technology in the Ethiopian banking context. For reasons of familiarity the researcher chose Dashen Bank S.C. to conduct the study.

1.2 Further Background to the Study Area

1.2.1 Banking History in Ethiopia

Shiferaw Bekele, (2001) in his article entitled 'The Evolution of Banking in Ethiopia' provides the following brief account on history of banking in the country. The article presents a concise summary of the origin and development of the banking sector in Ethiopia, from the earliest first bank of the country in the 19th Century to the banks of the modern time Ethiopia in the 20th Century.

Emperor Menelik II was the first to recognize the need and importance of a banking service in the late 19th Century. But lack of trained human resource in the sector, the relative poor economy of the country at the time and absence of the necessary administrative framework did not allow the

formation and operation of a local bank. Therefore, foreign banks had to be invited to begin addressing the emerging need of a banking service.

In 1905 the first bank of the land known as 'Bank of *Abyssinia*' was set up and it was owned by the National Bank of Egypt, which was an affiliate of the Bank of England. In spite of the establishment of the first bank in 1905, a considerable part of the country did not yet came out of the traditional bartering system which lasted as late as the beginning of the 20th century as the main form of economic transaction. In time, Bank of *Abyssinia* managed to grow slowly and branches started to be opened in the process of expanding services to the regional capitals. The *Harar* branch was established in 1906, the *Dire Dawa* branch in 1908 and the *Gore* branch in 1912.

Later, the Ethiopian government acquired 60% of the share of Bank of Abyssinia on purchase from the National Bank of Egypt while the remaining 40% continued to be held by foreigners. In August 28, 1931 Bank of Abyssinia was liquidated and a new central bank of the country was established under the name of Bank of Ethiopia. The Bank of Ethiopia can fairly be considered as a truly national bank of the country as it was fully owned by the national government. The Bank of Ethiopia was given all the tasks and responsibilities of a central bank and the first paper notes were issued in 1932. The progress of the bank was remarkable until its stride was cut short by the Italian invasion in 1936. The Italian period saw the introduction of Italian banks and insurance companies. There was a rapid expansion of the money market during the five years of Italian occupation owing to the large investment the Fascist government made in the country. However, upon the liberation of the country in 1941, the government had to start from scratch since the Fascists looted everything.

In 1942, State Bank of Ethiopia was set up and a new currency was designed. The steady improvement of the economy during the 1950's led to the expansion of the money market and opening of new banks as well as branches of foreign banks. Again in 1945 a new currency was issued. And in the same year the Agricultural Bank was established and later (in 1951) was converted to Development Bank of Ethiopia.

In 1963, the State Bank of Ethiopia was split into the National Bank of Ethiopia and Commercial Bank of Ethiopia. By the time the revolution broke out (in 1975) there were four commercial banks operating in Ethiopia, namely Commercial Bank of Ethiopia, The Addis Ababa Bank (private bank), *Banco di* Roma and *Banco di Napoli*. On January 1st 1975 the Provisional Military Administrative Council (Derg) nationalized all banks. The National Bank of Ethiopia was retained as a separate institution while the three commercial banks were merged with the Commercial Bank of Ethiopia.

Until 1994 government owned banks monopolized the banking sector. In 1994 the establishment of private banks became possible under the promulgation of Proclamation No. 84 of 1994 for Licensing and Supervision of Banking Business in Ethiopia. Today, there are eight commercial banks operating in the country namely Awash International Bank S.C., Bank of Abyssinia S.C., Commercial Bank of Ethiopia, Construction and Business Bank, Dashen Bank S.C., Development Bank of Ethiopia, Nib International Bank and Wegagen Bank. The services provided by these commercial banks include credit, mobilization of deposits, money transfer and international banking.

As at March 31, 2000 the total deposits of these banks amounted to Br. 19.2 billion and total loans advanced by the banks amounted to Br. 2.5 billion (NBE Quarterly Bulletin, 1999 Volume 15, No. 3).

1.2.2 Dashen Bank S.C.

Dashen Bank is one of the private banks established in Ethiopia since the promulgation of Proclamation No. 84 of 1994 that allowed the formation and operation of private banking business in Ethiopia. Dashen Bank was established on September 20, 1995 as a share company with an authorized and subscribed capital of Eth. Birr 50 million. The Bank provides both domestic and international banking services through a total of 22 Area Bank networks in the city of Addis Ababa and the regions. (In Dashen Bank, the outlets are named as "Area Banks" as opposed to the usual terminology of a "Branch". According to the Bank's Officials the change in the nomenclature is a unique feature of the bank, which signifies the qualitative delegation of responsibility and authority to each of the area banks.)

In the Bank's service of extending credit facilities, emphasis is given to businesses engaged in import, export, manufacturing, domestic trade and services, building and construction and agro industries. In money deposit function, the bank mobilizes various types of deposits including demand deposit, saving deposit, and time/fixed deposit. In addition to these two main functions, the bank is also engaged in international banking services and money transfer both locally and internationally.

Dashen Bank has reported having achieved the following results only in the first six months of its operation (Business Development Department of Dashen Bank, 2001)-

- It attracted 7237 depositors with a total deposit of Eth. Birr 152 million
- It had 422 borrowers and outstanding loan balance of Eth. Birr 106 million
- It created employment opportunity for more than 200 employees and
- The total assets of the company reached Eth. Birr 270 million

Five years after its establishment, the bank reported having achieved the following remarkable results (Business Development Department of Dashen Bank, 2001)-

- The deposit level grew to Eth. Br. 760 million with a total of 85,000 depositors
- It had 2100 borrowers and outstanding loan balance of Eth. Br. 575 million
- The staff size had grown over 600 employees and
- The total assets of the company reached Eth. Br. 1.2 billion

According to the Bank officials, computerization is another defining feature of Dashen Bank which was given due emphasis right from the beginning. More discussion on the automation effort is given under section 3.6 of Chapter 3.

1.3 Statement of the Problem

The underlying research problem that necessitated this research is the problem of high default rates in loan management of commercial banks. The commercials banks in Ethiopia are facing grave problems regarding the high default rates that are currently prevailing. An interview with one official of the National Bank of Ethiopia (NBE) revealed that the percentage of non-performing (irregular loans) to total outstanding loan balance of commercial banks to be 24.55% and 27.37% as at March 31, 2000 and September 30, 2000 respectively.

In the fourth quarter report for the fiscal year 1998/99, NBE has reported that the collection of loan by the banking system showed a decline of Eth. Birr 66.4 million which was a 6.4% decline compared to the third quarter of the same fiscal year. And compared to the amount collected during the same quarter of the previous year, the collection has declined by Eth. Birr 77.4 million, which is 7.3% (NBE, Quarterly Bulletin:1998/99, Vol. 14 No. 4).

As discussed in the background section of this research paper, a declining rate of loan collection is a threat not only for a viable existence of the financial institutions but also for the macro economic development of a country. Economic development programs can only continue to be financed if banks can maintain a positive net inflow of resources. Where banks fail in loan collection and disbursement of new loans decreases, the economic activity of particularly the private sector is directly affected and results in an overall decline of the economy.

Commercial banks have sought different approaches to minimize the default rate of loans but the main challenge remains to be one of designing and implementing an efficient mechanism for credit risk assessment. According to interviews held with bank officials, the most widely used

strategy followed by banks is the adoption of a very conservative approach in the loan approval process. Banks have increasingly become anxious to minimize credit risk and have increasingly tightened up their lending criteria. Such measures include depriving the authority of branches and approving loan request at head office level, a very strict collateral requirement that can adequately secure the loan amount and a considerable reduction of amounts to be granted in loan.

The problem of high default rate is recognized as a serious threat to the national economy to such an extent that the Federal Government has promulgated a law of foreclosure whereby commercial banks are authorized to sell collaterals and collect the proceeds without resorting to court of law. The proclamation under the name 'Property Mortgaged or Pledged with Banks' was promulgated in February 19, 1998 under proclamation number 97/98. Commercial banks have pushed for the promulgation of the foreclosure law considering it as one of the alternative approaches to deal with the problem of high default rates. When the draft foreclosure law was first heard of, it was met by a strong resistance from the private sector as contravening the principle of due process of law. On the other hand commercial banks pushed for a speedy adoption of the law to deal with the problem of high default rate.

The foreclosure law finally having been promulgated, commercial banks are now authorized to sell collaterals and repay the loan from the proceeds with out recourse to court of law. And hence the loan management approach of banks tends to highly emphasize on the collateral requirement. As it has only been less than three years since the foreclosure law was promulgated, whether or not it has helped to achieve the intended objective and the extent at which it has contributed to the intended objective is yet to be seen on the basis of empirical studies in the future.

In spite of legislative measures and the adoption of conservative approaches in granting loans, credit risk management continues to be a challenge for commercial banks of the country. While one may identify many reasons to explain the poor performance of loan collection, it is rather difficult to conclusively attribute the problem to a certain set of factors since the parameters involved are rather too many.

It is in this context that this research has sought to assess and experiment the potential applicability of data mining technology in supporting loan disbursement activity of commercial banks in Ethiopia.

The research was undertaken in a form of an experimental case study in one of the commercial banks in Ethiopia. For reasons of familiarity, the researcher chose Dashen Bank S. C. for conducting the experimental research. Case data was gathered from Dashen Bank S.C. on the basis of which the researcher built models that tested if the application of data mining technology can help to make a prediction with an acceptable rate of accuracy about the credit worthiness of borrowers.

As discussed in the background and literature review sections of this research paper, data mining technology has increasingly become popular in such areas as assessment of credit risks. The data mining activity suggested in addressing this type of problem is what is known as classification. As Moxon (1996) has put it, credit risk assessments are well suited for classification activity.

One data mining technique that is used for the purpose of credit scoring (assessment of credit risks) is the neural network technique. According to the writings of CorMac Technologies Inc. a bank manager uses intuition in making decision on whether or not to grant credit to a prospective

borrower. Intuition helps a bank manager to recognize certain similarities and patterns that his/her brain has become attuned to. He/She may never have seen an exact pattern before, but his/her intuition can detect similarities as well as dealing with the non-linearities. And the manager is probably unable to explain how his/her intuition works. But if we had a large number of data on several loan decisions, a neural network can be trained on these patterns i.e. the inner workings of the neural network have enough mathematical sophistication to reasonably simulate the expert's intuition.

In the present experimental research undertaking, neural network 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.

1.4 Objectives of the Research Undertaking

1.4.1 General Objective

The general objective of the research work is to explore the potential applicability of data mining technology in developing a model that can support the loan decision-making process at Dashen Bank S.C.

1.4.2 Specific Objectives

In order to achieve the above stated general objective, the research work has undertaken the following specific objectives:-

- Review literature on data mining at large and the application of techniques of neural network in particular.
- Explore different data mining software that support neural network technique;
- Identify and collect required data from Dashen Bank;
- Prepare data for analysis which include summarization, accounting for missing values, deriving other fields from existing ones and adjusting inconsistent data encoding;
- Build and test models;
- Report the result and make recommendations.

1.5 Research Methodology

For the purpose of this research undertaking the researcher has opted to use the methodology suggested by Berry and Linoff (1997) as provided in their work 'Data Mining Techniques: For Marketing Sales and Customer Support.' This methodology assumes that the business problem has already been identified and hence directly proceeds to the different data mining steps that need to be carried out in order to develop a model for the data-mining project.

The different steps suggested for a data mining project and how they were applied for the current research project are provided below:

A. Identifying sources of preclassified data

Sources of data were identified in consultation with credit department staff of Dashen Bank. Two sources of data were identified that were available in the form of two documents i.e. Loan

Approval Form (LAF) and Monthly Credit Report Form (Annex 1 and 2). The originators of these documents are the different area banks of Dashen Bank and these documents are made available to the head office for two purposes. The loan approval form is an important document in the loan decision process while the monthly credit reports are used for follow-up purposes. Both documents were available only in a manual format.

Sample area banks were selected since the limited time available for the research work did not allow collection of entire record of the bank's borrowers. From the twenty-two area banks that Dashen Bank has, eight area banks were chosen, by considering the relative length of time the area bank had been in operation in, the spatial distribution of the area banks and volume of data.

Thus, data from the above two documents were collected for the eight area banks. Discussions with bank experts were conducted several times in identifying the important variables to collect from the documents.

B. Preparing Data for Analysis

Data collected had to be massaged into a form that can allow the data mining tools to be used to the best advantage. As such, the collected data were cleansed into a form that was suitable for the particular neural network software used.

The data mining technique considered for the research undertaking was neural network. Therefore, different neural network software were evaluated mainly by considering price and complexity of the software. The Internet was the source of information on the different software and after considering a number of software the researcher decided on BrainMaker Neural

Network Software. This software was advertised to be the best selling neural network software and is reported to have addressed the specific area the research undertaking was concerned in.

In preparing data the collected records were summarized, inconsistent data encoding and missing values were accounted for, and new fields were derived from the already existing ones.

C. Build and Train the Computer Model

The software employed calls for dividing of the data into training and testing sets for the purpose of building models. The software automatically sets aside 10% of the data for testing. Numerous models were developed for the different area banks independently and for the combined data of some of the area banks.

1.6 Scope and Limitation

The scope of this research is to appraise the potential applicability of data mining in assessing credit risk at Dashen Bank S.C. While findings of the research work can fairly be considered as relevant in appraising the potential applicability of data mining in Ethiopian commercial banks at large, the scope of the present experimental research undertaking is strictly limited to appraising the possible application of data mining technology at Dashen Bank S.C.

The time that was available for the research work set a constraint on the amount of data collected. In addition, there was difficulty in obtaining applicable software. The search for an applicable and affordable neural network software that can be used to build and test models so that to assess possible application of data mining technology at Dashen Bank was rather time consuming.

1.7 Organization of the Thesis

The thesis is divided into five chapters. The first chapter is an introduction part which first sets out the background for the research work and further details the problem addressed and then puts the objective of the research. The chapter also discusses the methodology adopted for the study.

The second chapter discusses data mining technology and how it is related to the problem area this research is addressing. The chapter also discusses the application of data mining in the banks of other countries. The data mining technique used for this research i.e. neural network is also introduced at some detail in the same chapter.

The third chapter is devoted to further understanding of the problem area by studying the current credit procedures at Dashen Bank and the existing automation efforts. A discussion of the detailed problems identified during the survey concludes this chapter.

In the fourth chapter discussions on how the different data mining steps were carried out for this research are explained. These include data collection, data preparation, model building and testing. This chapter ends with the summary of results obtained. Finally, one chapter is devoted for the final concluding remarks and recommendations forwarded on the basis of the research findings.

Chapter 2

Data Mining and Neural Networks

2.1 Data Mining

2.1.1 Introduction

The ever increasing wide spread of inexpensive computers has made it possible for organizations to collect and store large collection of data. Al-Attar (1999) discuss that most organizations can currently be labeled as 'data rich'. But the data are usually used only to provide endless facts and figures. Such facts and figures do not represent knowledge and if anything could lead to information overload. Al-Attar (1999) further discuss that it is patterns in the data that represent knowledge and unfortunately most organizations are still knowledge poor. To bridge this gap different techniques are being developed. 'Because unsifted data represents a huge untapped investment, vendors of systems and software are rapidly moving to provide the tools that enterprises need to turn their mountains of data into valuable information. One of those tools is data mining' (Sun Microsystems Computer Company, 1997).

Edelstein (1998) also address the above concept by remarking that most organizations have accumulated large data while what they need is information. He then asserts that 'the newest, hottest technology to address these concerns is data mining.'

Connolly et. al. (1999) define data mining to be 'the process of extracting valid, previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions.' The authors stress that the focus of data mining is in finding information that is hidden and unexpected. Bigus (1996) also writes that the operative word in

data mining definition is discovery, which signifies that data mining is concerned with automated detection of new facts and relationships in data i.e. data mining tells users what they didn't know.

Sun Microsystems Computer Company (1997) write 'data mining is a process of extracting meaningful and unobvious information from databases that enables companies to efficiently solve business problems and give them a competitive edge.' And this has proven to be true for many sectors in the economy as there are many industries that have already highly benefited from data mining capabilities. Such industries include banking and finance, retail, health care and telecommunication (Ballenger et. al., 1999).

In carrying out a data mining activity there are steps to be followed. The general processes involved are to define the problem, select the data, prepare the data, mine the data, deploy the model, and then take business action (Saarenvitra, 2001).

The first step in data mining activity i.e. formulating of the problem is very crucial. Kestelyn (1997) cites Evangelos Simoudis, IBM's vice president for decision support as saying 'you can't set a data mining tool loose against a terabyte of data and just expect it to find something. Kestelyn (1997) further puts 'in other words, you need to have clear and unambiguous knowledge about the basic problem.'

The next step involves selecting of data, which is the main ingredient in data mining activity since the success of the data mining activity highly depends on the data (Kestelyn, 1997). Preparing the data is the step where much time is devoted. Saarenvitra (2001) estimates that this step can take up to 80% of the total project effort.

There are various techniques that are used to mine the data and discover important relationships. Some of the techniques are neural networks, decision trees, genetic algorithms and memory based reasoning (Berry and Linoff, 1997). These techniques use statistical and machine learning methods to search databases for patterns that describe relationships in the data or predict future values or behavior (Edelstein, 1997).

The model that is developed using a data mining technique would then be deployed i.e. the mathematical models would be implemented into operational systems (Saarenvitra, 2001). The last step is then to use the deployed model to achieve improved results to the business problem identified at the beginning of the process.

However, it does not mean that the above steps are always carried out sequentially. Data mining is an iterative process: 'the basic step (extract the data from a database, prepare and otherwise clean it, run the knowledge discovery algorithms on it, analyze the results and write it are repeated again and again. And each step can conceivably return to any previous one' (Skalak, 2001).

The current experimental research work also followed accepted stages of a data mining process to develop and test a model on the basis of the case data gathered from Dashen Bank. This process of developing a model is meant to identify both potentials and limitations of introducing data mining technology in assisting the credit decision process at Dashen Bank S.C.

2.1.2 Data Mining and Other Statistical Tools

There have existed many analytical tools that support a verification-based approach (Moxon, 1996). In such tools, the user puts a hypothesis about specific data relationships and then the tools

would help in verifying or disproving the hypothesis. Query and reporting tools will interrogate the data and report on any pattern (query) requested by the user. This discovery is possible if the users prompt it i.e. unless the user suspects a pattern they will never find it. A marginally better situation is encountered with the OLAP (online analytical processing) tools, which can be termed visualization driven since they assist the users in the process of pattern discovery by displaying multi-dimensional data graphically.

But the traditional tools have limitations in that results depend on the ability of the analyst to pose appropriate questions (Moxon, 1996). In addition, when the numbers of variables being analyzed are many it would be difficult to identify the important variables (Two Crows Corporation, 1999). Data mining is different because 'rather than verify hypothetical patterns, it uses the data itself to uncover such patterns' (Two Crows Corporation, 1999).

Edelstein (1997) also explain the difference between OLAP and data mining with the following two statements. 'When users employ OLAP and other query tools to explore data, they guide the exploration. However, when users employ data mining tools to explore data, the tools perform the exploration.'

However, it does not mean that data mining has replaced other techniques such as OLAP, query reporting etc. Graettinger (1999) wrote that 'data mining does not replace but rather complements and interlocks with other decision support system capabilities such as query and reporting, on-line analytical processing (OLAP), data visualization, and traditional statistical analysis.' For instance Edelstein (1997) discuss that OLAP can assist the early stage of data mining process by helping focus attention on important variables, identifying exception or finding interactions. Edelstein

(1997) further put that this is important since the better one understands the data the better would be the knowledge discovery (data mining) process.

2.1.3 Data Mining Activities

Some of the major tasks (activities) that could be carried out using data mining are classification, estimation, time series, affinity grouping (association), clustering and sequence discovery (Berry and Linoff, 1997; Moxon, 1996 and Edelstein, 1997).

Moxon (1996) writes that classification is perhaps the most common data-mining task. The writer explains that in classification task, a set of pre-classified examples are used to develop a model that can classify new data into one of the predefined set of classes (Berry and Linoff 1997). For this research, classification activity is to be carried out since a model is to be built by using the preclassified data of past borrowers i.e. there is a need to build a model consisting of independent variables (E.g. income, marital status) that can be used to determine a dependent variable (E.g. Credit risk) (Small and Edelstein, 1997). The model then would be employed in determining whether a prospective borrower would default or not. Moxon (1996) put that credit-risk applications are particularly well suited for classification activity. The writer also discusses that the data mining techniques that are frequently employed for classification task are decision tree and neural networks.

The second data mining activity is estimation. Classification deals with discrete outcomes like 'yes' or 'no' while estimation deals with continuously valued outcomes (Berry and Linoff 1997). For instance, instead of using a binary classifier to determine whether a loan applicant is a good credit risk, this approach generates a credit worthiness score (Moxon, 1996).

'Time series forecasting uses series of existing values and their attributes to forecast future values.' What makes time series forecasting different from other tasks is that there is dependence on time (Edelstein, 1997).

The other common data-mining task is affinity grouping, which is used in determining which things go together (Berry and Linoff, 1997). Association approaches are common in market basket analysis (Moxon, 1996). It is used to determine what products go together at one time purchase. Moxon (1996) put that affinities (associations) 'are expressed in terms of confidence rules such as 80 percent of all transactions in which beer was purchased also included potato chips.'

Sequence based analysis is a variant of the association task and the only difference is that the related items are spread over time (Edelstein, 1997). Here, the order in which different items were purchased is studied. For instance, a time series study could be made to study the sequence of the transactions of having an account number and holding of a credit card (Moxon, 1996).

The other task identified for data mining is clustering. 'Clustering is the task of segmenting a heterogeneous population into a number of more homogeneous subgroups or clusters' (Berry and Linoff, 1997). The authors also put that that there are no predefined classes rather the records are grouped together on the basis of self-similarity. Moxon (1996) writes that clustering is usually the first step in data mining analysis. Clustering would first be used to identify related records then other data mining tasks could be performed to explore further relationships.

2.1.4 Applications of Data Mining in Banks

Ballenger et. al. (1999) write that data mining is a 'horizontal application meaning it is not specific to any industry.' The writers add that the ingredients that are important for data mining are availability of data and the willingness to explore the possibility of hidden knowledge that resides in the data.

However, there are industries that have already made significant progress in the application of data mining. One of these industries is the banking industry (Benning et. al., and Ballenger et. al., 1999).

Riggen and Budansky (1997) write that 'more and more examples of the application of data mining techniques in the banking industry are emerging in trade press articles.' The writers further discuss that banks are using data mining to identify their most profitable customers, understand why those customer groups are profitable, and to predict how customer profitability behavior will react to changes to products, services, or fees. Ballenger et. al. (1999) on their part put that the banking and finance industries are using data mining in identifying underserved populations, evaluating credit risk, analyzing profitability, mortgage customer behavior, direct marketing, detecting credit card fraud etc. (Ballenger et. al. 1999).

Especially the application of data mining for credit risk assessment in banks, which is the area addressed in this research, has been discussed by many (Bigus, 1996; Oracle Corporation,1999; Benning et.al., 1999; Ballenger et. al., 1999 and Sun Microsystems Computer Company, 1997). 'Perhaps the most common application of data mining – and one of the ones that has been around

longest – is credit scoring. Credit scoring is used to predict the probability that a loan applicant or existing borrower will default or become delinquent' (Wasserman, 2000). Credit scoring has benefited both banks and consumers by reducing the time needed to approve loans and the costs of evaluating them.

From the literature it is clear that many banks have already made use of data mining applications. However, it is difficult to obtain information on their findings. Piatetsky-Shapiro writes that corporations (and especially banks and insurance companies) are not likely to publish if they find something of competitive advantage through knowledge discovery. Nevertheless, there are few banks whose experiences have been discussed by writers.

For instance, Bank of Montreal is reported to analyze mortgage customers' transactions in checking, saving and other accounts for insight into who is at risk of defaulting. And the bank was surprised to find out that some customers who consistently made their mortgage payments late were not necessarily at a high risk of defaulting i.e. there were certain type of customers who were in the habit of paying bills late but eventually fulfilled their obligations (Fabris, 1998).

And some banks are reported to be currently testing data mining tools in managing credit portfolios more efficiently. Jianmin Liu, vice president and project manager in credit risk management for Bank of America's mortgage division, discusses that data mining holds great promise in assessing the risk of a bank's entire portfolio of loan (Fabris, 1998).

According to references made regarding application of data mining technology in banks, Bank of America has been able to form detailed demographic views of the banking habits and financial assets of select groups of its customers. In addition, it is further discussed that Canadian Imperial

Bank of Commerce (CIBC), based in Toronto, is using data mining to support decision making across the bank (http://www.rpi.edu/~arunmk/dm1.html).

Riggen and Budansky (1997) on their part discuss that one bank had used data mining techniques to segment its most profitable customers, and to understand why certain customer segments were profitable. Interestingly the bank discovered that it was losing a small amount of money on customers that had only one product with the bank. Another bank in America, Fleet Bank, is using data mining to more narrowly define customer segments and marketing campaigns (Riggen and Budansky,1997).

Ballenger et. al. (1999) on their part discuss that most banks do not admit to using data mining and have policies not to discuss it. But some who admit using data mining technology are Bank of America, First USA, FCC National Bank, Federal Home Loan Mortgage, Chevy Chase Bank, US Bankcorp, USAA Federal Savings Bank.

2.1.5 Neural Network as a Data Mining Technique

There are several techniques that can be used as a tool to carry out a data-mining activity (classification, estimation, clustering etc.). The common ones are artificial neural networks, decision trees, genetic algorithms and nearest neighbor method (Liao, 1999).

As already stated in the objectives, the technique employed in this research undertaking is the neural network technique. A good readable account on discussion for the other techniques is available in the works of Berry and Linoff (1997).

'Neural networks are probably the most common data mining technique, perhaps synonymous with data mining to some readers' (Berry and Linoff, 1997). Neural networks are well suited for data mining tasks due to their ability to model complex, multi dimensional data' (Z-Solutions, 1999). Bigus (1996) also discuss that neural networks are one of the key technologies used for data mining.

Neural techniques have been cited to be applicable for the specific problem that this research work is addressing (Bigus, 1996; Fraser, 2000 and Stergiou and Siganos). According to the writings of PMSI (2001), a neural network can be used to build a model by collecting data describing people who borrowed money (age, income, married or single, etc.) along with whether there were any payment problems. This model would contain the relationship (if any) between each of the independent variables (age, income, married or single, etc.) and the outcome of the loans. The model developed can then be used for prediction of new customer data. Neural networks are discussed in more detail in the next section.

2.2 Neural Networks

2.2.1 Introduction

Neural network is 'an artificial representation of the human brain that tries to simulate its learning process' (Frohlich, 1999). The appropriate terminology for neural networks is artificial neural networks. But it is customary to drop the artificial and write only neural network (Stergiou, 2001).

Neural networks developed due to difficulty in creating the basic intelligence in computers using the conventional algorithmic approach (Pudi, 2001). Conventional computers are good at

following explicit instructions over and over again (Berry and Linoff, 1997). Stergiou and Siganos also put that conventional computers take an algorithmic approach where the computer has to follow steps of instructions in solving a problem. Using algorithmic problem solving approach scientists have been able to create machines that can solve complicated logical and mathematical problems. However, with the algorithmic approach, it proved difficult to create a machine that had general human intelligence. By general intelligence we mean every day tasks such as recognizing a face, recognizing a speech, making a cup of coffee etc (Grove, 1996).

The challenge for scientists to create intelligent machines encouraged artificial intelligence workers to consider the structure of the brain. 'To understand human intelligence and make programs that perform in an intelligent way we must copy the structure of the brain. This is the basic idea behind neural networks' (Grove, 1996).

Neural network, just like the brain, 'is composed of large number of highly interconnected processing elements working in parallel to solve a specific problem' (Siganos, 1996). These networks have the capacity to learn, memorize and create relationships amongst data.

Neural networks, unlike the conventional algorithmic computers, cannot be programmed to perform specific task. Neural networks learn from examples, rather than being told rules or mathematical formulas (Lawrence, 1994). For instance, to generate a model that performs sales forecast neural network only needs to be fed raw data related to the problem. The raw data could include past sales, prices, competitor's prices, and other economic variables. The neural networks then learn from these facts and produce a model that can be used to provide prediction of future sales when provided with the independent variables (Z Solutions, 1999).

However, 'Neural networks and conventional algorithmic computers are not in competition but complement each other. There are tasks more suited to an algorithmic approach like arithmetic operations and tasks that are more suited to neural networks' (Stergiou and Siganos, 1997). The writers also add that there are tasks that are best handled by combining the two approaches.

2.2.2 Brief History of Neural Networks

The original work on neural networks started even before the emergence of digital computers i.e. in the 1940s (Berry and Linoff, 1997). In 1943 McCulloch and Pitts made the first model of the biological neuron. The model described neuron as 'linear threshold computing unit with multiple inputs and a single output of either 0, if the nerve cell remains inactive, or 1, if the cell fires' (Fraser, 2000). The author further discusses that the neuron fires if the sum of the inputs exceeds a specified threshold.

In the 1950s, when digital computers became available, scientists implemented models called perceptrons (Berry and Linoff, 1997). Perceptrons are the first artificial neural networks, which are based on a unit called the perceptron, which produces an output scaled as 1 or –1 depending upon the weighted, linear combination of inputs' (Fraser, 2000).

Later, in the 1960s two scientists demonstrated basic theoretical deficiencies of the above networks (Berry and Linoff 1997). The scientists (Minsky and Papert) demonstrated that the perceptron could not represent simple functions, which were linearly inseparable (Stergiou and Siganos, 1997).

The above deficiency led to the decline in the study of neural networks in the 1970s (Fraser, 2000). Then, in 1982, John Hopfield invented back propagation, a way of training neural networks that sidestepped the theoretical pitfalls of earlier approaches.

This resulted in the renaissance of neural network research and in the 1980s neural network researchers moved from the lab to the commercial world where they have since been applied in virtually every industry (Berry and Linoff 1997).

Several factors contributed for the popularity of neural networks in the 1980s. First, computing power was available in abundance. Second, analysts became more comfortable with neural networks realizing that they are closely tied with familiar statistical methods. Third, data was available easily because of automation of most operations (Berry and Linoff 1997).

2.2.3 Application of Neural Networks

Neural networks should be applied in situations where traditional techniques have failed to give satisfactory results (Z Solutions, 1999). Frohlich (1999), also states that 'neural nets are being constructed to solve problems that cannot be solved using conventional algorithms.' Siganos (1996-97) on his part discusses that neural networks can be 'used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.'

Up to date, neural networks have already been put to use to solve different problem domains. Some of these domains of problems as described by Frohlich (1999), Berry and Linoff (1997) are:

- Pattern association
- Pattern classification
- Regularity detection
- Prediction
- Clustering
- Optimization problems etc.

Some of the specific application areas where neural networks have been put to use include (Stergiou and Siganos 1997; Lawrence, 1994, Fraser, 2000 and Smith, 1996):

- Credit risk assessment.
- Neural networks for handwritten character, speech, fingerprint and electrical signal recognition.
- Neural networks that detect hypertension and heart abnormalities.
- Neural network for automatic vehicle control
- Neural networks that are used in improving marketing mail shots etc.

2.2.4 Basic Structure of Neural Network

The structure of neural network is very similar to the structure of the neurons in our brain. Therefore, it is perhaps important first to see the structure of the brain.

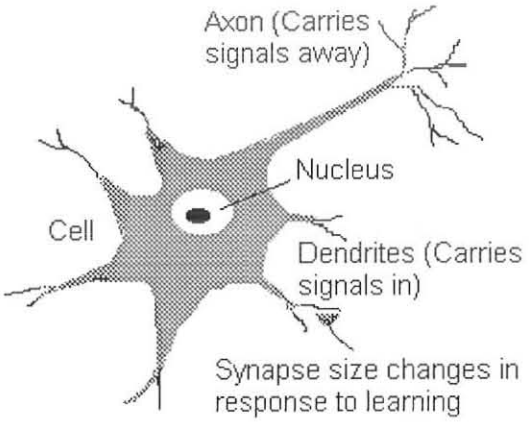


Figure 1: A simple neuron cell

The above figure shows the structure of a single neuron in the human brain. The neuron consists of 'dendrites for incoming information and axon with dendrites for outgoing information that is passed to connected neurons' (Frochlin, 1999). The writer discusses that information is transported between neurons in the form of electrical stimulation along the dendrites. The incoming information that reaches the dendrites are added up and passed along the axon to the dendrites at the other end. At the synapse transmission of signals from one neuron to the other takes place. 'The transmission of signals from one neuron to another at a synapse is a complex chemical process in which specific transmitter substances are released from the sending end of the junction.' If the received signal exceeds certain threshold the information would be passed to other neurons and the neuron is said to be activated. Otherwise if the stimulation is too low the information would not be passed any further and the neuron is said to be inhibited.

Grove (1996) discusses that the brain is composed of around 10^{11} neurons. And the writer further writes that these neurons are arranged in a rough layer like structure. Each neuron works as a processor and massive interaction between all cells and their parallel processing makes the brain's abilities possible (Frochlin, 1999). The early layers receive information from the sense organs

(eyes, ears). The final layers produce motor output (E.g. Moving arms and legs). The middle layer forms the associative layer. This layer is the least understood but is considered to be the most important part of the brain in humans.

Frohlich (1999) writes that neurons are adoptive i.e. the connection structure is changing all the time and learning takes place through these adoptions. Grove (1996) also put that the brain seems to learn in three ways by growing new axons, by removing axons and by changing the strengths of existing axons.

The simulation of the human brain is what we call artificial neural network or simply neural network (Smith, 1998). 'However, because our knowledge of neurons is incomplete and our computing power is limited, our models are necessarily gross idealizations of real networks of neurons' (Stergiou and Siganos, 1997).

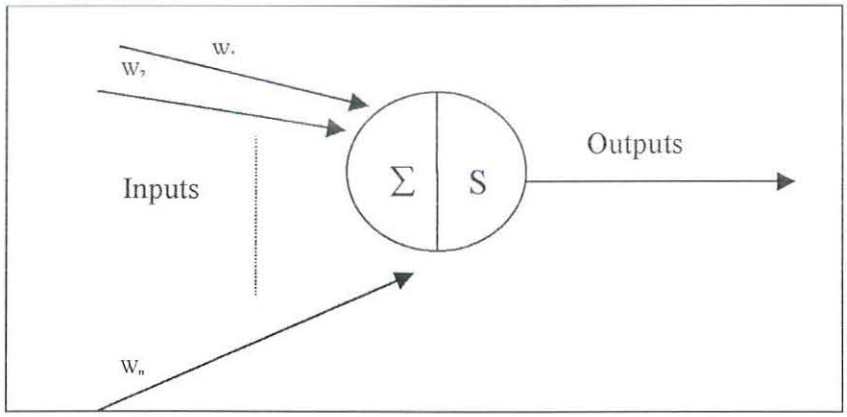


Figure 2: Artificial Neuron

Information is received from different inputs and are combined using a combination function, which is usually a summation of the different weights. Then a transfer function is used to calculate a single output between 0 and 1. The transfer function represents the non-linear characteristics exhibited in biological neurons.

Typical functions include threshold function, step transfer function, sigmoid function and Gaussian functions. (Lawrence, 1994) Among the different functions the common one is the sigmoid transfer function. (Fraser, 2000; Frohlich, 1999 New Wave Intelligent Business Systems Inc.) This is also the function used for this research work since the provider's of the software employed, put that they had not seen any problem that train fundamentally better with anything other than the standard sigmoid transfer function (California Scientific Software, 1998). A mathematical description on the different functions is available in the work of Lawrence (1994). The formula for sigmoid function is given by:

$$v = (1 + e^{-s})^{-1}$$

where s = sum of the inputs to the neuron

v = value of the neuron

The combination and transfer function together make up the activation function (Fraser, 2000). The resulting value from the activation function will be compared with certain threshold value. If the input exceeds the threshold value, the neuron will be activated, otherwise it will be inhibited. If activated, the neuron sends an output on its outgoing weights to all connected neurons and so on (Frohlich, 1999).

In artificial neural networks neurons are grouped in layers (Frohlich, 1999). There are basically three types of layers. These are the input, hidden and output layers which constitute of the input, hidden and output neurons respectively (Frohlich, 1999). Knowledge Technology Inc. defines the three terms as follows:

Input Layer: A layer of processing elements that receives the input to the neural net.

Hidden Layer: A layer of processing elements between a neural network's input layer and its output layer.

Output Layer: The layer of processing elements, which produce a neural net's output.

There could be a number of input, hidden and output neurons in their respective layers. For instance in the following neural network there are four input, two hidden and one output neurons. And the network has one input, one hidden and one output layer. In the discussion in Chapter Five the concept of networks size is often raised. And it is to mean the number of input, hidden and output neurons the network has and their interconnections. And in addition network size also indicates the number of hidden layers the network has.

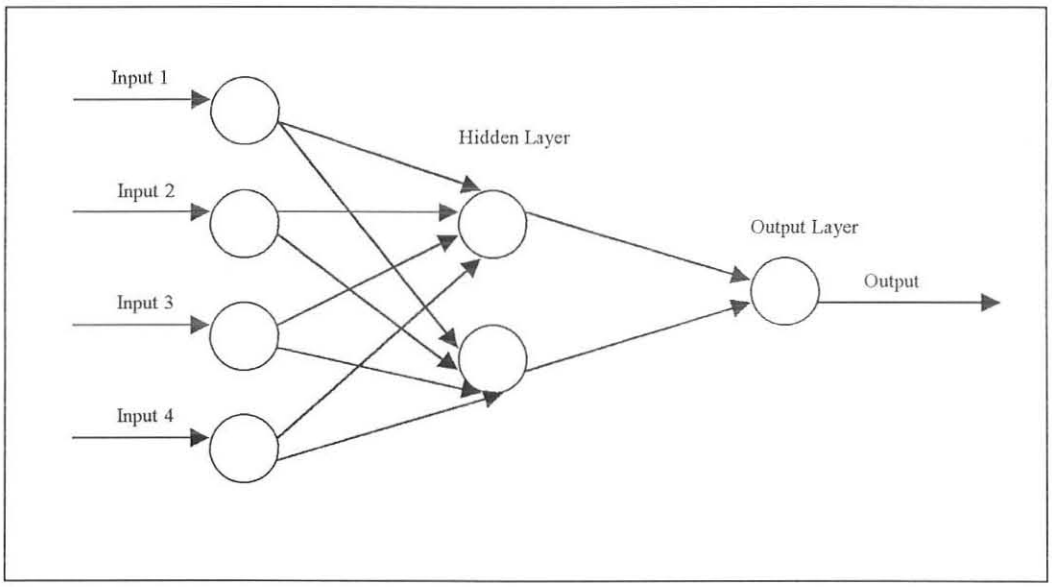


Figure 3: A simple feedforward neural network with one input layer, one hidden layer and one output layer

The way neurons are connected and the learning rule adopted determines the learning process in a neural network. These concepts are discussed in the next section.

2.2.5 Classification of Neural Networks

Lawrence (1994) describes that neural networks can be described 'in terms of the connections between them (topology) and its learning rule.'

2.2.5.1 Connections

Connection topologies define how data flows between the input, hidden and output layers (Bigus 1996). Lawrence (1994) describes connection topologies have an enormous effect on the operation of a network.

According to Bigus (1996) there are two major categories of connection topologies. These are feedforward and recurrent networks also known as feedback networks. The writer discusses that in feed forward networks information flows from input layers to zero, one or more succeeding hidden layers and then to the output layer i.e. data travel only in one way or there are no feedback or loops. First, data enters the neural network through the input neurons. The total input signal is then passed through an activation function in order to determine the output. The most widely used activation function, as discussed above, is the sigmoid function. This function converts an input value to an output value ranging from 0 to 1 (See Bigus, 1996).

Feedforward networks are used in situations where all the information to bear on a problem can be presented at once. For instance, for the research being undertaken the feedforward networks would be employed since a network is to be trained by providing information related to a borrower together with the outcome i.e. whether that borrower with the stated information defaulted on his/her loan.

The other types of networks are the recurrent networks (feedback networks). In feed back networks signals can travel in both directions by introducing loops in the network (Stergiou and Siganos,1997). Bigus (1996) also discuss that 'in recurrent networks, information about past inputs is fed back into and mixed with the inputs through recurrent or feedback connections for hidden or output units.' This makes the neural network to contain memory of past inputs. Such networks are used in situations where we need the neural network to somehow store a record of prior inputs and factor them in with the current data to produce an output.

2.2.5.2 Learning Methods

'The learning rule is the very heart of a neural network; it determines how the weights are adjusted as the neural network gains experience' (Lawrence, 1994). The writer further discusses that there are many different learning tools and mentions the Hebb's Rule and Delta rule to be among the well known ones.

Further, the writer discusses that many networks use some variation of the Delta rule for training. One of the variations of the Delta rule is the most widely used learning method i.e. back propagation. Back propagation is cited by many writers to be the most widely used training algorithm (Frohlich, 1999; Fraser, 2000 and Knowledge Technology, Inc., 2000).

The back propagation algorithm is also the algorithm that is best suited for the research project since it is stated to be the best algorithm for a classification activity (Bigus, 1996).

Back propagation algorithm is responsible in large part for the reemergence of neural networks in the mid 1980s i.e. after ten years of decline during the 1970s (Bigus, 1996).The back propagation algorithm uses the supervised learning approach during training. Supervised learning is one of the learning paradigms used while learning. The other learning paradigm is unsupervised learning.

In supervised learning the network learns through examples. The neural network is given a problem and it makes a classification or prediction. Then at that point the network would be given the correct answer. The learning algorithm takes the difference between the correct output and the prediction the neural network made and then uses the information to adjust the weights of the neural network so that the prediction next time would be closer to the correct answer. The neural network has to be given examples many times in order to learn and make correct predictions (Bigus 1996). 'A neural net is said to learn supervised, if the desired output is already known' Frohlich (1999).

The other learning paradigm is the unsupervised learning where a teacher is not required for training (Grove, 1996). In this paradigm there are no target outputs and it is impossible to determine what the result of the learning process will look like (Frohlich, 1999). The network is simply exposed to number of inputs and the network organizes itself in such a way as to come up with its own classification for inputs (Lawrence, 1994).

As mentioned, the back propagation uses the supervised training paradigm. (Knowledge Technology, Inc.) A basic back-propagation algorithm consists of three steps. First, the input pattern is presented to the network whereby the input pattern is propagated through the network until they reach the output units. This forward pass would produce the actual or predicted output. Then the desired output would be given as part of the training vector, so that the actual output can

be subtracted from the desired output in order to give the error signal. In the third step the errors are passed back through the neural network by computing the contribution of each hidden processing unit and deriving the corresponding adjustment needed to produce the correct output. The connection weights are then adjusted and the neural network is said to have learned from an experience (Bigus, 1996). These three steps are repeatedly carried out for every example in the data until the weights are no more adjusted. A good mathematical description of back-propagation algorithm is available in the work of Tveter (2000).

Chapter 3

The Existing Credit Approval Procedure at Dashen Bank

The purpose of this research undertaking is to experiment the potential applicability of data mining technology in supporting credit disbursement activities at Dashen Bank. To this end, it is appropriate to begin by properly understanding the current credit approval procedure of the bank and the major limitations thereon.

Therefore, this chapter will be devoted for discussing and understanding the existing credit approval procedure of Dashen Bank. The discussion will begin by introducing the different types of loan available at the bank to be followed by classification of loans and requirements for loan application. The chapter will then give a reasonably detailed account of the different stages of the loan approval procedure to be followed by a brief discussion on the credit follow-up mechanism.

The chapter further discusses the existing level of automation at Dashen Bank and on the basis of the over all survey of the existing procedure, the chapter concludes by discussing the problems identified during the survey.

3.1 Types of Loans

Dashen Bank extends four types of loans for the business community. Each of these loans is briefly described herein below.

Term Loan

A form of loan in which the amount borrowed will be paid on an installment basis over a fixed period of time. The loan period may be either short or medium term. Short-term loans are usually meant to meet the needs of working capital for a business whereas medium-term loans are made available for investment purposes.

Overdraft Facility

A form of credit facility by which the borrower is allowed to draw money in excess of his/her deposit, but only to the extent allowed by the bank under the overdraft facility agreement. The usual duration for an overdraft facility is six months.

Merchandise Loan:

A merchandise loan is granted against the security of merchandise in stock. This loan is usually extended for a short period of time (90 days) for customers that have a good track record of repayment and only to help them meet temporary shortage of working capital.

Letter of Credit Facility

A form of credit facility made available for importers. The facility is made available to importers in order to cover some percentage of the value of an opened letter of credit so that their working capital would not be tied before arrival of goods.

3.2 Classification of Loans

Loans extended by the bank are classified into different categories. The basis of the classification is the nature of the economic sector the business loan is intended to serve. Accordingly loans at Dashen Bank are classified into the following categories:-

- Agriculture
- Building and Construction
- Domestic Trade and Services
- Export
- Import
- Manufacturing

3.3 Requirements for Loan Application

According to interviews held with officials of the bank and on the basis of official document of the bank, the following are identified as the requirement for loan application.

- The application letter should indicate: -
    - Particulars about the loan applicant (name, address, work experience, etc.);
    - Address of the business;
    - The loan amount requested and the business purpose for which the loan is intended;
    - Proposed period and schedule of repayment;
    - The nature of the security to be provided and proof of ownership.

- Renewed trade license for the concerned trade sector and an investment certificate where applicable;

- Balance sheet, income statement and cash flow projection where available (Audited financial statements are encouraged). For businesses that do not keep books of accounts, they may use a form known as Financial Credit Report (FCR) prepared by the bank to gather the required financial information about new loan applicants. (The format for FCR is attached as Annex 3) Where a loan request is supported by a project study, the project document (or the feasibility study) should provide a projected financial statements as per the study;

- The Memorandum and Articles of Association for businesses established in the form of a company or other forms of business organizations;

- A feasibility study for new businesses, where available;

- Supporting documents for items listed in the balance sheets;

- Evidences of securities such as land holding certificates for buildings and ownership booklet for motor vehicles;

3.4 The Loan Approval Procedure

A loan request passes through different stages of an evaluative process before it finally gets approved and disbursement of the loan fund is authorized. In the interest of a better clarity of the activities involved in the various stages of the process, the researcher has divided the loan approval procedure into eight stages.

Each of these stages and the salient features of the activities involved are described herein below. In the actual practice of the loan approval procedure at the bank, two or more of the following stages may usually be undertaken simultaneously in order to save the loan request processing time.

3.4.1 Initial Review of Documents Submitted by the Prospective Borrower

The first step is to review the application filed by the customer to ascertain whether or not all the required documentations have been filed to consider the request for the loan. As discussed above, any application for loan should be supported by all the necessary documentation including the trade license, company profile, financial statement, and project studies, if any. Thus, the process begins by such a preliminary review of documents to ensure that all the required documentations are properly filed in order to begin the credit evaluation process. Ensuring proper documentation of all the required information is quite vital because the subsequent evaluative process of the loan depends on the documents available in the file.

3.4.2 Customer Visit

Once it is ascertained that all the required documents are properly filed, the next step of the process would be to visit the business of the prospective borrower, where applicable. The purpose of this visit is to assess the overall situation of the business and get a sense of, among other things, the customer's income and the business worth.

3.4.3 Analyzing Financial Statements

Analyzing the financial statements is an instrument to assess the financial soundness of a business. Therefore, the third important step of the loan approval process will be devoted to analyzing the financial statements i.e. the balance sheet and the income statement. Where the borrower is a business that keeps proper books of accounts and has an audited financial statement, the audited financial reports are considered as sufficient financial documents to ascertain the financial soundness of the business. If the financial statements are not audited, they can still serve as provisional financial documents to make a preliminary assessment.

Where there are no books of accounts, the required financial statements shall be prepared on the basis of the available information and entered in a form prepared by the bank for this purpose called Financial Credit Report (FCR). The FCR is a format prepared by the bank to enter all the required information in a situation where there are no financial statements. The FCR has an application part, a financial statement part and a description part (Format is attached as Annex 3).

The application part provides basic information collected from the loan application documents where as the financial statement part gives information about assets, liabilities, capital, expense

and income. Such financial information will help to determine the profitability and viability of the business under consideration. The description part of the FCR provides description for items in a balance sheet.

In the absence of audited financial statements, the bank will make the necessary verification regarding the items reported in the provisional financial statements or the FCR.

3.4.4 Collecting Credit Information

At this stage of the process, the bank would be engaged in collecting all available credit information on the prospective borrower. Credit information is basically about past credit history and it also includes any other relevant information such as, about reputation, personal integrity, and character of the loan applicant. The information is supposed to be solicited from a wide range of possible sources of information including banks, other financial institutions and business community. However, in practice credit information is mainly solicited from other commercial banks and hence, the bank would distribute information request to other commercial banks to find out whether the prospective borrower has another credit relation with other banks and reports of the credit relation, if any. The credit information inquiry further requires a declaration of any liability a loan applicant owes to third parties.

3.4.5 Determining the Collateral

A business loan is usually secured by furnishing a sufficient guarantee for the repayment of the loan money. The types of collaterals normally accepted by the bank are fixed assets such as

buildings and vehicles. There are few exceptions where a loan may be granted with out a collateral or with the security of a personal guarantor. Thus, at this stage of the process, the bank shall assess and determine the acceptability of the security provided by the loan applicant to guarantee the repayment of the loan money.

3.4.6. Recommendation of Loan

Loans are normally approved at the head office. And the responsibility of the Area Bank is to make an informed recommendation based on collected data and prior experience.

According to the bank experts a loan approval process is expected to be evaluated by a set of criteria called the '5 Cs', which stands for character, capacity, condition, capital and collateral. Their meanings are provided herein below:-

Character: Trying to determine the applicant's willingness to pay the loan. This can be done by studying the past behavior of applicant, general background, risk management ability, public views etc.

Capacity: This involves identifying of the applicant's ability to pay the loan. This is important so as to avoid possible under financing or over financing. Financial position indicators such as net working capital, current ratio, income, etc. are important in this regard.

Condition: General economic, political and social conditions of the area the applicant's business is located in.

Capital: Refers to ownership equity contribution and is important in assessing ownership interest.

Collateral: Refers to assessment of acceptability of the collateral, the loan to collateral proportion, and the marketability of the collateral in case of default.

Generally the above five factors are considered when recommending an applicant for a loan. But since it is hard to put explicit rules for the combination of these factors the banker's experience becomes very important.

The Area Banks put their recommendation in a form known as Loan Approval Form and pass it to the head office for further evaluation and final approval (Format is attached as Annex 1). But there are few Area Banks who are given the discretion to approve loans up to a certain limit at the Area bank level.

The Loan Approval Form includes the financial statements and summarized information about borrower's business. Area Banks outside Addis Ababa fax the Loan Approval Form to the Head Office whereas the Area Banks within Addis Ababa send the whole document (file) to the Head Office for consideration by the Credit Department.

3.4.7 Loan Approval

Credit analysts of the Credit Department study documents submitted to the head office. The credit analysts check whether the Area Banks have properly followed the procedures and properly evaluated the loan request in accordance with the '5 Cs' requirements discussed above. The experience and judgment of the analysts is quite important in coming to the final recommendation

on whether to recommend a loan to a specific customer because, as already said above, there are very many factors that affect the loan decision process like financial position, past performance, condition, collateral etc. And the intuition on how a combination of these different factors should be judged comes only with experience.

The final decision to approve or decline a loan can be made by one of the three committees at the Head office level. These committees are Credit Department Loans Committee, Head office Credit Committee and Board of Directors Credit Committee. The discretionary limit of each of the three committees is specified in the internal policy manual of the Bank.

3.4.8 Registration of Collaterals, Insurance and Loan Contract

Once a loan request gets approved, what may be considered as the final step of the process is getting the collaterals registered and insured, getting the contract prepared and signed, and finally disbursing the loan money.

Collaterals are registered with the appropriate governmental body to ensure that the property does not get disposed without the knowledge of the bank (Immovable properties such as buildings are registered with the Municipality where as motor vehicles are registered with the Road Transport Authority). Collaterals shall also be insured against all possible risks in order to safe guard the interest of the bank.

The bank has a standard loan contract which sets out the rights and obligations of both the bank and borrowers. A term loan agreement, among other things, provides for terms and conditions of the loan, particulars on pledge or mortgage of property, a detailed description of the property

mortgaged or pledged, and the conditions under which the loan contract may be cancelled or terminated.

3.5 Credit Follow-up

Credit management does not end by approving a loan request and disbursing the cash. It requires designing an appropriate mechanism to follow-up on regular repayment of the loan as per the agreed schedule and terms of the loan contract.

The follow-up process is an ongoing activity that continues until the loan is finally repaid back with all the interest due thereon. The follow-up process helps foresee problems relating to regular repayment of loans and take timely measures to solve the problems, or at least to minimize possible losses. The follow-up activity includes both a written communication with the customer as well as direct personal visit of the customer.

Dashen Bank also has a policy of credit follow-up to ensure that loans are utilized for the intended purposes and repayment is effected in accordance with the loan agreement. The Bank actually has credit information and follow-up division at the Head office level and this division undertake the task of regular follow-up on disbursed loans. According to interviews with the Bank officials, the staffs of this division make a regular visit to area banks and consult with borrowers as well. In certain instances such visits are reported to have resulted in positive outputs by increasing the rate of loan repayment following the visit. However ensuring regular repayment of loan money in accordance with the loan agreement remains to be a challenge for the bank.

3.6 Automation Efforts at Dashen Bank

Dashen Bank officials state that one of the qualities, which make Dashen Bank unique from the other commercial banks in Ethiopia, is the effort put to automate operational activities right from the start. According to officials of the bank the search for and acquisition of relevant banking software were handled well ahead of the bank's operation launching date. As a result, some of the city area banks' operations were automated from the beginning (Business Development Department, 2001). And according to an IT (information technology) professional at Dashen Bank head office, presently operational activities are automated in all the twenty-two area banks.

The IT professional explained that the software that is put to use at the bank is called MicroBanker. The facilities provided by this software are many, which in addition to automating of operational activities includes preparation of different reports for management consumption, facility for verification of signature and also supports remote banking.

Presently, the area where the software has been utilized is in automating the saving and current account operations. In addition, it has also been used in automating the credit operation. Since the focus of this research is in the credit area, the researcher had made some inquiry as to which parts of the credit operation are automated. According to an IT department staff at the bank, MicroBanker holds borrowers information mainly in two files. These are the loan file and collateral file. The first file that is the loan file has fields including 'Amount Granted', 'Date Granted', 'Expiry Date', 'Arrear' etc. And the collateral file holds information related to collateral such as 'Security Value', 'Sum Insured' and the like.

Currently, the available two files are not used to generate reports that can be used by credit department. This is explained to be due to inadequacy of the borrowers' information that is available in these two files. It was also pointed out that it is possible to incorporate some new fields but the issue does not seem to have been properly addressed. In addition, the database that a MicroBanker supports is reported to be very inflexible. This has made report generation difficult not only in credit areas but in all other areas of the banking activity.

To address this problem IT professionals in the bank have considered shifting to another software that supports an Oracle database. One IT professional at the bank stated that this is believed to alleviate the problems in preparing different reports that could be used for decision purposes.

Data mining is discussed to benefit much from the availability of data warehouse i.e. collection of data from many different sources that is stored in a common format with consistent definitions for keys and fields (Berry and Linoff, 1997). But the database that is available now at Dashen Bank is a distributed database where remote login is possible through dialing. In the Ethiopian context, a centralized database is reported to be a challenge since networking of area banks by itself is challenging due to the poor telecommunication facility. However, these issues are being addressed by the telecommunication agency and the problem is being slowly alleviated at least in Addis Ababa. The IT professional further commented that efforts to network some of the area banks in Addis Ababa are already underway.

During automation efforts, the IT professional discussed, that there are some limited problems. These include difficulty in convincing users to shift from the practices they have been accustomed to for a long time. And the other problem is the small involvement users are making in terms of improving and developing a more efficient use of the technology. Users are only

ready to accept the minimum level of use without any further suggestions for improvement in which the automation effort is more responsive to their needs.

The purpose of this research work was to assess potential applicability of data mining technology in supporting credit decision at Dashen Bank. Therefore, an extensive study on the information technology department was not conducted except to understand the efforts put towards automation of credit data since possible application of data mining at Dashen Bank would be highly facilitated by availability of data in an electronic format. And the interview with IT professionals at Dashen Bank indicated that there is due attention in the area of computerization and all efforts are being exerted in automating the different banking operations.

3.7 Problems Identified as a Result of the Survey

The overall survey of the existing procedure for credit approval process at Dashen Bank has asserted the problem situation foreseen at the beginning of the study. Like in other commercial banks, loan collection is below the desired level. It is appropriate to note at this juncture that not much information, and particularly figure amounts, can be obtained due to the nature of the business, which requires maintaining a level of business confidentiality.

According to published reports of the bank, however, it is stated that the provision for doubtful debts at the end of the fiscal year June 30, 2000; stood to the tune of Eth. Birr 17.6 million whereas loans under litigation was to the tune of Eth. Birr 43.6. The same is true for overdue loans, which was Eth. Birr 8 million as at June 30, 2000 (Dashen Bank: Annual Report, 1999/2000). Such is an evident indication of the level of loan collection problem at the bank.

The most critical elements of the loan approval process are the stages where a given loan request is evaluated, recommended and get approved. As discussed herein above in the survey of the existing procedure, the most critical element of the loan recommendation stage is a subjective evaluation of the bank experts on what are termed as the '5 C's' for loan consideration.

While the '5 C's' are important considerations for a loan recommendation they do not specifically spell out more detailed elements for consideration and hence heavily relies on a very general subjective evaluation of the bank experts. Thus, what can be considered as one of the major problems of the existing loan approval procedure is the limited number of factors taken in to consideration for loan evaluation. The introduction of data mining technology helps in considering a number of different combinations of variables (factors) and assess which are important for credit decision. The introduction of data mining technology could further help in shortening the rather prolonged way in which a loan request has to pass thereby making the procedure more speedy and efficient.

The approach taken by Dashen Bank to deal with the problem of high default rate was to introduce tight lending criteria and approve loans at the head office level. While this might have helped to critically analyze loan requests at the head office level, it at the same time has a problem of creating a highly centralized system, which is an impediment for an efficient operation of a business.

Chapter 4

Data Collection, Preparation and Model Building

This chapter details the different data mining steps that were carried out for this research undertaking. The steps include identifying and collecting the data, preparation of data for analysis, model building and model testing. The initial step in any data mining undertaking i.e. clear definition of the problem has been addressed in the first chapter under section 1.3 and in the third chapter under section 3.7. Before going into the particulars of what was carried out for the different data mining steps, this chapter begins with an introduction of the software that was employed for model building and testing. The description is rather detailed so that to avoid repetition in the main parts.

BrainMaker Neural Network Software

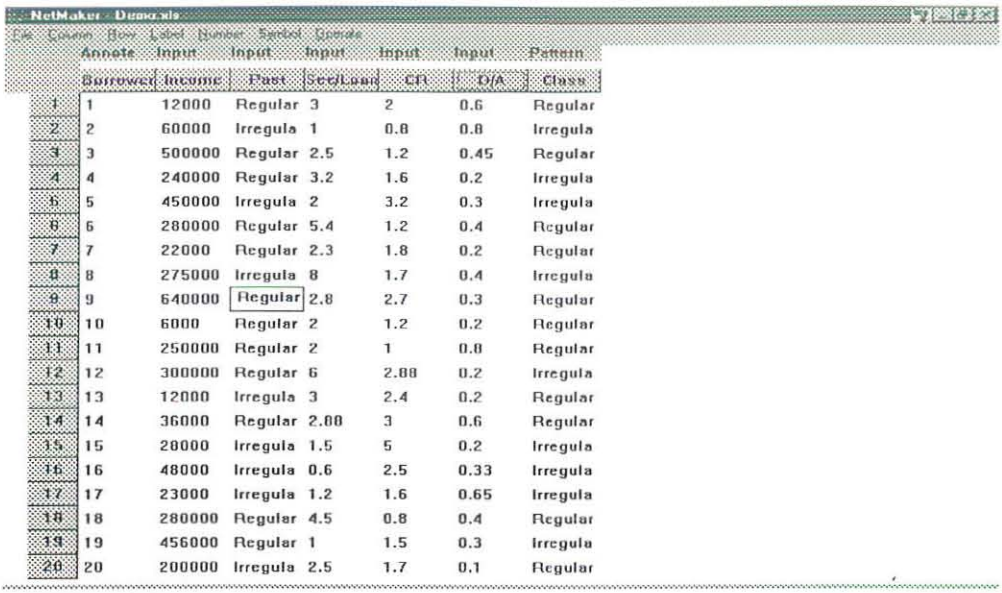
BrainMaker Neural Network Software is developed by California Scientific Software. (http://www.calsci.com/). The software vendors state that BrainMaker is the world's best-selling software for developing neural networks with more than 25,000 systems sold (http://www.calsci.com/).

BrainMaker uses back propagation algorithm in developing a model i.e. the network is trained by presenting a set of facts over and over again. Each time the network is presented with a fact it produces an answer of what it thinks the output should be. Then this answer is compared with the actual output and weights would be adjusted internally based on the error. BrainMaker goes through all the training list (records) addressing each fact in turn and making the necessary corrections. When the entire list of facts has been presented, BrainMaker starts over at the

beginning of the list. The training process is repeated until the network gets all the facts correct or until training is interrupted. Each time through the entire fact is called a 'run' (California Scientific Software, 1998).

BrainMaker has two programs called NetMaker and BrainMaker. NetMaker makes building and training neural networks easy by importing data and automatically creating BrainMaker's neural network files. In broad terms 'NetMaker is used to create BrainMaker files, convert and perform operations on data, and graph and analyze data' (California Scientific Software 1998).

NetMaker can import data from Lotus, Excel, dBase, MetaStock, ASCII and binary files. Both numeric and text data can be accepted by NetMaker and transformed into a representation that the neural network can understand i.e. in the range of 0 to 1. The imported files are seen on NetMaker as a spreadsheet. The model developer can perform calculations and can visually analyze data while in NetMaker program. Upon completion of manipulation of data, the network builder would identify which fields would be used as inputs (independent variables) and which as pattern (dependent variable) (California Scientific Software, 1998). For this research, example of an independent variable (input) could be 'Asset' value and a pattern is the classification field that indicates whether the borrower defaulted or not. In addition, a field can be labeled as annotation, which is a label used to represent fields that are not used as independent (input) or dependent (pattern) variables but as an identification of a particular record (California Scientific Software). For instance a borrower identification number could be labeled as annotation.



	Borrower	Income	Past	Sec/Loan	CR	D/A	Class
	Annote	Input	Input	Input	Input	Input	Param
1	1	12000	Regular	3	2	0.6	Regular
2	2	60000	Irregula	1	0.8	0.8	Irregula
3	3	500000	Regular	2.5	1.2	0.45	Regular
4	4	240000	Regular	3.2	1.6	0.2	Irregula
5	5	450000	Irregula	2	3.2	0.3	Irregula
6	6	280000	Regular	5.4	1.2	0.4	Regular
7	7	22000	Regular	2.3	1.8	0.2	Regular
8	8	275000	Irregula	8	1.7	0.4	Irregula
9	9	640000	Regular	2.8	2.7	0.3	Regular
10	10	6000	Regular	2	1.2	0.2	Regular
11	11	250000	Regular	2	1	0.8	Regular
12	12	300000	Regular	6	2.88	0.2	Irregula
13	13	12000	Irregula	3	2.4	0.2	Regular
14	14	36000	Regular	2.88	3	0.6	Regular
15	15	20000	Irregula	1.5	5	0.2	Irregula
16	16	48000	Irregula	0.6	2.5	0.33	Irregula
17	17	23000	Irregula	1.2	1.6	0.65	Irregula
18	18	280000	Regular	4.5	0.8	0.4	Regular
19	19	456000	Regular	1	1.5	0.3	Irregula
20	20	200000	Irregula	2.5	1.7	0.1	Regular

Figure 4: NetMaker Screen with Hypothetical Records of Borrowers

The prepared file is then saved with a .dat extension (California Scientific Software). The saved file is the basis for creating the BrainMaker files that are used to train and test a network. There are three BrainMaker files that need to be created. The first one is the definition file (*.DEF) file, which has the definition information for training such as what columns are inputs (independent variables), patterns (dependent variable) and how the information is displayed. The second is the fact file (*.FCT), which by default, constitutes 90% of the prepared data for training. The last one is the test file (*.TST), which by default has 10% of the prepared data for testing (California Scientific Software).

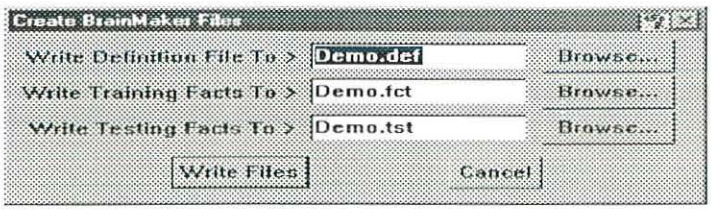


Figure 5: Screen for The Three BrainMaker Files

After the above three files are created the model developer would be ready to move to the BrainMaker program. BrainMaker program has default values for the parameters that are essential in neural network training and testing. The three most important parameters are training tolerance, learning rate and smoothing factor (momentum) (Bigus, 1996). The meanings and importance of each is described below.

Training Tolerance: Training tolerance specifies how accurate the neural network output must be to be considered correct. For example one can have an output value ranging from –20 to 30. The range for these values is 50. If a training tolerance of 0.1 is specified it would be equivalent to +-5 (10% of 50). Thus, if the training pattern is 10 and the neural network output is 15, it will be considered correct and no internal changes will be made to the network since a difference of 5 (15 minus 10) is within the training tolerance range of +-5. The software providers recommend beginning with a loose tolerance and lowering it as the network gets most of the facts correct (Bigus, 1996; California Scientific Software, 1998, Lawrence, 1994). The training tolerance concept is also applicable for testing tolerance, which is used to determine when to accept the output for the test facts as correct.

Learning Rate: In neural networks training is carried out by comparing the neural network output and the actual value and adjusting the weights depending on the error value. Learning rate determines how big a change must be made towards the correct value i.e. do we take a giant step towards the correct value (large learning rate) or small step (small learning rate). A very high learning rate is not preferred since there would be giant oscillations as the network makes large adjustments for one pattern and another large change for the next pattern. It is recommended to lower learning rate at the beginning (Bigus 1996).

Smoothing Factor (Momentum): This is a parameter that goes hand in hand with learning rate. 'The momentum parameter causes the errors from previous training patterns to be averaged together over time and added to the current error. So if the error on a single pattern forces a large change in the direction of the neural network weights, this effect can be mitigated by averaging the errors from the previous training patterns' (Bigus, 1996). In BrainMaker the default value for smoothing factor is 0.9 and the software providers state that 'adjusting the smoothing factor has not been found to reduce training time or improve prediction in every case' (California Scientific Software 1998). Therefore, in this research undertaking, the smoothing factor value had not been changed for any of the models.

The default values, in BrainMaker, for the above parameters are:

Training tolerance	0.1
Learning rate	1
Smoothing factor	0.9

Table 1: Default Parameters in Brain Maker for Training Tolerance, Learning Rate and Smoothing Factor

There are also other parameters that can be changed such as when to stop training, when to stop and make a test and how many hidden neurons and hidden layers to use etc (California Scientific Software). The default value for number of hidden layers is one and default value for number of hidden neurons is the average of the output and input neurons but in cases where number of outputs are few the number of hidden neurons are made equal as the number of input neurons. Both these values i.e. number of hidden layers and number of hidden neurons can be changed. The software providers do not recommend change in number of hidden layers. But in the case of

hidden neurons, they state that there is no hard and fast rule but recommend the following as a practical guideline (California Scientific Software, 1998).

Hidden neurons = (input neurons + output neurons)/2

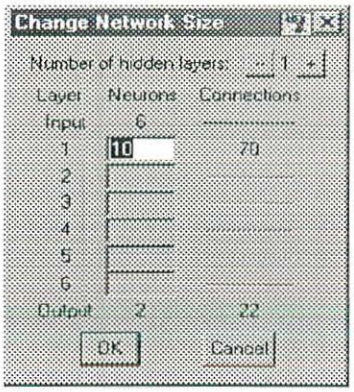


Figure 6: Screen for Network size in BrainMaker

Figure 6 shows a description of a network size in BrainMaker. This network size was created for the hypothetical borrowers' data put in Figure 4. In that table i.e. Figure 4 five columns were marked as inputs. These are 'Income', 'Past', 'Sec/Loan', 'CR' and 'D/A'. All the fields are numeric except for one field i.e. 'Past'. Each of these numeric fields is assigned one input neuron while the text fields are assigned one input neuron for each unique symbol. Therefore, in this example there are four input neurons for the four numeric fields and two input neurons for the symbolic field (The symbolic field has two unique possible outcomes i.e. regular and irregular). Thus, all together there are six input neurons. Likewise, for the output neurons since it is a symbolic field and there are two unique possible outcomes we have two output neurons.

The other concept in the network size is the connection numbers. Figure 6 shows that there are 70 connections between input and hidden neurons. Each input neuron is connected to all the hidden neurons, which bring the connection numbers to sixty (6*10). But as seen from above the total

connections between the input and hidden neurons are seventy. This is due to the presence of a bias unit. In a back propagation training a bias unit is added to the input and hidden layers. The values of the bias units is always one and they are important in avoiding the possibility of having a network where it would be impossible to change the weights in cases where all inputs are zero (a good mathematical description on how back-propagation works is available in the work of Tveter, 2000). In the above network, the presence of the bias unit in the hidden layer increases the number of connections to seventy. Also, the connections between hidden and output neurons are twenty (2*10) without considering bias unit. But since there is one bias unit in the hidden layer the total connections becomes 22. In summary the above network has 70 weights connecting the input to the hidden nodes and there are 22 weights connecting the hidden to the output nodes.

The other concept that is considered important in neural networks is which function to use. BrainMaker supports four kinds of functions namely sigmoid, threshold, step and linear functions. However, as discussed in the literature review part (under section 2.2.4), the most common function is the sigmoid function. And the software providers also state that they '...have seen no problems which train fundamentally better with anything other than the standard sigmoid transfer function' (California Scientific Software, 1998). Therefore, for all the trainings in this research work it is the sigmoid function that was used.

To start training the command is 'Operate/Train Network'. Upon the command a screen like in Figure 7 appears. While training progresses statistical information are provided on the screen such as which fact the BrainMaker is processing at a specific time, the number of facts which met and did not meet the training tolerance, the number of run (epoch) etc. There are also two graphs that display the progress of the training. The first is a histogram that shows the distribution of error over an entire run. The horizontal axis represents the error level and the vertical axis signifies the

number of output values at that level. As training progresses and fewer facts are classified as incorrect, the bars (solid boxes) move to the left. The second graph shows the progress of the error rate as network trains. In this graph the horizontal axis shows the run number while the vertical axis represent the overall error level (RMS error). For a good training the error value (RMS error) would decrease as the number of runs increases (California Scientific Software, 1998).

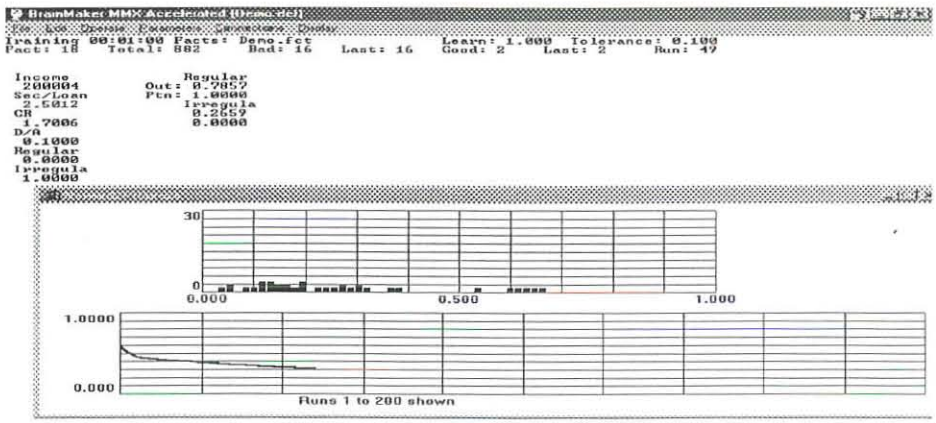


Figure 7: BrainMaker screen While in Training

Training can be stopped at any time before the instruction to stop training is met. And the model developer can test the network and save it if it results in an acceptable prediction rate. Or it is possible to wait until training stops and test the model (network). The default for stopping training is when the incorrect classification of facts in a run (epoch) becomes zero. The software vendors discuss that better network (model) is usually obtained before the criteria for stopping training are met (Lawrence, 1994). Therefore, it is advised that the model developer periodically saves a network.

BrainMaker Neural Network Software has the facility whereby data are classified into two sets i.e. training and testing set. The software automatically puts aside 10% of the records into a testing file. These data would be used to test the accuracy of the network and would not be seen by the network during training (California Scientific Software 1998). This step i.e. dividing of the data into training and testing records is carried out for every network that is experimented with.

Dividing of data set into training and testing sets is also the approach suggested for model building by different writers such as Bigus (1996) and Lawrence (1994).

Among the saved network the one with the best result would be selected and a running fact file would be developed that can be used in predicting future records (California Scientific Software).

The above in brief explains the steps involved in building a model using the BrainMaker Neural Network Software. In Figure 8 the different steps that are stated to be requirement so that to be able to build models using BrainMaker Neural Network software are put.

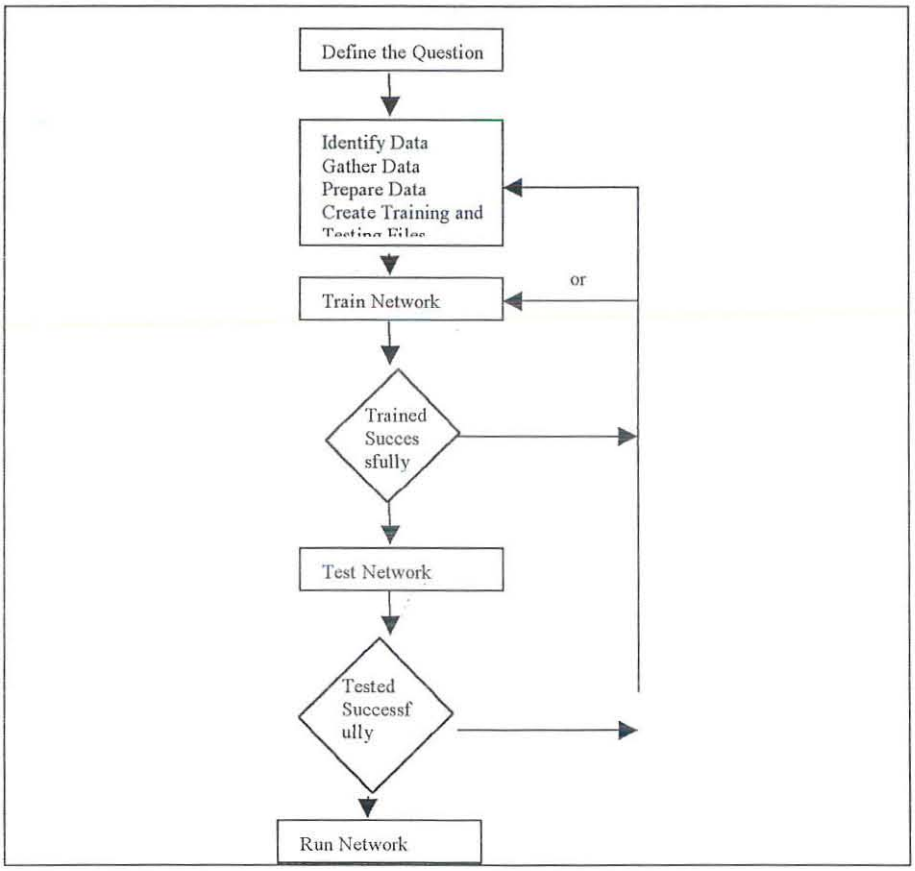


Figure 8: Essential Steps for Model Building as put by Providers of Brain Maker Neural Network Software

Brain Maker software's applicability has been experimented in different areas including business, medicine, manufacturing, speech recognition, optical character recognition, sports etc (California Scientific Software). The specific task, which is being undertaken in this project i.e. evaluation of

credit worthiness of loan applicants (credit scoring) is also reported to have already been experimented by BrainMaker. (http://www.calsci.com/CreditScoring.html). The model was developed using data of not more than 100 applicants and the resulting models are stated to have achieved accuracy rate of 75-80%. This experiment had eight inputs namely Own/Rent home, Years with Employer, Credit Cards, Store Account, Bank Account, Occupation, Previous Account and Credit Bureau. And the output had three possible outcomes for the loans namely delinquent, charged-off and paid-off.

For the above reasons, BrainMaker was considered ideal for this project. The researcher had considered a number of other software before deciding on BrainMaker. Some of them are NeuNet Pro (http://www.cormactech.com/neunet/), NeuroSolutions4 (http://www.nd.com/products/nsv3.htm), EasyNN (http://www.tropheus.demon.co.uk/easynn.htm) and ThinksPro (http://www.sigma-research.com/bookshelf/rtthinks.htm). These software were not considered applicable for different reasons, which mainly are price and complexity considerations.

The steps followed for this project in brief are to identify the problem, collect data, prepare data and train and test models. The BrainMaker software is employed in training and testing of models.

First, different independent variables that were suspected to affect regular repayment of the loan were identified in consultation with bank experts. Then sample area banks were selected as a source of data for training and testing the models. The collected data were used to obtain the inputs (independent variables) as well as the pattern (dependent variable). Examples of inputs are current ratio, asset value, trade sector etc. and the dependent variable (pattern) was the particular classification a specific loan fell to. Three possible classifications were created in consultation

with bank experts. These are regular, substandard and doubtful loans. The meanings of each is described in section 4.1.1.

Numerous networks (models) were thus trained and tested by using the historical data collected and the researcher was successful in developing a number of good networks as discussed in section 4.3

4.1 Identifying and Collection of Preclassified data

Berry and Linoff (1997) state that the basic requirement for knowledge discovery is good data. And the ideal source of data is identified to be the corporate data warehouse. The authors describe data warehouse to be collection of data from many different sources that is stored in a common format with consistent definitions for keys and fields. Unfortunately a corporate data warehouse is not available at Dashen Bank S.C.

The only available source of information that held past data on people who defaulted on a loan and those who did not was existing in a manual format. At Dashen Bank head office there are basically two kinds of documents that held such information. The originators of these documents are the different area banks. The first is the loan approval form that becomes available prior to disbursement of the loan. This form, as discussed in Chapter 3 section 3.4, has summarized information on a potential borrower. These include financial information (balance sheet and income statement), whether the borrower is a new customer or not, years in business etc. (Format is attached as Annex 1)

The other sources of information were the monthly credit reports that reach the head office from the different area banks every month. The primary purpose of this report is to show the amount collected from each customer and whether the borrower is behind schedule or not. (Format is attached as Annex 2)

The above two documents together provided the required information to experiment on the possible application of data mining activity at Dashen Bank.

In order to collect the information, sample area banks were chosen considering the time available for the research undertaking. Presently, Dashen Bank has twenty-two area banks but the area banks that were considered for sampling were thirteen. The thirteen area banks were chosen taking into account the years they have been in operation. All the thirteen area banks had been in operation for more than four years while the other nine area banks were opened in the past two years. Their relative added years in operation made the thirteen area banks better candidates for data mining activity since data mining requires large data.

The thirteen area banks are found in different parts of the country, which enabled a fairly good geographic representation.

Area Banks	**Location**
Awassa and Dilla Area Banks	Southern Ethiopia
Bahr Dar, Dessie and Mekelle Area Banks	Northern Ethiopia
Jimma Area Bank	Western Ethiopia
Dire Dawa and Nazareth Area Banks	Eastern Ethiopia
Golla, Kality, Kerra, Main and Tana Area Banks	Addis Ababa

Table 2: Spatial Distribution of Area Banks Considered for Sampling

A report obtained from Dashen Bank Credit Department shows that as at February 28, 2001 40% of Dashen Bank's borrowers were from Addis Ababa. The researcher thus was concerned that random sampling from this area would significantly reduce the proportion of records from other areas. Therefore a purposive sampling method was used where two area banks in Addis, with the smallest number of records were taken. These were Kality and Tana Area Banks.

But for all the other regions high volume and availability of data were the considerations in choosing the samples. Accordingly, Dilla, Mekelle, Jimma and Nazareth Area Banks were chosen.

From among the four types of loans that are available at Dashen Bank, 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 installments 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 area banks. Moreover, term loans are the more common loans at the Bank and they constitute major proportion of the total outstanding loans. Report obtained from Dashen Bank Credit Department shows that, as at April 30, 2001 the composition of term loans is 58% of the total loan.

The software chosen i.e. Brain Maker has a facility to import data from Lotus, Excel, dBase, MetaStock, ASCII and binary files. Among the choices available collection of data using Excel was preferred due to ease in calculation and graphical facilities.

Collection and keying in of information consumed quite a considerable time of about one and a half month. All the information was available in a manual form and thus had to be collected

manually and fed into a computer. In Credit Department at the head office, both sources of information i.e. the loan approval form and the monthly credit reports are chronologically filed in box files and each area bank's data are filed independently. Fields to be collected from the documents were identified in consultation with the bank experts basing their experiences (The format used for collecting data with some hypothetical records is attached as Annex 4).

The following fields were collected from the loan approval form of each borrower
(N.B. Explanation for some of the following terms is provided in the glossary).

- Asset
- Total liability
- Current Asset
- Current Liability
- Yearly income
- Business Establishment Year
- Sex
- Trade Sector
- Number of Prior Loans (Number of loans the borrower settled in the past)

Fields that were available in the loan approval form but not considered include 'Facility Requested', 'Date of Application', 'Property Owned By', 'Other Line of Business' etc. (Annex 1).

And the monthly credit report provided the following information about each borrower.

- Name of Customer
- Data Granted
- Expiry Date
- Amount Granted
- Security Type
- Security 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)
- Area

Fields that were not taken from this document (Monthly Credit Report Form) include 'Extension Date, 'Amount Collected', 'Amount Outstanding', 'Arrears' and 'Repayment Amount' (Annex 2).

The above variables were collected with different purposes in mind. Some of the variables were to be used as independent variables for model building while others were to be used to derive other important variables. And there were variables that were used for both purposes. The derivation of the other independent variables is discussed below in section 4.2.4.

The dependent variable i.e. classification of a loan, was inferred from the monthly credit reports. This is the variable that was the most time consuming to obtain 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 checked to see whether the borrower was regular in repayment or not.

A given loan was classified in one of the following three categories. This classification was developed in consultation with the bank experts. All commercial banks in Ethiopia report the status of their loans to National Bank of Ethiopia under four headings. These are regular, substandard, doubtful and loss loans. These classifications go from the good (regular) loans to the worst loans (loss). The classification adopted for this research is also a small modification of these classifications. Instead of having two classifications for the loans that were in arrears they were joined into one and put under the classification 'Substandard'. And the label for the 'Loss' loans under National Bank's classification was replaced for this research with 'Doubtful'. According to the bank experts the classification put under loss loans are not really considered as a loss for the banks as the name suggested. That is why the label 'Doubtful' was preferred.

Classification	**Explanation**
Regular	Loans repaid with regular repayment
Substandard	Loans settled on time but had arrears of more than 3 months
Doubtful	Loans with an overdue balance that is worth a three-month repayment. And the loan has to stay as overdue for more than four months to be classified as doubtful. This category also includes loans that were pending under litigation and loans which had been rescheduled but still not settled.

Table 3: Description for the Three Classification of a Loan

The other point that had to be addressed, during collection of data, was which time range to consider. Dashen Bank started its credit operation in February 1996 with ten area banks. And it was in 1997 that all the thirteen area banks mentioned above became operational. Therefore, for uniformity sake, loans granted from January 1997 onwards up to August 1999 were taken. In addition, loans granted in 1996 were excluded since it was the first year of operation for the bank and many credit procedures, rules and regulations were not yet refined and developed and hence the loans disbursed in 1996 do not represent the realities that exist now.

Loans after August 1999 were not considered because most of the loans are still outstanding and their outcome is still not known. It has already been discussed that the record of borrowers are collected in order to be able to develop a model, which can predict the likelihood that a future borrower would default, or not. And this model is developed by learning from past history of borrowers' performances. Therefore, loans whose fate i.e. whose classification are not already known cannot be used.

The number of records collected from the six sample area banks is summarized in the following table i.e. Table 4.

Area Banks	**Number of Records Collected**
Dilla	126
Jimma	107
Kality	51
Mekelle	415
Nazareth	79

Tana	79
Total	**857**

Table 4: Number of Records Collected from the Six Sample Area Banks

The software vendors recommend that number of facts should be:

Training facts = 2 * (inputs+hidden+outputs) to 10*(inputs+hiddens+outputs)

As will be discussed in the next sections, the maximum network size obtained had 88 inputs, 88 hidden and 3 output neurons and according to the above formula, number of records should preferably be between 358 and 1790. And the collected data i.e.857 was already within this range though closer to the lower side. Therefore, before going to the next further steps the researcher wanted to add some more records so that total records would be close to the average number of facts recommended by the software providers. Accordingly, data of two area banks was decided to be added. From the thirteen area banks considered the ones left out were Bahr Dar, Dessie, Dilla, Dire Dawa, Golla, Kerra and Main Area Banks. And for reasons stated above, the researcher did not want to take more samples from Addis Ababa and hence Golla, Kerra and Main Area Banks were excluded. From the remaining four area banks Awassa and Bahr Dar were added considering number of records and availability of data. Number of borrowers' records collected from Awassa and Bahr Dar Area Banks were 102 and 43 respectively. The inclusion of the two area banks increased the number of collected records from 857 to 1002.

4.2 Preparing Data for Analysis

This is the second step of the data preparation step as described in the data mining methodology of Berry and Linoff (1997). The authors describe that collected data has to be massaged into a form that will allow the data mining tools to be used to best advantage. And the transformation depends on the technique to be used and the particular software package to be employed. However, they also identify five steps that are common to almost all techniques and these steps are applied as in the following.

4.2.1 Summarization

The reason for summarization is that there might exist only few examples at the finest level of detail. This became important for the 'Trade Sector' field, which had more than 54 categories as collected from the loan approval form of the different area banks. But this kind of very fine distribution of observations would make it impossible to have sufficient records in one bin (Berry and Linoff, 1997). For instance some of the sectors that were put in one bin, considering their similarity, along with their new labels are given below:

Small Scale Manufacturing	**Food Items**	**Clothes**
Black Smith	Botchery	Boutique
Gold and Silver Smith	Food Items Retail	Shoes
Metal and Wood Workshop	Bakery	Tailoring
Carpentry	Pepper and Spice	Textile
Metal Workshop	Honey	Woolen Thread
	Salt	Leather and Shoes
		Leather Products

After lumping up related sectors together the resulting category of trade sectors became nineteen:

- Fertilizers
- Building Materials
- Electronics
- Sundry Goods
- Clothes
- Cosmetics and Jewelry
- Small scale Manufacturing
- Building and Construction
- Hotel
- Big Scale Manufacturing
- Furniture and other house hold Goods
- Cereals
- Clinic and Pharmacy
- Coffee
- Photography
- Import
- Spare Parts
- Transport
- Food Items

4.2.2 Inconsistent Data Encoding

'When information on the same topic is collected from different sources, the various sources often represent the same data different ways' (Berry and Linoff, 1997).

As explained in section 4.1 above, the two types of documents that served as sources of information are made available to the Head Office from the different area banks. And these area banks use non-uniform data encoding for some of the fields. This was a problem for three fields i.e. 'Trade Sector', 'Business Establishment Date' and 'Security Type'. Other fields such as Term of Payment' and 'Sex' also needed minor adjustments due to differences in data encoding.

For instance some area banks represented 'Term of Payment' as 'Monthly' others as 'M' or 'EM' (every month). But all three represent the same thing. And values, which represent the same thing in different ways, had to be given a uniform value.

For the Trade Sector' field also, there were many instances where the same sector was encoded by different area banks differently (E.g. Sundry goods, sundry articles, sundry, shop, etc. represented the same kind of business).

The problem with 'Business Establishment Date' was that some of the values for this field were represented using the Ethiopian calendar while for others the Gregorian calendar was used. Thus, to make values uniform all the values were changed to the Ethiopian calendar.

The other fields that had data encoding problem were corrected by uniformly following the following representations.

Field Name: Security Type

Building	Loans secured against building only
Vehicle	Loans secured against vehicle only
BV	Loans secured by both building and vehicle
PG	Loans secured against personal guarantor
BP	Loans secured by both building and personal guarantee
VP	Loans secured by both vehicle and personal guarantee

Field Name: Sex

Male	Male borrower

Female	Female borrower
MF	The loan is in the name of both male and female borrowers
CP	A loan disbursed to a company or partnership

Field Name: Term of Payment

Monthly

Bimonthly

Quarterly

4.2.3 Missing Values

In the course of collecting records many missing values were experienced. Two Crows Corporation (1999) suggests possible ways for handling such missing values and it was their approach that was adopted for this project.

For continuous variables (variables with numerical value such as age) it is suggested that missing values be replaced with the mean value for that field (Two Crows Corporation, 1999). This approach was used for the 'Years in Business' field. The average value for this field was 8.9 years. Therefore, all missing values for 'Years in Business' field were replaced with 9 years.

But for the other continuous fields such as 'Asset', 'Liability' and 'Capital' it was difficult to use this method i.e. the researcher as well as the bank experts did not find it logical to assign an average business worth or liability of other businesses to a business with missing data. For instance, it would be unreasonable to calculate the mean asset value and assign this to a missing value. Therefore, for the continuous variables, which represented financial value another approach

had to be adopted. In the case of second time borrowers financial information of the previous loan was considered. But in the case of new borrowers the whole record had to be deleted since it was difficult to assign an arbitrary value for the missing financial information. But these instances were only few and not more than ten.

And one field representing yearly income, however important, had to be completely discarded for three reasons. First, the problem of bias between estimators were exaggerated for this field. Second, it was difficult to determine if the values provided represented monthly or yearly figure. When the Loan Approval Form (LAF) is used for decision purposes these confusions are cleared by communicating the bank official who is responsible in preparing the LAF. But for this research purpose this was difficult to do considering the cost of communication and the time gap since the person responsible has prepared the LAF. The third reason for exclusion of the 'Income' field was the instances of many missing values for this field.

The other types of fields are the categorical fields. Such fields can be grouped into nominal and ordinal variables. The ordinal variables are categorical variables whose values have meaningful order (E.g. High, medium and low). And nominal variables refer to categorical variables whose values are unordered (E.g. Postal Codes) (Two Crows Corporation, 1999). The suggested method for handling missing data for categorical variables is to take the median for ordinal variables and the modal values for nominal values (Two Crows Corporation, 1999).

Variables whose missing values were handled in the above way include Trade Sector', Security Type', 'Term of Payment', 'Sex' and 'Classification'. All the mentioned fields are nominal variables except for the last one, which is an ordinal variable. For the nominal variables the

missing data were filled with the modal value. The modal values for the four nominal values is given in the following table.

Field Name	Modal Value
Trade Sector	Clothes
Security Type	Building
Term of Payment	Monthly
Sex	Male

For the last field i.e. 'Classification' field, Two Crows Corporation approach indicate that missing values be replaced with median value as it is an ordinal variable. The three possible values for this field are three as discussed in Section 4.1 which are 'Regular', 'Substandard' and 'Doubtful'. And the classification goes from the good ones to the bad ones. The median value here is 'Substandard' thus missing values for this field were replaced with this value i.e. 'Substandard.'

4.2.4 Deriving Other Fields from Existing Ones

Berry and Linoff (1997) write that 'by adding fields that represent relationships in the data that we know from experience are likely to be important, we can increase the chance that the knowledge discovery process will yield useful result.'

According to the bank officials and a financial management book consulted (Higgins, 1998), there are many ratios and values that are considered essential in determining the credit worthiness of an individual. These include current ratio, debt to asset ratio and net working capital. These three financial indicators were derived from the existing fields and added as new fields. In addition,

other fields were also computed and included after a discussion held with Dashen Bank credit department staff.

The total fields considered then became 26. The field names along with their explanations and source are given below i.e. Table 5.

No	Name of Variable	Description	Source
1	Area	Name of the area where the Area Bank is located	Crédit Report Form
2	Loan No.	The number of loan for the specific borrower (1st loan, 2nd loan, 3rd loan etc.)	Computed
3	Month	Month loan was granted	Computed
4	Duration	The duration of loan in number of days	Computed
5	Yearly Payment	The estimated amount to be paid in a year	Computed
6	Amount Granted	Amount granted	Credit Report Form
7	Loan/Time Ratio	Loan amount divided by the loan duration	Computed
8	Asset	Total asset of the borrower	Loan Approval Form
9	Capital	Total capital of the borrower	Loan Approval Form
10	Current Asset	Total current asset of the borrower	Loan Approval Form
11	Current Liability	Total current liability of the borrower	Loan Approval Form

12	Net Working Capital	Current asset – Current Liability	Computed
13	Liability	Total liability of the borrower	Loan Approval Form
14	Debt/Asset Ratio	Liability value divided by asset value	Computed
15	A. Debt/Asset Ratio	The anticipated debt/asset ratio after considering the new loan to be granted.	Computed
16	A. Current Ratio	The anticipated current ratio after considering the new loan to be granted.	Computed
17	Security Type	Type of security (E.g. Building, vehicle, personal guarantee)	Credit Report Form
18	Security Value	Estimated value of the security	Credit Report Form
19	Security/Loan Ratio	Security value divided by amount granted	Computed
20	Sex	Sex of the borrower	Loan Approval Form
21	Trade Sector	The kind of business borrower is engaged in E.g. Hotel, cereals etc.	Loan Approval Form
22	Years in Business	The number of years the borrower has been in business	Computed
23	Term of Payment	Monthly, bimonthly or quarterly	Credit Report Form
24	No. of Prior Loans	The number of loans borrower has settled in the past	Loan Approval Form
25	Per. of Prior Loans	Performance of past loans i.e. whether past loans were regular or not	Credit Report Form

26	How Early	Indicates whether a previous loan was settled on time or before its due date.	Credit Report Form

Table 5: List of the Independent Variables (Inputs) Considered for Model Building (For Easy Reference this Table is Attached as Annex 4)

Explanations on how the computed fields were derived are given below.

Loan No. = (Per. of Prior Loan + 1)

Month: 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)

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 = (1994 - Business Establishment Year)

4.2.5 Preparing the Data into a form that is Acceptable to the Neural Network

Neural networks accept values only in the range of 0 to 1 or –1 to +1. Therefore all values in the data set have to be represented with values ranging from 0 to 1 or –1 to +1.

The Brain Maker software that was used for this project has a facility to automatically transform values into a form that can be understood by the neural network i.e. in the range of 0 to 1. As discussed in the first section of this chapter one input neuron is assigned for every numeric field. And the values in the numeric field are scaled down to the range of 0 to 1.

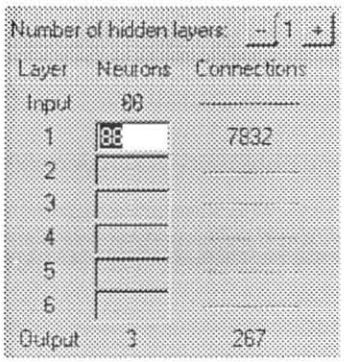
In the case of text fields the number of input neurons are equal to the number of unique fields in that field. For instance, in the 'Term of Payment' field there are three unique symbols. These are 'Monthly', 'Bimonthly' and 'Quarterly'. Therefore, three input neurons are assigned for this field and the presence of a symbol in input data turns that neuron on (California Scientific Software, 1998).

4.3 Building and Training of Models

The data prepared in an Excel format was imported to NetMaker (See Annex 6 for further details on the prepared data). This data had the records of the eight area banks constituting of 1002 records.

All the default parameters were accepted as they were. As explained in the BrainMaker software description part, the parameters include training tolerance, learning rate, smoothing factor,

number of hidden neurons etc. These parameters can be changed but for the first trial the defaults were taken since the software providers state that the default parameters are sufficient for most problems (California Scientific Software, 1998). The number of hidden neurons was also accepted as suggested i.e. 83. This network had the following network size. The explanation for the interpretation of network size is available in the first section of this chapter.



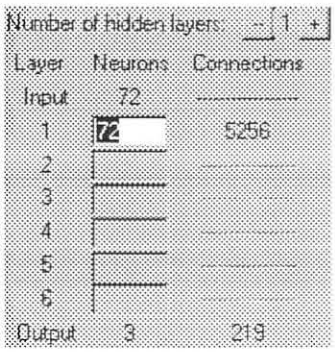
Then upon the command 'Operate/Train Neural Network' training resumed. However, after about 800 facts it became difficult for the network to learn the other 202 facts and the two displays, which are supported by Brain Maker in visualizing the progress of the network, indicated that there was no improvement. Therefore, this attempt had to be abandoned. However, during training a network had been saved and this network resulted in poor precision rate of 51% (51 of the 100 test facts were classified correct) at testing tolerance of 0.4. And at testing tolerance of 0.2 the saved network had precision of 46% (46 of the 100 test facts were classified correct). And upon the decision to stop training another network was saved. This network had precision rate of 50% and 42% at testing tolerances of 0.4 and 0.2 respectively.

The researcher then tried varying the different parameters. The parameters were changed in accordance with a suggestion provided by different writers as described in the first section of this paper.

Training tolerance	0.3
Learning Rate	0.6
Smoothing Factor	0.9
Number of hidden neurons	45

This network did not converge but was still tested at different intervals and resulted in a less performing network than the above. Three of these networks tested with accuracy of only 36%, 42% and 43% at a loose testing tolerance of 0.4. And at a tighter testing tolerance of 0.2 the networks resulted in precision of 42%, 32% and 34%.

Next varying the composition of the inputs was tested if it could result in a better performance. This was done in consultation with experts in the area who had better experience and understanding as to which variables are more essential. However, a better result was not obtained. One of these trials was a network which had 20 of the 26 variables listed in Annex 5. The excluded five variables were 'Loan/Time Ratio', 'Security Type', 'Sex', 'Term of Payment' and 'No. of Prior Loans'. And the default parameters were used. This network had the following network size.



Several networks were saved during training. However, the highest precision obtained from these networks was only 44% at testing tolerance of 0.4.

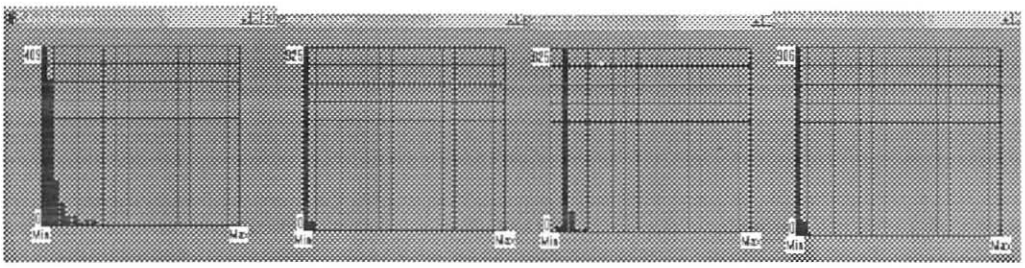
As shown above, the different trials did not result in a very encouraging result. Therefore, the researcher tried to determine possible explanations. Upon visualization of the data using a histogram it was observed that most of the variables had outlier values especially on the side of the higher values i.e. there were records whose values were very far from the commonly found values. Lawrence (1996) discusses that when outliers exist in a data set 'the neural network will have more trouble distinguishing between the common values if it has to take into consideration some unusual and extreme values as well.'

Upon further analysis it was observed that the customers who had an outlier value for one field would most likely have an outlier value for another field. For instance, a borrower with very high asset value is likely to have a potential to obtain a large amount of loan and have a high valued security (collateral). This makes most of the values for that particular customer to be outliers. Thus, it was clear that removing of records with an outlier for a particular field would reduce the problem of having unevenly distributed data.

One such field was 'Total Asset Value'. This field had maximum value of Br. 45 million while the minimum value was Br. 10,400. However 90% of the borrowers had total asset value of less than Br. 1 million. Therefore, the records of all borrowers with total asset value of more than Br. 1 million were excluded decreasing the number of records to 898 from the original number of records of 1002. The resulting file had 557 (62%), 156 (17%) and 185 (21%) records classified as regular, substandard and doubtful respectively. However, there were still outliers for some of the fields especially 'Security Value' and 'Security/Loan' ratio. According to discussion with the experts, this was mainly due to differences in estimation in areas within and outside Addis. Buildings in Addis Ababa are estimated at a very high value compared to buildings outside Addis

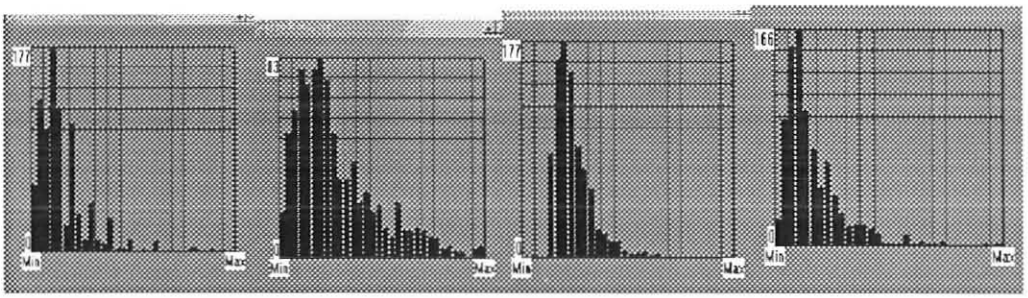
Ababa. For instance, the average building value for records of borrowers in Addis Ababa was Br. 1,086,465 while for the outlying branches it was Br. 172,711.

The problem of having such outliers were fixed with the BrainMaker facility. This facility has a way of visualizing the data using a histogram and then providing new values to be used as minimum and maximum figures. Thus, values outside the new ranges would take the newly set values. (California Scientific Software, 1998) The maximum value for 'Security/Loan Ratio' was 36.8 while the minimum was 0. The maximum value was then reduced to 15. And for the 'Security Value' field the minimum value was zero and the maximum value was Eth. Birr 1.07 million. The maximum value was then reduced to 900,000. The histograms for four fields before and after adjustment of outliers clearly show the improvement achieved by the adjustment.



Amount Granted Asset Net Working Capital Security Value

Histograms for Four Fields before Removing The Outliers



Amount Granted Asset Net Working Capital Security Value

Histograms for Four Fields after Removing The Outliers

The resulting file of 898 borrowers was trained using the original 26 fields listed in Annex 5 as independent variables (input) and classification column as dependent variable (pattern). And for this training, all the default parameters were used. However, training did not progress well and two of the networks saved resulted in accuracy of 37% and 43% even at a loose testing tolerance of 0.4.

The other problem that was then suspected is that the uneven representation of the outcomes might have resulted in a poor performance of the network. It has already been put that the distribution for the three possible outcomes is 557 (62%), 156 (17%) and 185 (21%) for regular, substandard and doubtful loans respectively. Lawrence (1998) put that 'it is especially important for the output column to be well distributed. If your collection of data contains predominantly one case over another or severely lacks examples of a particular outcome, you should collect more examples of the minimally represented cases.'

The classification column had two categories for irregular loans. These are substandard loans that represent irregularly repaid loans which however were settled on time and there were the doubtful loans whose due date has significantly passed before they were settled. But basically both two types of classifications represent irregularly paid loans so it seemed reasonable to put the two classifications together. This resulted in a record that has 557 (62%) regular loans and 341 (38%) irregular loans. Though the new arrangement still had uneven distribution of the outcomes still it was checked if it could result in a good network (model).

The first trial was carried out with all the 26 inputs listed in Annex5 and with all the default parameters. The best network that was obtained from this trial had precision rate of 60% at testing tolerance of 0.4.

The default values were then changed based on discussions put in the first section of this chapter.

Training tolerance	0.3
Learning Rate	0.6
Smoothing Factor	0.9
Number of hidden neurons	44

Two of the above parameters were changed during training. Training tolerance was reduced to 0.2 while the learning rate was gradually increased to 1. The best network that resulted from the above trial had precision rate of 64% at testing tolerance of 0.4.

Another point considered was that similar facts might have been grouped together. The software providers put that neural networks might have trouble grasping the overall solution if the facts are grouped by similar inputs or output. (Lawrence, 1994) In such cases it is recommended that rows be shuffled and this facility is available in BrainMaker. Therefore, another set of records were experimented with after shuffling the facts. The test results from this training was also not satisfactory and the highest precision rate was only 61% even at a loose testing tolerance of 0.4. Variations of independent variables, together with the shuffling, were also experimented but still an encouraging result was not obtained.

Then the researcher considered building of a model for each Area Bank independently. This idea seemed feasible for one basic reason. The practice in Ethiopian banks is to make estimation for financial information such as asset, liability, capital and income instead of using reliable information from proper records. This is a general problem of all banks due to lack of proper books of accounts. This is believed to affect the training progress since it would be difficult to detect patterns from values that are not measured by uniform measurement but in fact depend on the estimator's bias i.e. some estimators could be very conservative while others may not. But if only one area bank, preferably one that has been managed by a single area bank manager through out its operation life is used this bias could be reduced. Awassa Area Bank met these criteria and therefore the researcher started experimenting with the data from Awassa Area Bank, which had 102 records. In addition, Awassa Area Bank was favored because it was possible to obtain fair representation of regular and irregular loans since the collected data from this area bank had 49 regular loans and 53 irregular loans.

In consultation with the bank experts, ten variables that were considered more important were chosen for the model building. These were 'Loan No.', 'Amount Granted', 'Asset', 'Net Working Capital', 'A. Debt/Asset Ratio', 'A. Current Ratio', 'Security Value', 'Security Ratio', 'Years in Business' and 'Per. of Prior Loans'. 'While in principle some data mining algorithms will automatically ignore irrelevant variables and properly account for related columns, in practice it is wise to avoid depending solely on a tool. Often your knowledge of the problem domain helps you make many of these selections correctly' (Edelstein, 1998). This is why the number of independent variables was reduced to ten variables that were considered important among the 26 fields in consultation with bank experts.

The training for this network did not progress well and after about half of the facts have been learnt the RMS error stopped decreasing. The analysis of the variables used for the above model building indicated that most of the variables used were financial values. As already stated above there was a doubt that these financial values may not be dependable, as they are not measured by uniform measurement. Thus, the researcher considered using more non-financial variables for the model building. The suggested good candidate by the bank experts was 'Trade Sector.' Thus, trade sector together with the above-identified ten variables was used for the second model building. And indeed an improved performance was obtained that resulted in a network that had precision rate of 80% (8 of the 10 test cases were classified correct). This was achieved even at a tight testing tolerance of 0.1. Hence, for further trials the researcher decided to use both financial and non-financial values. Numerous networks were then developed by varying the number and composition of inputs (variables).

One of these trials resulted in a network that had precision rate of 90% (with 9 of the 10 test cases classified correct) at testing tolerance of 0.4 and precision rate of 80% (8 of the 10 test cases classified correct) at testing tolerance of 0.3 and 0.2. The variables used for this network were 'Loan No.', 'Month', 'Loan/Time ratio', 'Amount Granted', 'Asset', 'Net Working Capital', 'A. Debt/Asset Ratio', 'A. Current Ratio', 'Security Value', 'Security/Loan Ratio', 'Trade Sector', 'Years in Business' and 'Per. of Prior Loans'. Then the researcher tried to determine if a better network could be obtained by excluding one of the independent variables in turn and training different networks. The exclusion of some of the variables resulted in a less performing networks, which signified the importance of the excluded variable. And for some variables their exclusion resulted in a better performing network. A summary of the precision rate observed for the different variables at different testing tolerances is provided below.

Testing Tolerance → / **Excluded Variable ↓**	0.4	0.3	0.2	0.1
Loan No.	90%	90%	90%	90%
Month	70%	70%	70%	70%
Loan/Time Ratio	100%	90%	80%	70%
Granted	90%	90%	80%	80%
Asset	100%	90%	90%	90%
Net Working Capital	90%	90%	90%	90%
A. Debt/Asset Ratio	90%	80%	80%	60%
A. Current Ratio	90%	90%	90%	90%
Security Value	90%	90%	90%	70%
Security Ratio	90%	90%	90%	70%
Trade Sector	80%	80%	80%	80%
Years in Business	90%	80%	80%	70%
Per. of Prior Loan	90%	90%	90%	90%

Variables, which were identified to be important in the above way, were 'Month', 'Amount Granted', 'A. Debt/Asset Ratio', 'Trade Sector', and 'Years in Business.' Incidentally all these variables are free from the estimator's bias except for one i.e. 'A.Debt/Asset ratio'. And even for this variable the estimator's bias is reduced, as it is a ratio. This observation becomes more evident when we notice that the best network resulted when 'Asset' was excluded which is a variable clearly affected by the estimator's bias. Therefore, for further trials the researcher decided to concentrate more on variables, which were free from estimator's bias.

Other trials were carried out to assess if a better result could be obtained. Some of these trials constituted of varying of parameters. For instance, two models were developed using the variables that were identified to have been important for the model building but by employing different parameters. The first network developed in such way had as inputs 13 variables namely 'Loan No.' 'Month', 'Loan/Time ratio', 'Amount Granted', 'Asset', 'Net Working Capital', 'A. Debt/Asset Ratio', 'A. Current Ratio', 'Security Value', Security/Loan Ratio', 'Trade Sector', 'Years in Business' and 'Per. of Prior Loans'. This network had 38 inputs and 2 outputs and was trained with the following parameters. The parameters were changed on the basis of the suggestions discussed on the first section of this paper.

Training tolerance	0.3
Learning Rate	0.6
Smoothing Factor	0.9
Number of hidden neurons	20

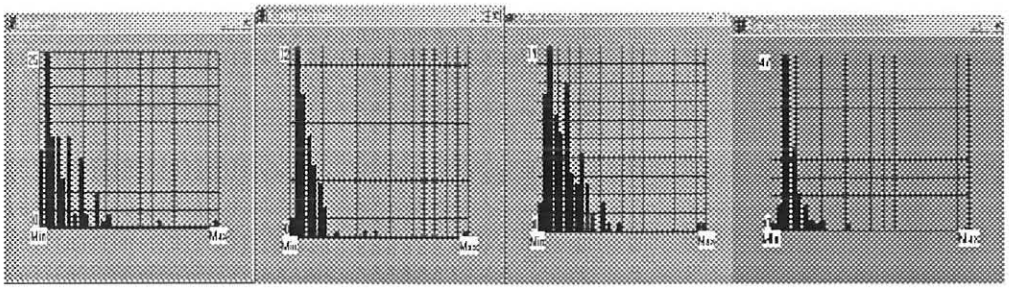
During training, learning rate was gradually increased to 1.The model developed with the above parameters and inputs resulted in a less performance. This network tested with accuracy of 100% (i.e. all the 10 test cases were classified correct) both at 0.4 and 0.3 testing tolerances but when the testing tolerance was reduced the performance significantly decreased and the network resulted in precision rate of only 40% at testing tolerances of 0.2 and 0.1. This is explained by the fact that the network was trained with loose training tolerances.

The parameters were then slightly varied and tested. For instance training tolerance was reduced to 0.1 while the other parameters remained the same as in the above network. This network

resulted in a performance that was identical to a previous network that was trained with the same inputs but with the default parameters i.e. at testing tolerance of 0.4 accuracy was 100% (all 10 test cases were classified correct) while for testing tolerances of 0.3, 0.2 and 0.1 accuracy was 90% (9 of the 10 test cases were classified correct). Numerous other experiments were also carried out by varying the parameters on the basis of suggestions made by writers as explained in the first section of this paper. However, it did not become possible to create a better network than those that already were created using the default parameters.

From all the above experiments it became clear that varying of the parameters was not resulting in a significant change to the performance of the network. Thus, for the following experiments the researcher concentrated more on varying the inputs, shuffling of records and accounting for outliers than varying of parameters. The software providers also put that the default parameters are adequate for most problems (California Scientific Software, 1998).

One more trial was carried out for Awassa Area Bank's records to see if performance can be further improved from the previous networks. And this was to check for the existence of outliers, which makes the network unable to distinguish between common values that lie in the middle ranges as discussed above. Therefore, by using a histogram the existence of outliers were checked. Four fields namely 'Amount Granted', 'Net Working Capital', 'Security Value' and 'Security/Loan Ratio' were identified to have outliers on the maximum side as shown in the following histograms.

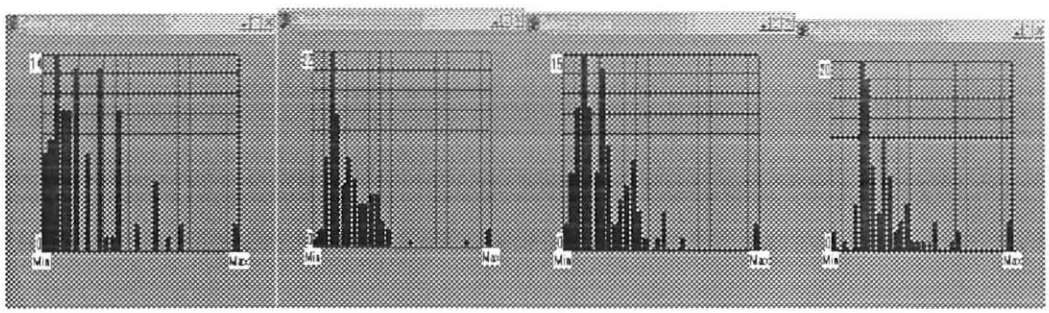


Amount Granted | Net Working Capital | Security Value | Security/Loan Ratio

The existence of these outliers was adjusted by making use of the brain maker facility, which assigns a certain reduced value to all the extreme values. The reduced value is to be assigned by the model developer. Accordingly, the original high values for each field were assigned other reduced values as in below.

Field Name	**Original Max. Value**	**Changed Max. Value**
Granted	870,000	500,000
Net Working Capital	2,103,000	1,000,000
Security Value	1,278,000	1,000.000
Security/Loan Ratio	18.23	7

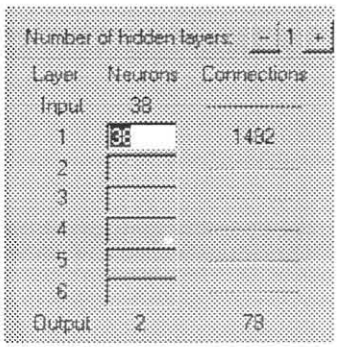
The histograms for the above four fields after adjustment is as in below.



Amount Granted Net Working Capital Security Value Security/Loan Ratio

However, the above did not result in a better performing network. The converged network from this trial tested with accuracy of 90% (9 of the 10 test cases classified correct) with testing tolerance of 0.4, 0.3 and 0.2. And with testing tolerance of 0.1 the network tested with accuracy of 70% (7 of the 10 test cases classified correct).

As discussed above, many models were developed using the data of Awassa Area Bank and the network which had the best results was the one which was trained with twelve inputs namely 'Loan No.', 'Month', 'Loan/Time ratio', 'Amount Granted', 'Net Working Capital', 'A. Debt/Asset Ratio', 'A. Current Ratio', 'Security Value', 'Security/Loan Ratio', 'Trade Sector', 'Years in Business' and 'Per. of Prior Loans'. This network was trained with the default parameters given in Table 1 and had the following network size.



The accuracy rate of this network was 100% (all 10 test cases classified correct) at testing tolerance of 0.4 and accuracy of 90% (9 of the 10 test cases classified correct) at testing tolerances of 0.3, 0.2 and 0.1

The encouraging results obtained from Awassa Area Bank implied that the above assumptions i.e. building of a model for each area bank independently might be proper. The next good candidate was Jimma Area Bank since, as in Awassa, the area bank was managed by one manager for most of its operation life. From the records of borrowers collected from this area bank 93% were processed by this manager.

The collected records had 107 records constituting of 68 regular loans and 39 irregular loans. With 64% of the data classified in one category i.e. regular it seemed that it would be difficult for

the network to learn the pattern of the rarer cases i.e. the irregular loans. As already discussed above Lawrence (1998) put that it is important for the output column to be well distributed.

The approach would have been to reduce the records of the regular outputs but this was not feasible since the numbers of records were already very few. Therefore, the researcher decided on experimenting on the data as it is.

The first experiment was conducted by using the 11 inputs that were determined to be important from the results obtained from Awassa Area Bank. These are 'Loan No.', 'Month', 'Amount Granted', 'Net Working Capital', 'A. Debt/Asset Ratio', 'A. Current Ratio', 'Security Value', 'Security/Loan Ratio', 'Trade Sector', 'Years in Business' and 'Per. of Prior Loans'. The model thus developed had accuracy rate of 73% (with 8 of the 11 test cases classified correct) at a testing tolerance of 0.4 and 64% (with 7 of the 11 test cases classified correct) at testing tolerance of 0.3 and 0.2 and at a testing tolerance of 0.1 the network tested with accuracy of 54% (6 of the 11 test cases classified correct). A variety of different combinations of variables were also tried but improvement was not obtained.

The researcher then tried to determine if there were variables that are typically important for this area bank. Discussion with the bank experts suggested that probably one variable could be tried. This field was labeled 'How Early' and it signified how early before its due date a prior loan was settled. The experts seem to notice that there were instances where a particular borrower obtains a loan amount and pay it very early and then considering the good repayment habit of the borrower the bank would release a larger loan amount. And in such cases it has been observed that some borrowers default after securing the larger loan amount. Thus, the new field 'How Early' was added to the previous eleven fields and a network was trained taking the twelve fields as input.

Testing Tolerance → / **Excluded Variable** ↓	0.4	0.3	0.2	0.1
Loan No.	64%	64%	64%	64%
Month	54%	54%	54%	54%
Amount Granted	64%	64%	64%	64%
Net Working Capital	73%	73%	73%	73%
A. Debt/ Asset Ratio	73%	73%	73%	73%
A. Current Ratio	73%	73%	73%	73%
Sec. Value	64%	64%	64%	54%
Security/Loan Ratio	64%	64%	64%	54%
Trade Sector	45%	45%	45%	36%
Years in Business	73%	73%	73%	64%
Per. of Prior Loans	64%	64%	64%	64%
Yearly Payment	73%	64%	64%	64%
How Early	54%	54%	54%	54%

From the results, it is observed that exclusion of the variables 'Loan No.', 'Month', 'Amount Granted', Security Value', 'Security/Loan Ratio', 'Trade Sector', 'Per. of Prior Loans' and 'How Early' resulted in poor performing networks. This result is similar to the results obtained from Awassa Area Bank. Except for security value all the variables were free from estimator's bias.

And better performing networks resulted when 'Net Working Capital', 'A. Debt/Asset Ratio', and 'A. Current Ratio' were excluded. This also supports the above assumptions that financial values used for this research work are not good candidates for model building due to estimator's bias.

Upon analysis of the test results from the above networks it was seen that three facts were constantly being misclassified even at a loose testing tolerance of 0.4. The confusion matrix is as shown below. 'For classification problems, a confusion matrix is a very useful tool for understanding results. A confusion matrix shows the counts of the actual versus predicted class values. It not only shows how well the model predicts but also presents the details necessary to pinpoint where things may have gone wrong' (Edelstein, 1998).

		Actual		
		Regular	Irregular	**Total**
Predicted	Regular	5	3	**8**
	Irregular	0	3	**3**
	Total	**5**	**6**	**11**

The above confusion matrix shows that all the three misclassified loans are irregular loans that were classified as regular. This seems to be the problem of having few cases for the irregular loans making it difficult for the network to detect the pattern of irregular loans. As already stated above it is very important for the output column to be well distributed (Lawrence 1998). This assumption was checked with another area bank namely Dilla Area Bank.

The collected data for Dilla Area Bank had 126 records of which 77(62%) and 49(39%) were classified as regular and irregular loans respectively. Depending on the results from Awassa and Jimma Area Banks, 12 variables were chosen to be used for the first experiment. These were 'Loan No.', 'Month', 'Granted', 'Net Working Capital', 'A. Debt/Asset Ratio', 'A. Current Ratio', 'Security Value' and 'Security/Loan Ratio', 'Trade Sector', 'Years in Business', 'Per. of Prior Loans' and 'How Early'. This experiment resulted in precision rate of 62% (8 of the 13 test

cases were classified correct) and at a precision rate of 0.1 the network resulted in precision rate of 54% (6 of the 13 test facts were classified correct). Then the researcher started experimenting with different combinations of variables and the best result that was obtained from the experiments was a network that had precision rate of 69% (9 of the 13 test facts were classified correct) even at a testing tolerance of 0.1. This network was trained with the variables 'Loan No.', 'Amount Granted', 'Asset', 'Net Working Capital', A. Debt/Asset Ratio', 'A. Current Ratio', 'Security Value', 'Security/Loan Ratio', 'Years in Business', 'Month', 'Trade Sector ' and 'Past Loan'.

Compared to the results of Awassa, the resulting networks from Dilla Area Bank's data were less performing. Numerous trials were further made to improve the performance of the network. For instance, the records of Dilla were shuffled to see if better performing networks could result. First the rows were shuffled three times and the 12 variables identified to have resulted in a 69% (9 of the 13 test facts were classified correct) precision rate at testing tolerance of 0.1 were used for the model building. This experiment resulted in a network that had precision rate of 77% (10 of the 13 test records were classified correct) at 0.4 and 0.3 testing tolerances and 69% (9 of the 13 test facts were classified correct) precision rate at 0.2 and 0.1 testing tolerances.

Another network was also developed by shuffling the records seven times and using the same variables as in the above network. This network had precision rate of 69% (9 of the 13 test facts were classified correct) with 0.4, 0.3, 0.2 and 0.1 testing tolerances.

The last trial made for Dilla's records was to see if there were outliers in the data which make distinguishing of values in the middle ranges difficult. The histogram showed that there were outliers for many of the variables especially on the higher side. Therefore, six records with asset

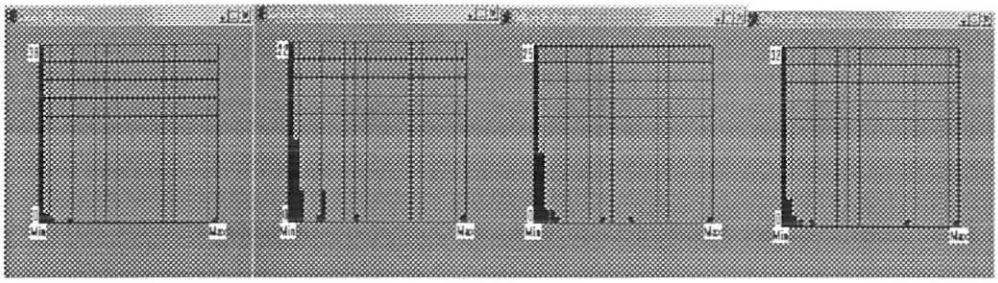
value of more than one million were removed from the file reducing the number of records to 119. This was seen to greatly improve the existence of outliers. The resulting records were then trained by using the 11 variables identified above. This network had precision rate of 67% (with 8 of the 12 test facts classified correct) at testing tolerance of 0.4 and 0.1. Shuffling of the rows was then experimented but resulted in a similar result.

Though, the results from Dilla were not discouraging when compared to Awassa performance of the networks was not very satisfactory.

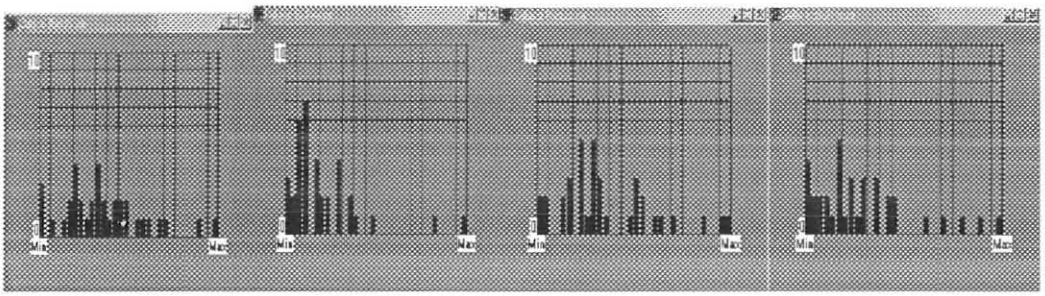
The results from Jimma and Dilla Area Banks signified that equal distribution of the outcomes i.e. regular and irregular is important for model building purposes. Most of the collected records from the Area Banks had more regular outcomes than the irregular. But the collection from Awassa and Bahr Dar Area Banks were the exceptions. Awassa Area Bank's records have already been experimented with and had resulted in a network with accuracy rate of 100%. And the collected records for the next candidate i.e. Bahr Dar Area Bank were few of not more than 43. Nevertheless, the researcher decided on experimenting on these few records.

During collection of data the researcher had noticed that the great majority of Bahr Dar borrowers' had small businesses whose total worth was mostly less than Br. one million. But there were very few exceptions from these. For instance, the minimum asset value was Br. 61,000 while the highest was Br. 38.14 million. As discussed above having such kind of data makes training difficult. This is because the scaling down of data within the range of 0 to 1 would make distinguishing of values in the middle ranges difficult. And it has also been discussed that removing the outliers for one financial value would remove the outliers for the other values i.e. a borrower who has high business worth is most likely to have highly valued security value, loan

amount etc. Therefore, the researcher removed all records, which had total asset value greater than Br. One million. And upon the exclusion of the outliers the resulting table had 37 rows. The histogram for four fields before and after removing of the outliers clearly show the extent of outlier values.



Asset Amount Granted Net Working Capital Security Value



Asset Amount Granted Net Working Capital Security Value

The other consideration was regarding how many variables to use for the training. Bigus (1996) write that the number of records should increase with increase in number of connections. Therefore the limited number of records that were available for Bahr Dar Area Bank necessitated a significant reduction of the independent variables to be used since the number of connections would increase with increase in the number of the independent variables. From the results of experiments with other Area Banks some of the variables that had been commonly identified as more important than others were selected and experimented with. In addition, a new variable that has not so far been used was included. This variable signified the year that the loan was availed. This was considered to be added so that to reduce the estimator's bias, which was expected to be

high since different managers had run the area bank at different years. And after few trials it was possible to develop a model (network) with 100% accuracy i.e. all the four test cases were classified correct. This network was trained with nine variables namely 'Loan No.', 'Amount Granted', 'A. Debt/Asset Ratio', 'A. Current Ratio', 'Security/Loan Ratio', 'Year', 'Month', 'Trade Sector' and 'Past Loan. And the network was trained with the default parameters given in Table 1.

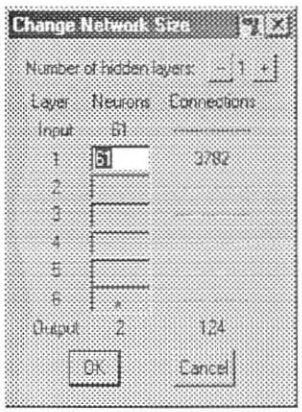
Here, it would have been interesting to develop a model with data from the area banks that have not been included in the sample by collecting proportionate data from both regular and irregular cases. But the limited time that was left for the research work did not allow another collection and preparation of data.

The last experiment to develop a model was carried out with combined records of three area banks. Though from the above it seemed that a model for a particular area bank is preferably developed independently still there was a reason that made development of a model for different area banks together more preferable. Data mining requires large collection of data and the possibility of having all the records for the different Area Banks together would significantly increase the records that would be available for model building.

The experiment to develop the model for the records of the three area banks namely Awassa, Bahr Dar, and Dilla Area Banks was started by consolidating their records. The three area banks were chosen considering the fact that from the four area banks that has been experimented with the best models were developed from the records of these three area banks.

The combined data of the three area banks had 271 records. To handle the problem of outliers first records with asset value of more than Br. one million were excluded from the file. This resulted in a file that had 248 records of which 137 (56%) were regular and 111 (44%) irregular loans. The percentages showed that there are fair representations of both outcomes.

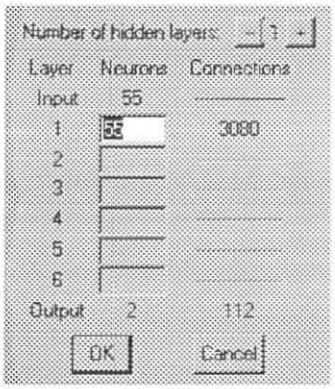
As discussed in another part the software providers suggest that there is subtle relationship between number of records and the number of input neurons i.e. as the number of input neurons increases it is suggested that number of records should also increase. Thus, for the data from the three area banks it was possible to use many input since the records were many. The first network that was experimented with had 16 variables. The network size for this experiment was:



One of the models saved before the convergence of the network resulted in precision rate of 76% (19 of the 25 test cases were classified correct) at a testing tolerance of 0.4. This result was encouraging therefore the researcher started experimenting with different combinations of variables.

After numerous trials the best results that was obtained was from a network that was trained with the following inputs: 'Loan No.', 'Yearly Payment', 'Amount Granted', 'Total Asset', 'Net Working Capital', 'A. Current Ratio', 'Security Value', 'Security/Loan Ratio', 'Years in

Business', 'Area', 'Month', Trade Sector' and 'Per. of Prior Loan'. This network was trained with the default parameters given in Table 1 and had the following network size.



Several networks were saved before the network converged. The precision rate of the different networks is summarized below.

Network Number	**Testing Tolerance**			
	0.4	0.3	0.2	0.1
1	88%	80%	76%	56%
2	84%	84%	76%	68%
3	84%	80%	80%	60%
4	80%	76%	68%	68%
5	84%	68%	68%	68%
6	80%	76%	68%	68%
7 (The converged network)	72%	72%	76%	72%

From the different networks (models) that were saved the best one was the third network where there was precision rate of 84% (21 of the 25 test facts were classified correct) at testing tolerance

of 0.4 and precision rate of 80% (20 of the 25 test cases were classified correct) at testing tolerance of 0.3 and 0.2.

Also, the first network had good precision i.e. 88% (22 of the 25 test cases were classified correct) at testing tolerance of 0.4 and 80% (20 of the 25 test cases were classified correct) at testing tolerance of 0.3. However, for all the networks saved it is seen that there is a significant reduction in precision rate as testing tolerance become very tight.

4.4 Summary of Results

In this research work numerous models were developed for an area bank independently and for the combined data of different area banks. The performance of some of the models were encouraging indicating the possible application of data mining technology in supporting credit decision making at Dashen Bank. Some of these models are put below. All the models discussed below are trained with the default parameters i.e. training tolerance of 0.1, learning rate of 1 and smoothing factor of 0.9. Though a number of other trials were made by varying the values of the parameters, for all experiments the best one was achieved when using the default parameters.

The first good result came from Awassa Area Bank which had 102 records. The best model that resulted from Awassa Area Bank's data was trained by using twelve inputs namely 'Loan No.', 'Month', 'Loan/Time Ratio', 'Amount Granted', 'Net Working Capital', 'A. Debt/Asset Ratio', 'A. Current Ratio', 'Security Value', 'Security/Loan Ratio', 'Trade Sector', 'Years in Business' and 'Per. of Prior Loans'.

The accuracy rate of this network was 100% (all 10 test cases were classified correct) at testing tolerance of 0.4 and accuracy of 90% (9 of the 10 test cases were classified correct) at testing tolerances of 0.3, 0.2 and 0.1.

Satisfactory results were also obtained from Dilla and Jimma Area Bank but compared to Awassa Area Bank results were not very good. Jimma Area Bank had 107 records. The best model developed for the Area Bank was one developed with 12 variables namely, 'Loan No.', 'Month', 'Amount Granted', 'Net Working Capital', 'A. Debt/Asset Ratio', 'A. Current Ratio', 'Security Value', 'Security/Loan Ratio', 'Trade Sector', 'Years in Business', 'Per. of Prior Loans', 'Yearly Payment' and 'How Early'. This network had precision rate of 73% (8 of the 11 test facts were classified correct) at testing tolerances of 0.4, 0.3 and 0.2. And at testing tolerance of 0.1 the network tested with accuracy rate of 64% (7 of the 11 test facts were classified correct.)

And for Dilla Area Bank the best model was obtained in an experiment where the rows were shuffled three times. This network was trained with the variables 'Loan No.', 'Amount Granted', 'Asset', 'Net Working Capital', 'A. Debt/Asset Ratio', 'A. Current Ratio', 'Security Value', 'Security/Loan Ratio', 'Years in Business', 'Month', 'Trade Sector' and 'Per. of Past Loan'. The precision for this network was 77% (10 of the 13 test facts were classified correct) at testing tolerance of 0.4 and 0.3. And at testing tolerance of 0.2 and 0.1 the network had precision of 69% (9 of the 13 test facts were classified correct).

The less performance of the networks from Jimma and Dilla Area Banks were considered to be attributed to the fact that the files of the two Area Banks had uneven distribution of the possible outcomes i.e. regular or irregular loans.

Thus, a file with proportional distribution of the possible outcomes was chosen and experimented with. This file held the records of Bahr Dar Area Bank. The few records that were available i.e. 43 were experimented with. And it was possible to develop a model that classified all the four test facts correctly at testing tolerance of 0.4. And at testing tolerances of 0.3, 0.2 and 0.1 three of the four test facts were classified correct. This network was trained with the variables 'Loan No.', 'Amount Granted', 'A. Debt/Asset Ratio', 'A. Current Ratio', 'Security/Loan Ratio', 'Year', 'Month', 'Trade Sector', and 'Per. of Past Loan'.

The last experiment was carried out for the combined data of three area banks namely Awassa, Dilla and Bahr Dar Area Banks. The results from numerous experiments resulted in a number of good models. Two of these were trained using the variables 'Loan No.', 'Yearly Payment', 'Amount Granted', 'Total Asset', 'Net Working Capital', 'A. Current Ratio', 'Security Value', 'Security/Loan Ratio', 'Years in Business', 'Area', 'Month', Trade Sector' and 'Per. of Prior Loan'. The two networks were saved at different intervals.

The first network had precision rate of 84% (21 of the 25 test facts were classified correct) at testing tolerance of 0.4 and precision rate of 80% (20 of the 25 test facts were classified correct) at testing tolerance of 0.3 and 0.2. The second network had precision rate of 88% (22 of the 25 test facts were classified correct) at testing tolerance of 0.4 and 80% (20 of the 25 test facts were classified correct) at testing tolerance of 0.3.

Overall the model building process demonstrated that data mining using neural networks could be considered for Dashen Bank in supporting loan disbursement activity. Especially if it was possible to obtain detailed and reliable financial and other information from borrowers' records it is believed that better networks (models) could be developed. Also, it is important to have a data

warehouse where a centralized data is available electronically. Otherwise collecting data manually, in a model building process where data needs to be updated at an interval, would not be efficient.

The model building process revealed information on the importance of different variables that were not at first suspected to be very important. It was observed that non-financial values such as 'Month', 'Trade Sector', and 'and 'Per. of Prior Loan' were important variables than the financial values. This was explained by the fact that commercial banks in Ethiopia use financial values, which are not measured by uniform measurement.

In developing models it was also observed that experts opinion and suggestions were of paramount importance and thus is not a task that should be left out only to IT professionals. Al-Attar (1999) discusses that users who understand the business must guide data mining tools. Therefore, assistance of area experts should be duly sought in any data mining effort.

Chapter 5

Conclusion and Recommendations

5.1 Conclusion

Today, businesses have too much data and experiences being generated everyday but the problem is in comprehending these data and experiences in order to put them into use. Data mining is one technique that is being applied in bridging this gap between availability of large volume of data and the limitation of the analyzing tools.

Data mining technique has already been tested and it's results appreciated in different sectors. One of these is the banking sector where data mining application has been used in assessing credit worthiness of customers, detecting credit card frauds, direct marketing and analyzing profitability of credit card users etc.

The objective of this research undertaking was to assess the possible application of data mining technology in the Ethiopian banking context, and particularly at Dashen Bank, by developing a model that could help predict whether a potential borrower would default or not. Such a model could then be applied in assisting the credit decision process.

The methodology employed followed three basic steps; data collection, data preparation and model building/testing. These steps were not strictly followed sequentially but there were frequent instances where there was a need to go back and forth between the different steps.

Numerous trials had to be made in order to come up with a model that made good prediction. The best models that were developed were those developed for Awassa, Bahr Dar and Dilla Area Banks. In addition, a good model was developed using consolidated data of three area banks namely Awassa, Bahr Dar and Dilla Area Banks. This model had accuracy rate of 88%.

During the course of the model building there were many important findings that were observed. The research revealed that some variables were consistently observed to be important variables for model building. These variables include 'Amount Granted', 'Trade Sector', 'Month loan was granted', 'Performance of Past Loan, 'Anticipated Debt/Asset ratio', 'Anticipated Current Ratio' and the 'Number of Loans the borrower has settled in the past'. During the research work, the researcher has come to observe that some of these variables are not considered as very important ones by experts of the bank. The researcher, therefore, hopes that the findings of this research work will in the future help to give due emphasis for variables that were hitherto given less attention.

The nature of the above listed variables also indicated another insight i.e. variables, which were purely determined by estimator's judgment, are not good candidates for model building. And this indicates the importance of having uniform measurement for the financial values.

The encouraging results obtained indicate that data mining application is really a technology that should be considered in assisting credit decision at Dashen Bank. However, there are issues that should be considered before the technology can be efficient and feasible in the Ethiopian context. These issues are discussed in the next section.

It is important to note that application of data mining technology should not be understood as removing the need for experts who have an experiential knowledge in understanding the intricate details of the business. Without such expertise, it is even impossible to develop and update the kind of model under consideration. In addition, it is also up to the experts to assess the results obtained by the work of such a model and determine its feasibility. Small (1997) state that 'no analysis technique can replace experience and knowledge of the business and its markets.'

5.2 Recommendations

On the basis of the findings of this research work, the researcher would like to make the following set of recommendations in relation to the possible application of data mining technology in supporting credit disbursement activities at Dashen bank. Although the current research work was only in one commercial bank, namely Dashen bank, it is the considered opinion of the researcher that the basic findings of the research work and the attendant conclusions are fairly applicable to other commercial banks as well.

The researcher would also like to note that this research, as an academic exercise, should only be considered as a preliminary effort to assess the applicability of data mining technology in Ethiopian commercial banks. Accordingly, the findings of this research undertaking can fairly be considered as a contribution towards a more in depth and comprehensive study in the area. The researcher hopes that the findings of this research work will help to initiate more research in the area and some of the following recommendations relate to points for consideration in a similar future research work.

- ***Making the required data available in a computerized form***

Data mining, like any other techniques, would need to be efficient if it is to be put to use and a data mining technique does not appear to be feasible where data is available only in manual formats. Therefore, for an efficient application of data mining technology pertinent data need to be available in a computerized format. The importance of electronic data becomes vital since data mining technique requires frequent updating of data. Hence, Dashen Bank should address the issue of having an information system where relevant data can be easily accessed electronically.

- ***Encourage keeping proper books of accounts***

The issue of keeping proper books of accounts among Ethiopian business people is an important concern that needs to be addressed. The business community by and large, does not maintain proper books of accounts. The lack of proper books of accounts has resulted in many obstacles such as difficulty for the businessperson to assess if his/her business is progressing well or not. And there is a general problem for government to assess tax payments. Lack of proper books of accounts remains to be one of the problems in credit appraisal in the banking sector, which in turn was a problem that affected the present research work. Data mining determines relationship based on past data and as a result the lack of uniform measurement obviously affects the model that is developed. Hence, it is really important to address such issues if the new state of the art technologies that are coming up in the market is to be utilized in the Ethiopian context.

- ***Devote sufficient time for building and training an appropriate model***

This research undertaking has yielded an encouraging result with a rather short period of time. The researcher has noted that the accuracy rate tends to increase to a desirable level with more and more trials of different combinations of the variables. It is, therefore, recommended that sufficient time be taken to build and train an appropriate model with a desirable level of accuracy

in prediction. While the results obtained from this research work is encouraging, it is believed that with more time for an appropriate model building and with time given to evaluate the results of the model, the more chance that a better model would be developed. Two Crows Corporation (1999) also state that 'the more the model builder can 'play' with the data, build models, evaluate results, and work with the data some more (in a given unit of time), the better the resulting model will be.'

- ***Consider including as many relevant variables as possible***

It has been observed that some variables that were considered as marginal factors by the researcher are found in practice to be rather important variables in determining the credit worthiness of borrowers. And the wisdom in choosing variables that have a reasonable degree of relevance for the appraisal process is available among the bank experts and their assistance should be duly sought. In addition, it may be important to include variables that are available outside the organization. For instance, during the course of the research work it has been discovered that past performance of borrowers is important in determining their credit worthiness. So perhaps it could be beneficial to take past performance of borrowers with other commercial banks as a variable. Other variables that may be available outside the organization could be economic indicators such as inflation rate, per capita income etc.

- ***Introduce more detailed classification of borrower's category***

Detailed categorical classification of borrowers would provide better information in assisting the credit decision. For instance, if there was a classification for borrower who repaid irregularly but on time and another classification for borrowers who paid after the loan due date, it is possible to get better information for the decision process. The loans with arrears may indicate that such

borrowers should probably be given different repayment schedules and that they are not very high risks but for those who pay after due date special caution has to be taken. It is, therefore, recommended that for further research work related to data mining application in credit risk assessment more detailed classification be introduced to categorize borrowers in to different categories in accordance with their specific characteristics.

- ***Studying the applicability of other data mining techniques***

The neural network technique considered here is only one of the different data mining techniques. The experiment with this technique has shown an encouraging result. But writers discuss that there are other techniques such as decision tree, which have also been beneficial in credit risk assessments. Hence, it is recommended that other techniques should also be tested to see if they could be more applicable than neural network technique.

Reference:

Al-Attar, Akeel. "Data Mining –Beyond Algorithms." Attar Software Limited. (1999): Online. Internet. Available URL: http://www.attar.com/tutor/mining.htm

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

Benning, Stacey; Michelle Denning, Cooch Janquint and Paul Russel. "Data Mining." (1999) University of Iowa: College of Business. Online. Internet. Available URL: http://www.biz.uiowa.edu/class/6k180_park/Student-Reports/sbenning/

Beryy, Micheal J.A. and Gordon Linoff. 'Data Mining Techniques, For Marketing Sales and Customer Support.' New York: John Wiley and Sons Inc, 1997.

Bigus, Joseph P. 'Data Mining with Neural Networks, Solving Business Problems from Application Development to Decision Support.' New York: McGraw Hill, 1996.

Business Development Department of Dashen Bank. "The Evolution of Banking in Ethiopia." (2001) Dashen Bank, Complimentary Issue, 5th Year Anniversary.

California Scientific Software, 1999, http://www.calsci.com/

- - -. 'BrainMaker, User's Guide and Reference Manual.' 1998

Connolly, Thomas; Carolyn Begg and Anne Strachan (1999). 'Database Systems, A Practical Approach to Design Implementation and Management.' New Jersey: Addison-Wesley.

CorMac Technologies Inc. 'What is a Neural Network' Online. Internet Available URL: http://www.cormactech.com/neunet/helpfile/4100what.htm

Dashen Bank Annual Report 1999/2000

"Data Mining", Online. Internet Available URL: http://www.rpi.edu/~arunmk/dm1.html

Edelstein, Herb. "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. (Summer 1998): Online. Internet. Available URL: http://www.db2mag.com/98smEdel.htm

- - -. " Data Mining: *Exploiting the Hidden Trends in Your Data.*" Database Programming & Design Magazine. (Spring 1997): Online. Internet. Available URL: http://grwy.online.ha.cn/sweetheart/9701edel.htm

- - -. "Mining for Gold." Information Week. (April 21, 1997): Online. Internet. Available URL: http://www.twocrows.com/iwk9704.htm

Fabris, Peter. "Data Mining" (1998) CIO Magazine, "Advanced Navigation." (1998): Online. Internet. Available URL: http://www.cio.com/archive/051598_mining_content.htmlhttp://www.cio.com/archive/051598_mining.html

Fraser, Christopher M. "Neural Networks: Literature Review From a Statistical Perspective." Hayward Statistics. California State University, Hayward. (Spring 2000): Online. Internet. Available URL: http://www.telecom.csuhayward.edu/~stat/Neural/CFProjNN.htm

Frohlich, Jochen. "Neural Net Overview." (1999): Online. Internet. Available URL: http://rfhs8012.fh-regensburg.de/~saj39122/jfroehl/diplom/e-1-text.html

Graettinger, Tim. 'Digging Up $$$ with Data Mining –An Executive Guide' (1999) Discovery Corps, Inc. Online. Internet. Available URL: http://www.tdan.com/i010ht01.htm

Grove, Tom D. "Neural Nets – part I: *Why are people more intelligent than machines*?" (2001): Online. Internet. Available URL: http://umtii.fme.vutbr.cz/MECH/NN/tomgr1.html

Higgins, Robert C. "Analysis for Financial Management." Irwin/McGraw-Hill, 1997.

Kestelyn, Justin. "Extracting Fact from Fiction: Does Data Mining Fit in your Enterprise?" Database Programming & Design Magazine. (Spring 1997): Online. Internet. Available URL:

Knowledge Technology Inc. "PC AI – Glossary of Terms." (2001): Online. Internet. Available URL: http://www.primenet.com/pcai/New_Home_Page/glossary/pcai_glossary.html

Lawrence, Jeannette (1994). 'Introduction to Neural Networks, Design, Theory, and Applications.' Nevada City: California Scientific Software Press, 1994.

Liao, Weidong. "Data Mining on the Internet." (May 1999): Online. Internet. Available URL: http://trident.mcs.kent.edu/~javed/DL/surveys/IAD99s-datamining/

Moxon, Bruce. "Defining Data Mining." Miller Freeman, Inc. (August 1996): Online. Internet. Available URL: http://www.dbmsmag.com/9608d53.html

National Bank of Ethiopia Quarterly Bulletin, Volume 15, No. 3

New Wave Intelligent Business Systems, NIBS Inc. "Neural Network Computing." (2001): Online. Internet. Available URL: http://web.singnet.com.sg/~midaz/Intronn.htm

Oracle Corporation. "Why Mine Data? An Executive Guide: *Using Business Intelligence to Attract and Retain Customers.*" An Oracle White Paper. (June 1999): Online. Internet. Available URL: www.oracle.com

Piatetsky-Shapiro, Gregory. "Knowledge Discovery in Real Databases, A Report on the IJCAI-89 Workshop." GTE Laboratories Incorporated. Online. Internet. Available URL:

PMSI, "The Saxon 4.3 Software Suite: a Simple yet Complete Neural Network Toolbox." (2001) : Online: Internet. Available URL: http://www.pmsi.fr/sxccomma.htm

Pudi, Vikram. "Neural Networks." Indian Institute of Science. (2001):Online. Internet. Available URL: http://dsl.serc.iisc.ernet.in/~vikram/nn_intro.html

Riggen, Russ and Budansky, Vladimir. "Retaining Profitable Customers Through Data Mining." Profitability Bulletin. Publication of Ernst and Young's Financial Services Consulting. (1997):Online. Internet. Available URL: http://www.biznetwork.com/comm/fs/fsne3e.htm

Saarenvirta, Gary. "Operation Data Mining." Database Programming & Design Magazine. (Summer 2001): Online. Internet. Available URL: http://www.db2mag.com/db_area/archives/2001/q2/saarenvirta.shtml

Shiferaw Bekele. "The Evolution of Banking in Ethiopia." (2001) Dashen Bank, Complimentary Issue, 5th Year Anniversary.

Siganos, Dimitrois. "Why Neural Networks." (1997): Online. Internet. Available URL: http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol1/ds12/article1.html

Skalak, David. "Data Mining Blunders Exposed." Database Programming & Design_Magazine. (Summer 2001): Online. Internet. Available URL: http://www.db2mag.com/db_area/archives/2001/q2/miner.shtml

Small, Robert D. "Debunking Data Mining Myths." Information Week. (January 20, 1997): Online. Internet. Available URL: http://www.twocrows.com/iwk9701.htm

Small, Robert D., and Herbert A. Edelstein. "Scalable Data Mining." Two Crows Corp. (1997): Online. Internet. Available URL: http://www.twocrows.com/whitep.htm

Smith, Leslie. "An Introduction to Neural Networks." Center for Cognitive and Computational Neuroscience. University of Stirling. (1996): Online. Internet. Available URL: http://www.cs.stir.ac.uk/~lss/NNIntro/InvSlides.html

Stergiou, Chris. "What is a Neural Network?" (2001): Online. Internet. Available URL: http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol1/cs11/article1.html

Stergiou, Christos, and Dimitros Siganos. "Neural Networks." Online. Internet. Available URL: http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

Sun Microsystems Computer Company. "Scalable Data Mining with Sun™ Ultra™ Enterprise Servers." (April 1997): Online. Internet. Available URL: http://www.gca.net/solutions/whitepapers/sun/sun_data_mining.html

Tveter, Donald R. "The Backprop Algorithm." (Year) Online. Internet. Available URL: http://www.dontveter.com/bpr/public2.html

Two Crows Corporation. "Introduction to Data Mining and Knowledge Discovery." Two Crows. ISBN: 1-892095-02-5. (1999): Online. Internet. Available URL: http://www.twocrows.com/

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

Z Solutions. 'Neural Networks and Data Mining.' (1999) Online. Internet. Available URL: http://www.zsolutions.com/sowhy.htm

Z Solutions. "Managers Guide to Neural Networks: *Extracting Knowledge from Information.*" Online. Internet. Available URL: http://www.zsolutions.com/amanager.htm

Bibliography:

Bluce, James L., Charles L. Wilson, and Omid Omidvar. "Neural Network Classification and Dynamical Systems." (1996): Online. Internet. Available URL: http://math.nist.gov/mcsd/Reports/96/yearly/node31.html

Chung, Michael H. and Paul Gray. 'Current Issues in Data Mining' Online. Internet. Available URL: http://www.csulb.edu/~imats/datamining.htm

Edelstein, Herb. "Predicting Customer Behavior with Neural Nets." Database Programming & Design Magazine. (Spring 1997): Online. Internet. Available URL: http://grwy.online.ha.cn/sweetheart/9701eds2.htm

Gobena Michael. "Flight Revenue Information Support System for Ethiopian Airlines.' (2000) A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree of M. Sc. I. S. Addis Ababa University, Addis Ababa.

Gray, Michael, Craig Coen, and Kevin Frost. "Mutual Fund Net Asset Value Forecasting Using Neural Networks." Online. Internet. Available URL: http://library.cmsu.edu/cis3630/misweb.htm

Haykin, Simon. 'Neural Networks, A Comprehensive Foundation' New Jersey: Prentice Hall Inc, 1999.

Hermiz, Keith., and Stefanos Manganaris. "Beyond Beer and Diapers." Database Programming & Design Magazine. (Winter 1999): Online. Internet. Available URL: http://www.db2mag.com/winter99/miner.shtml

Makulowich, John. "Data Mining: Development Gain Attention." Miller Freeman, Inc. (August 1996): Online. Internet. Available URL: http://www.kdnuggets.com/press/wt97/

Oracle Corporation. "Data Selection Guidelines for Data Mining Evaluation." An Oracle White Paper. (December 2000): Online. Internet. Available URL: www.oracle.com

- - -. "Oracle's Data Mining Solutions." An Oracle White Paper. (December 2000): Online. Internet. Available URL: www.oracle.com

Palace, Bill. "Data Mining: What is Data Mining" 1996 Online. Internet. Available URL: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

Stergiou, Christos. "Neural Networks, the Human Brain and Learning." (2001): Online. Internet. Available URL: http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol2/cs11/article2.html

Whiting, Rick, and Jeff Sweat. "Profitable Customers, Business are using IT to identify high-yield clients and formulating new strategies for dealing with those that aren't." Tech search: Information Week. Issue: 727. (March 29, 1999): Online. Internet. Available URL: http://www.informationweek.com/727/customer.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: assets 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 stockholders' 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 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

C:\LAF\LAF 2000\LAF FORM.doc

DASHEN BANK

LOAN APPROVAL FORM

Area Bank *Dashen Main Area Bank* LAF No *dMB/LAF/ /2000* Date _____

1. Name of Applicant _____
2. Facility Requested for T/L Birr _____ Mdse. Loan Birr _____
   O/D Birr _____ L/C _____ @ _____
   Others Birr _____ Date of Application _____
3. Purpose (briefly describe) _____
4. Major line of business _____
   Address Town _____ Wer. ____ K. ____ H. No. ____ @ _____ [x-] ____
   Distance from the Area Bank _____ K. M. Date of establishment _____
   Other line of business _____ Trade License renewed for _____
5. Security offered Type a. Bldg. ☐ b. Vehicle ☐ c. P.Gtee ☐ (attach FCR of guarantor)
   d. Clean ☐ e. Fin. Gtee ☐ f. combination of ____ & ____
   Property estimated value Birr _____ Date of estimation _____
   Estimated by _____
   Property owned by applicant ☐ third party ☐ applicant and third party ☐
   Mode of security holding 1st degree ☐ 2nd degree ☐
   Specify 1st degree holder and magnitude of credit availed _____
6. Comparative financial statements as checked by the Area Bank Manager to be filled in the absence of audited financial statements.
   6.1 Balance sheet _____

	Asset	Current year as at	Previous year as at
A.	Cash on hand & Bank		
B.	Receivable & Prepayments		
C.	Stock		
D.	Total current Assets (A+B+C)		
E.	Fixed Asset		
F.	+ Total. Assets (D+E)		

Liability		Current year as at	Previous year as at
G.	Payable including Sundries Associated Enterprise		
H.	Taxes		
I.	Bank Loans		
J.	* Total Current Liability (G+H+I)		
K.	Long Term Debt		
L.	Total Liability (J+K)		
M.	Capital Capital and retained Earnings		
N.	Total Liability & Capital (L+M)		
O.	Net Working Capital (D-J)		
P.	Capital and Reserve (F-L)		

6.2 Income Statement

Birr ('000)

	Current year	Previous year 19
Sales (Income)		
Add Other income		
Less Cost of Sales		
Gross Margin		
Less expenses		
Profit Before Tax		
Less Profit Tax		
Net Income (Annual)		

6.3 Give reasons for significant rise/fall in the financial position of the customer

_____

_____

_____

7. Composition of items in stock

_____

_____

_____

8. Description of receivable

_____

9 Description of payables & repayments

_____

10 Turnover of O/D or C/A

	Current Year		Previous Year
Birr	_____	Birr	_____

11. Foreign Banking Transitions channelled through DB

Current Year		Previous Year	
Authorised	Utilised	Authorised	Utilised

12. Existing Facility with

			Expiry Date	*Status*	*Balance*
Dashen Bank	T/L	_____	_____		_____
	O/D	_____	_____		_____
	L/C Birr	_____	_____		_____
Other Banks Specify	T/L Birr	_____	_____		_____
	O/D Birr	_____	_____		_____
	Other Birr	_____	_____		_____

13. General Information about the customer and observation of the business situation of applicant.

14. Information about past loans if any.

_____

_____

15. Recommendation/Decision of Area Bank Loan Committee

Signature

16 Decision of Head Office Loans Committee (give reasons for variation if any)

Annex 2

Monthly Return

F.CR-3

DASHEN ----------- Area Bank
Schedule of Term Loans
As at ---------------

TYPE OF LOAN....................

Name of Customer	Amount Granted	Date Granted	Extension Date	Due Date	During the Month		Amount Outstanding		Arrears	Repayment		Security	
					Amount Granted	Amount Collected	Current	Overdue		Term	Mode	Type	Value

Annex 3

ዳሽን ባንክ አ.ማ.

DASHEN BANK S. Co.

የፋይናንስ መግለጫ ቅጽ

FINANCIAL CREDIT REPORT

ቅርንጫፍ ቀን

Branch _____ 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. የ.መ.ሣ.ቁ. P.O. BOX

የሥራ BUSINESS _____ _____ _____ _____ _____ _____

የመኖሪያ ቤት RESIDENCE _____ _____ _____ _____ _____ _____

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

የማዘጋጃ ቤት MUNICIPAL _____ ሌሎች (ዝርዝር) OTHERS (SPECIFY) _____

የተብረት ሥራ ማኅበር-የመዘገበው ድርጅት፣አእማድ CO-OPERATIVE REGISTERED BY HASIDA _____ ሌሎች OTHERS _____

5. አመልካች የተፈቀደላቸው የሥራ ዓይነት(ዝርዝር) 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. የሠራተኛ 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 (ዕቅድ) PROPOSED MODE OF REPAYMENT _____

12. አመልካች ከባንኩ ጋር ያለው፣ ግንኙነት፣ በአስቀማጭነት፣ RELATION WITH BANK as DEPOSITOR _____ የቅርንጫፍ ባንክ ስምና የሂሣብ ቁጥር NAME OF BRANCH BANK A/C No. _____

በዋስትና as GUARANTOR _____ 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	ያስመዘገቡት 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 ከዕቃ ሽያጭ Receivables a) Accounts			
	ከተሰፋ ሠነዶች b) Notes			
3	በሱት ወይም በመጋዘን 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 ገዥ Payable: a) Accounts			
	የተሰፋ 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			

LINE No.	ገቢ INCOME	AMOUNT	KEY
22	ሽያጭ Sales		
23	የተሸጠው ዕቃ ዋጋ Cost of Goods Sold:		
	መነሻ ዕቃዎች Beginning inventory, as at Birr		
	ሲደመር የተገዙ ዕቃዎች Add-purchases for the period "		
	ሲቀነስ መጪረሻ የቀሩ 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 Premise Rent		
28	ስልክ፣ መብራትና ውኃ Utilities		
29	ማደሻና ጥገና Maintenance & Repair		
30	መድን Insurance		
31	የአገልግሎት ተቀናሽ Depreciation		
32	የግል (የግል መኖሪያ ቤትን ኪራይ 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 ዕዳ Guarantee Libility (in total)		

የአመልካች ፊርማ
Applicant's Garantor's Signature ----------

	የ ት ን ተ ና ሚ ዛ ና ች ANALYTICAL & COMPARATIVE RATIOS	በዚህ ዓመት This Year	ባለፈው ዓመት Last Year	
39	የተጣራ መንቀሳቀሻ ካፒታል Net Working Capital			
40	ተንቀሳቃሽ ሀብት ከጊዜያዊ ዕዳ ማነጻጸሪያ Current Ration			
41	የሽያጭና የዱቤ ማነጻጸሪያ Sales to Receivable Ratio			
42	የሽያጭና የመንቀሳቀሻ ሀብት ማነጻጸሪያ Sales to Current Asset Ratio			
43	ትርፍ ከተንቀሳቃሽ ሀብት ማነጻጸሪያ Income Before Tax as % of Current Asset			
44	ጠቅላላ ዕዳ ከተጣራ ሀብት ማነጻጸሪያ Total Debt to Worth Ratio			

አዘጋጁው
Processed by ----------

የ ሂ ሳ ብ ሠንጠረዥ ማብራሪያ
EXPLANATORY NOTES ON FINANCIAL STATEMENTS

በቅርንጫፉ ኃላፊ የሚሞላ
(to be filed by the Branch)

KEY	DESCRIPTION	AMOUNT

Annex 4: Format for Collecting Borrowers' Data with Hypothetical Records

No.	Area	Name of Borrower	Date Granted	Expiry Date	Amount Granted	Asset	Current Asset	Current Liability	Liability	Yearly Income	Security Type
1	Awassa	Abdu Kebede	2/15/1997	2/14/1998	200,000	230,000	67,000	12,000	24,000	27,000	Building
2	Awassa	Getachew Abera	5/1/1997	5/1/1998	100,000	329,611	234,555	128,000	150,000	56,432	BP
3	Bahr Dar	Debebe Yiktaw	2/27/1997	2/26/1998	50,000	345,000	123,000	-	-	12,800	Building
4	Dilla	Adey Belete	2/21/1997	8/20/1998	200,000	576,643	325,000	120,000	184,567	15,125	PG
5	Kality	ABC PLC	3/15/1997	3/14/1998	100,000	1,779,910	576,000	134,765	189,000	81,775	BV
6	Mekelle	Girma Hailu	4/18/1997	4/17/1998	80,000	151,334	58,000	-	-	23,000	Building
7	Nazareth	Azeb Teklu	4/22/1997	10/21/1998	80,000	258,375	67,800	12,000	28,756	42,567	Building

Annex 4: Format for Collecting Borrowers' Data with Hypothetical Records (Ctd)

Security Value	Sex	Trade Sector	Business Establishment Year	Term of Payment (monthly, quarterly)	No. of Prior Loans Settled	Performance of Prior Loan	Remaining loan life at the time of settlement (for prior loans)	Classification
230,000	Male	Coffee	1978	Quarterly	1	NA	0	Regular
165,000	Male	Cereals	1986	Monthly	-	NA	0	Regular
65,000	Male	Clothes	1989	Monthly	2	Dob	0	Substandard
-	Female	Hotel	1979	Monthly	3	Reg1	23%	Regular
285,000	CP	Manufacturing	1956	Monthly	2	Sub2	0	Substandard
115,000	Male	Clothes	1975	Monthly	1	Reg1	33%	Regular
115,000	Female	Electronics	1965	Quarterly	1	Reg1	14%	Substandard

Annex 5: List of the Independent Variables (Inputs) Used for Model Building along with their Descriptions

No.	Name of Variable	Description
1	Area	Name of the area where the Area Bank is located
2	Loan No.	The number of loan for the specific borrower (1st loan, 2nd loan, 3rd loan etc.)
3	Month	Month loan was granted
4	Duration	The duration of loan in number of days
5	Yearly Payment	The estimated amount to be paid in a year
6	Amount Granted	Amount granted
7	Loan/Time Ratio	Loan amount divided by the loan duration
8	Asset	Total asset of the borrower
9	Capital	Total capital of the borrower
10	Current Asset	Total current asset of the borrower
11	Current Liability	Total current liability of the borrower
12	Net Working Capital	Current asset – Current Liability
13	Liability	Total liability of the borrower
14	Debt/Asset Ratio	Liability value divided by asset value
15	A. Debt/Asset Ratio	The anticipated debt/asset ratio after considering the new loan to be granted.
16	A. Current Ratio	The anticipated current ratio after considering the new loan to be granted.

17	Security Type	Type of security (E.g. Building, vehicle, personal guarantee)
18	Security Value	Estimated value of the security
19	Security/Loan Ratio	Security value divided by amount granted
20	Sex	Sex of the borrower
21	Trade Sector	The kind of business borrower is engaged in E.g. Hotel, cereals etc.
22	Years in Business	The number of years the borrower has been in business
23	Term of Payment	Monthly, bimonthly or quarterly
24	No. of Prior Loans	The number of loans borrower has settled in the past
25	Per. of Prior Loans	Performance of past loans i.e. whether past loans were regular or not
26	How Early	Indicates whether a previous loan was settled on time or before its due date.

Annex 6: Format for the Prepared Data with Hypothetical Records

No.	Area	Name of Customer	Loan No.	Month	Date Granted	Expiry Date	Duration	Yearly Payment	Amount Granted	Loan/Time Ratio	Asset
1	Awassa	Abdu Kebede	2	Feb	2/15/1997	2/14/1998	364.00	200,549.45	200,000	549	230,000
2	Awassa	Getachew Abera	1	May	5/1/1997	5/1/1998	365.00	100,000.00	100,000	274	329,611
3	Bahr Dar	Debebe Yiktaw	3	Feb	2/27/1997	2/26/1998	364.00	50,137.36	50,000	137	345,000
4	Dilla	Adey Belete	4	Feb	2/21/1997	8/20/1998	545.00	133,944.95	200,000	367	576,643
5	Kality	ABC PLC	3	Mar	3/15/1997	3/14/1998	364.00	100,274.73	100,000	275	1,779,910
6	Mekelle	Girma Hailu	2	Apr	4/18/1997	4/17/1998	364.00	80,219.78	80,000	220	151,334
7	Nazareth	Azeb Teklu	2	Apr	4/22/1997	10/21/1998	547.00	53,382.08	80,000	146	258,375

Annex 6: Format for the Prepared Data with Hypothetical Records(Ctd)

Capital	Net Working Capital	Current Asset	Current Liability	Liability	Debt/Asset Ratio	Anticipated Liability	Anticipated Asset	A. Debt/Asset Ratio	Anticipated Current Liability	Anticipated Current Asset
206,000	55,000	67,000	12,000	24,000	0.10	224,000	430,000	0.52	212,549	267,000.00
179,611	106,555	234,555	128,000	150,000	0.46	250,000	429,611	0.58	228,000	334,555.00
345,000	123,000	123,000	-	-	-	50,000	395,000	0.13	50,137	173,000.00
392,076	205,000	325,000	120,000	184,567	0.32	384,567	776,643	0.50	253,945	525,000.00
1,590,910	441,235	576,000	134,765	189,000	0.11	289,000	1,879,910	0.15	235,040	676,000.00
151,334	58,000	58,000	-	-	-	80,000	231,334	0.35	80,220	138,000.00
229,619	55,800	67,800	12,000	28,756	0.11	108,756	338,375	0.32	65,382	147,800.00

Annex 6 (Ctd)

A. Current Ratio	Security Type	Security Value	Security/Loan Ratio	Sex	Trade Sector	Business Establishment Date	Years in Business	Term of Payment	Per. Of Prior Loans	How Early Was Prior Loan Settled	No. of Prior Loans	Classification
1.26	Building	230,000	1.15	Male	Coffee	1978	16	Quarterly	NA	TM	1	Regular
1.47	BP	165,000	1.65	Male	Cereals	1986	8	Monthly	NA	TM	-	Regular
3.45	Building	65,000	1.30	Male	Clothes	1989	5	Monthly	Dob	TM	2	Substandard
2.07	PG	-	-	Female	Hotel	1979	15	Monthly	Reg1	ER	3	Regular
2.88	BV	285,000	2.85	CP	Manufacturi	1956	38	Monthly	Sub2	TM	2	Substandard
1.72	Building	115,000	1.44	Male	Clothes	1975	19	Monthly	Reg1	VE	1	Regular
2.26	Building	115,000	1.44	Female	Electronics	1965	29	Quarterly	Reg1	TM	1	Substandard

DECLARATION

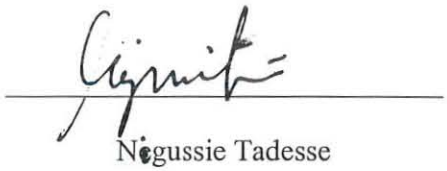
I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other University, and that all sources of material used for the thesis have been duly acknowledged.

Askale Worku

July 2001

The thesis has been submitted for examination with our approval as University Advisors

Tesfaye Birru



Negussie Tadesse

July 2001