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

POSSIBLE APPLICATION OF DATA MINING TECHNOLOGY IN SUPPORTING TERM LOAN RISK ASSESSMENT: THE CASE OF UNITED BANK S.C.

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

BY

SAMSON TADESSE

JANUARY, 2009

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

POSSIBLE APPLICATION OF DATA MINING TECHNOLOGY IN SUPPORTING TERM LOAN RISK ASSESSMENT: THE CASE OF UNITED BANK S.C.

BY

SAMSON TADESSE

Name and Signature of Members of the Examining Board

Chair person, Examining Board	Signature	Date
DR MANOJ VNV		
Advisor	Signature	Date
Chair person, Faculty	Signature	Date
Chair person, Graduate Council	Signature	Date

Dedication

This paper is dedicated to my grand mother W/ro Alganesh Hassen,to my father Ato Tadesse Redi,to my mother W/ro Lishan Dawd,to my sister Makda Tadesse, to my Brothers Girma (Degole),Elias,Tsigae and to my best friends Jecy and Smith.

Acknowledgment

First, special thanks go to GOD for all His mercy, grace, and support through my entire life. With his infinite help and grace everything is possible.

I would like to express my deepest sense of gratitude to my advisor Pro.Manoj VNV, from faculty of Technology, Department of Electrical and Computer Engineering for his guidance and encouragement through out this research. I am also grateful to all staff members of United Bank Share Company, particularly to Ato Girum Tsigaye, Ato Alelelegne Merkbu, Ato Addisu Negatu, W/ro Tsigereda Tadesse, W/ro Genet Asfaw, W/ro Alemtsehay Berhanu, W/rt Reza Demissie, Ato Tameru Agegn, Ato Tadesse Bezabhi, Ato Wallelign Tedla, Ato Wallelign Mamo and Ato Feysel Mohammed for their comment and suggestion. They were all helpful and informative.

I am greatly indebted to my brothers Girma Dawd (Degole) and Elias Kassaye for their support during the last two years of my stay in Addis Ababa University.

Finally I would like thank my best friends Kalab Admasu and Dejene seid, my class mates, all staff members of Information Science Department and those who are not mentioned in name for their worthwhile contribution and support.

Table of Contents

Dedication i

Acknowledgment ii

Table of Contents iii

List of Figure vii

List of Table viii

List of Appendices ix

List of Abbreviations x

Abstract xi

Chapter One

Introduction 1

1.1. Background 1
1.2. Statement of the problem 7
1.3. Related Literature Review 10
1.4. Objectives 11
1.4.1 General objective 11
1.4.2 Specific objectives 11
1.5. Research Methodology 12
1.5.1. Literature Review 12
1.5.2. Business understanding 12
1.5.3. Data Collection 12
1.5.4. Data Preparation 12
1.5.5. Training and Building Model 13
1.5.6. Performance Evaluation 13
1.6. Scope and Limitation 13
1.7. Research Contribution 14
1.8. Thesis organization 14

Chapter Two

Data Mining 16

2.1. Introduction to Data Mining 16
2.2. Data Mining and Knowledge Discovery 19

2.2.1. Business/Research Understanding 20
2.2.2. Data Understanding 20
2.2.3. Data Preparation 20
2.2.4. Data Modeling 20
2.2.5. Evaluation of the Model 21
2.2.6. Deploy the model 21
2.3. Data Mining and Data Ware Housing 22
2.3.1. Design of Data Ware House 24
2.4. Data Mining and Online Analytical Processing 25
2.5. Data Mining Activities 26
2.5.1. Characterization and Discrimination 27
2.5.2. Association Analysis 28
2.5.3. Classification and Prediction 28
2.5.4. Cluster Analysis 28
2.5.5. Outlier Analysis 29
2.5.6. Evolution Analysis 29
2.6. Data Mining Application 29
2.6.1. Application of Data Mining Technology in Banks 31
2.7. Data Mining Techniques 33
2.7.1. Overview of Neural Network 33
2.7.1.1. Structure of Neural Network 35
2.7.1.2. Function of Neural Network 37
2.7.1.3. Classification of Neural Network 40
2.7.1.4. Training Neural Network 40
2.7.1.4.1. The Back-Propagation Learning Algorithm 41
2.7.1.5. Advantage and Disadvantages of Neural Network 42
2.7.2. Overview of Decision Tree 43
2.7.2.1. Decision Tree Structure 43
2.7.2.2. Decision Tree Algorithm 44
2.7.2.3. Attribute Selection in Decision Tree 45
2.7.2.4. Decision Tree Induction 48
2.7.2.4.1. Tree Growing 48
2.7.2.4.2. Pruning 49
2.7.2.5. Understanding the Output of Decision Tree 49

Chapter Three

The Current Loan Approval Procedure at United Bank S.C. 51

3.1. Types of Loans 51
3.1.1. Term Loans 51
3.1.2. Overdraft Facility 51
3.1.3. Letter of credit facility 52
3.1.4. Merchandise Loans 52
3.1.5. Pre- Shipment Credit Facility 52
3.1.6. Advance on Export Bills 53
3.1.7. Letter of Guarantees 53
3.2. Economic Sectors to be served 53

3.3. Credit Functions Structure and Approval Authority 54
3.4. Requirement for Loan Request 56
3.5. The Loan Approval Processes 56
3.5.1. Evaluation of the Document.....57
3.5.2. Business Visit.....58
3.5.3. Analysis of Financial Statement 58
3.5.4. Collecting Previous Credit History 58
3.5.5. Evaluation of Collateral 58
3.5.6. Recommendation and Approval of the Loan 59
3.5.7. Disbursing the Loan Money.....60
3.6. Credit Follow-Up 60
3.7. Observation of the Survey.....61

Chapter Four

Experimentation..... 63

4.1. Business understanding.....64
4.2. Data understanding 64
4.3. Data Preparation.....64
4.3.1. Data collection 65
4.3.2. Defining the Target Class.....67
4.3.3 Data preprocessing 68
4.3.3.1. Handling Missing Values.....68
4.3.3.2. Data Summarization.....70
4.3.3.3. Inconsistent Data Handling.....71
4.3.3.4. Deriving New Attributes from Existing Attributes 72
4.4. Modeling 74
4.4.1. Data Mining Software Selection 74
4.4.2. Attribute Selection and Model Building Using Decision Tree.....77
4.4.2.1. Data Organization and Preparation 77
4.4.2.2. Attribute Selection and Decision Tree Model Building.....77
4.4.2.3. Attribute Selection Result 82
4.4.2.4. Generating rules from Decision tree 85
4.4.3. Neural Network Model Building 87
4.4.3.1. Data Organization and Preparation 87
4.4.3.2. Selection and Interpretation of Neural Network 91
4.5. Evaluation of Decision Tree and Neural Network.....93
4.6. Deploy the model.....94

Chapter Five

Conclusion and Recommendation..... 95

5.1. Conclusion 95
5.2. Recommendation 98

References **99**

Glossary of Terms **104**

List of Figure

Figure 2. 1: CRISP-DM reference model 22

Figure 2. 2: A simple Neural Network with input, hidden and output neurons 36

Figure 2. 3:The unit of an Artificial Neural Network 38

Figure 2. 4: Logistic function of Neural Network 39

Figure 2. 5: Hyperbolic tangent function of Neural Network 39

Figure 2. 6: A simple decision tree structure 44

Figure 3.1: Credit function structure 54

Figure 3.2: Flow chart for loan approval process. 57

Figure 4.1: The flow diagram of the experimentation 63

Figure 4.2: Weka's preprocessing interface 75

List of Table

Table 4. 1: Collected records from twenty branches 65

Table 4. 2: Loan customer classification in United Bank S.C. 67

Table 4. 3: Attributes treated by nominal case for missing values.....69

Table 4. 4: Distribution of records with respect to payment performance 69

Table 4. 5: Inconsistent data handling for term of payment 71

Table 4. 6: Inconsistent data handling for Collateral type 72

Table 4. 7: The initial Total attributes 73

Table 4. 8: Attributes selected by attribute selection feature of Decision Tree.....78

Table 4. 9: The confusion matrix of 5 attributes 79

Table 4. 10: The confusion matrix of 22 attributes 80

Table 4. 11: The selected attributes by J48 decision tree 81

Table 4. 12: The results found by using different attributes combination 82

Table 4. 13: The confusion matrix for the 14 attributes. 83

Table 4. 14: The results found by using different percentage split value 84

Table 4. 15: The results found by varying learning rate.....89

Table 4. 16: The results found by varying the number of neurons.....89

Table 4. 17: The results found by varying number of training time.....90

Table 4. 18: The results found by varying the percentage split mode.....91

Table 4. 19: The confusion matrix of a model with learning rate 0.2.....92

Table 4. 20: The confusion matrix of a model with number of neurons 16.....92

Table 4. 21: The confusion matrix of a model with number of training time 1500.....92

List of Appendices

Annex 1: Decision Tree constructed Using 14 attributes **105**

Annex 2: List of initial attributes **108**

Annex 3: Commercial Credit Report **109**

Annex 4: Loan Approval Form (LAF) **113**

Annex 5: Monthly Loans and Advance Return Form **115**

Annex 6: Sample data for Loan Customers **116**

List of Abbreviations

AI: -Artificial intellegence

ARFF: -Attribute-relation file Format

CART: - Classification and Regression tree

CCR: -Commercial Credit Report

CHAID: - Chi-squared Automatic interaction detection

CRISP-DM: -Cross industry standard process for data mining

CSV: -Comma-separated value

KDD: - Knowledge discovery in database

LAF: -Loan approval form

NBE: - National Bank of Ethiopia

OLAP: - On-line analythical processing

SQL : - Structure query language

S.C: - Share company

WEKA: - Waikato environment for knowledge analysis

Abstract

A Commercial Bank is a financial intermediary that holds deposits for individuals and businesses in the form of checking and savings accounts and certificates of deposit of varying maturities while it issues loans in the form of personal and business as well as mortgages. It arises due to a debtor's non-payment of a loan or other line of credit.

In order to control and manage the risk, banks normally have discipline called risk management. Hence it is very important to develop and implement an effective technology that can support risk management. This research focused on the application of data mining techniques in supporting loan risk assessment taking as case study United Bank Share Company. It used two data mining techniques namely, decision tree and neural network.

Different decision tree models using j48 algorithm were constructed during the experiments and among them a tree with overall accuracy of 95.65% with conceivable rule was selected. The important attributes that were identified by the selected decision tree were: Networking capital, Current Ratio, Total Asset, TL/TA, Current Liability, Collateral Value, Years in Business, Number of prior term loans settled, Performance of term PriorLoans, Collateral Type, Credit Relationship with other bank, Trade Sector, Performance in other types of loan and Current Asset.

Based on the above selected attributes different types of neural network models with multilayer perceptron algorithm were constructed and a model that maximizes the accuracy in predicting poor payment performance was selected with over all accuracy of 92.83%.

When evaluation was done, the overall accuracy of decision tree found better than the neural network even if further research is needed. In addition the result of decision tree is more interpretable than neural network. In general the result showed the possible application of data mining in loan risk assessment term loan.

- **Market risk**: - is the risk that the value of an investment will decrease due to moves in market factors.
- **Liquidity risk**: - arises from situations in which a party interested in trading an asset cannot do it because nobody in the market wants to trade that asset.
- **Operational risk**: - is defined as the risk of loss resulting from inadequate or failed internal processes, people and systems, or from external events.
- **Compliance risk**: - is the current and prospective risk to earnings or capital arising from violations of or nonconformance with laws, rules and regulations.
- **Regulatory risk**: - The risk of a change in laws and regulations that will materially impact a security, business sector or market.
- **Reputation risk**: - is the risk that negative publicity regarding an institution's business practices will lead to a loss of revenue [12].

Among the stated types of risks, the researcher would focus on loan risk. It is a type of risk that challenges the banking sector and arises from the potential that an obligor is either unwilling to perform commitments in relation to lending, trading, settlement and other financial transactions.

Normally loan risk emanates from the bank's interaction with individuals, corporate, financial institutions or a sovereign. It could also stem from activities both on and off balance sheet.

In relation to other risks, loan risk not necessarily occurs in isolation. The same source that endangers loan risk for the institution may also expose it to other risk. For instance a bad portfolio may attract liquidity problem [33].

But currently the innovation of data mining technology with its great potential in identifying various interesting patterns, made organizations to control data resources for strategic planning and decision making in their domain area.

The theme of this research is focused on the application of data mining technology in supporting term loan risk assessment. It used decision tree and neural network techniques in order to predict the payment performance of prospective borrower. As a result the bank's loan experts can make decision easily.

Decision tree would help in identify a handful of the most important attributes in addition to the predicting capability and the ability of neural network to simulate learning and discover relationship and correlation within data as the model gains experience and makes them very adaptable for use in financial sector, including banks [4].

The foregoing discussions infer that the application of decision tree and neural network has paramount importance in building a predictive model. Detail descriptions on each technique will be presented in chapter two.

In Ethiopia until 1994, government owned banks monopolized the bank sector. But after 1994, the establishment of private commercial banks became possible and the number has been increasing. This in turn create an opportunities for the societies to get services from different banks.

Nowadays these commercial banks gaining more customers and became a good competitors due to better services. Among the different private commercial banks the focus of this research paper is United Bank S.C.

United Bank was incorporated as a Share Company on 10 September 1998, in accordance with the commercial code of Ethiopia of 1960 and the licensing and supervision of banking business proclamation No. 84/1994.The bank obtained a banking services license from the National Bank of Ethiopia and is registered with the Trade, Industry and Tourism Bureau of the Addis Ababa City Administration.

United Bank over the years built itself into a progressive and modern banking institution, endowed with a strong financial structure and strong management, as well as a large and ever-increasing customers and correspondent base. At the end of June 2007, United Bank reported a net profit with a return on equity of 35.87%.

Today, it is a full service bank that offers its customers a full range of commercial banking services with a net work that includes 36 branches. Its priority in the coming years is to strengthen its capital base, maximizing return on equity and benefit from the latest technology in order to keep abreast with the latest developments in the local and international financial services industry.

The bank has the following major objectives;

- Mobilizing all types of deposits (saving, demand and time) and pay interest-bearing accounts.
- Providing loans and advance to its customers, including long term investment/project financing.
- Providing domestic and international money transfer services.
- Providing international banking services.
- Buying and selling travelers' cheques and foreign currency notes.

- Providing deposit services in foreign currency for Ethiopian nationals and foreign nationals of Ethiopian origin.

Related to loan, the bank provides loan to business investment who are engaged in profitable business establishments to support their working capital and investment needs. The economic sectors which are beneficiaries of loan service from United Bank S.C. are;

- Domestic Trade
- International Trade
- Factories
- Service Sector
- Transport Sector
- Construction Sector
- Hotels and Tourism

1.2. Statement of the problem

Loan is the process of transferring monetary elements to once who satisfies the requirement of getting loans according to the rules and regulation of the bank. After certain period of time it should be returned to the bank according to the loan agreements that have been made between the two parties.

Loan comprises a very large portion of a bank's total assets, and it also forms one of the most essential operations of a bank as a result the strength of a bank is judged by the soundness of its different kinds of loans. In connection with this, it is obvious fact that the economic development of a country greatly supported by the net inflow resource from a bank.

But currently a declining rate of loan collection and existence of default rate particularly in term loan are the treat for the bank sector in order not to give the services to the maximum limit and the development of a country like Ethiopia.

In order to control the high defaulted rate, commercial banks have sought different approaches and techniques but still there is a challenge on how to implement loan risk assessment due to many reasons. For example, even though voluminous data about borrowers is stored, but almost all banks in Ethiopia do not have well organized tools for these data analysis and identify some knowledge from the data for decision making. Besides, this crucial data is found in manual format.

Managing loan request by a financial institution like bank is one of the important activities that require delicate care. The bank can have loan officers to make credit decisions or recommendations for the bank.

These officers are given some basic rules to guide them in evaluating the trustworthiness of loan applications. Based on the rules and guide lines, these officers able to decide whether an application is loan worthy or not.

Even if it is possible to predict the performance of prospective borrower to some extent by officers, it is not possible to be sure based on certain attributes and cases. The analysis made by using traditional methods focuses on problems with much more manageable number of attributes and cases than may be encountered in real world, hence they have limited capacity to discover new and unanticipated patterns and relationships that are hidden in conventional databases [31]. In addition, these officers could not handle the risk assessment manually as the volume of the data collected and stored increases as time passes by.

In addition to the above the capability of humans to judge the loan applications trustworthiness may be affected by the following reasons;

- The presence of a physical or emotional condition can affect the decision making process.
- Personal connections with the applicants might distort the judgmental capability

In general, lending is inherently risky but the risk can be minimized and controlled through the application of data mining technology.

Current studies clearly depicted that this technology has much applicability in financial sectors due to its ability to uncover and mine relevant information from huge accumulated data. It provides a variety of useful tools for discovering none obvious relationships in historical data. As ensuring those relationships discovered will generalize to the future data that can be used by the loan officers in assisting in rejecting or accepting a loan request.

Based on the foregoing discussion there is a need to undertake a research and the following major questions guided the research work:

- Is data mining applicable for loan risk assessment particularly in term loan.
- Which attributes must be given more emphasis to reject or accept a loan request of prospective borrower.
- Which data mining techniques from neural network or decision tree more efficient in supporting term loan risk assessment.

1.3. Related Literature Review

In the Ethiopian context related to this research, two researches were done, namely Data mining Application in Support of Loans Disbursement Activity at Dashen Bank done by Askale Worku [3], and Possible Application of Data mining technology in Support of Credit Risk Assessment: The Case of NIB International Bank S.C. done by Mertework Shawel [26].

Each of them concluded that the result obtained was encouraging and showed the applicability of data mining technology in the banking sector in particular in loan approval process. The research done by Askale Worku [3] used a dataset with 898 records,25 independent attributes and one class label with values regular and irregular. The researcher constructed different neural network models by varying the values of the parameters and combination of attributes. Finally the model whose overall accuracy of 88% using 13 attributes combination was selected as the best model.

The research done by Mertework Shawel [26]used a dataset with 922 records,28 independent attributes and one class label with values regular and irregular.

The researcher constructed different decision tree models by varying the combination of attributes. Finally the tree, with accuracy of 93.69% and conceivable rules was selected. This tree used 9 attributes for model building.

Some of the recommendations forwarded by these two researchers were; conducting the experiment with large dataset, using both decision tree and neural network techniques on the same dataset to know which techniques will be the ultimate solution for loan risk assessment and introduce more detailed classification of customer category.

Hence this research was conducted to supplement the above two stated researches by using the decision tree and neural network techniques on dataset with 5006 records and by introducing a class label with three categories for the case of United Bank S.C.

1.4. Objectives

1.4.1 General objective

The general objective of the research is to examine the possible application of data mining technology in supporting of term loan risk assessment.

1.4.2 Specific objectives

To accomplish the general objective, the specific objectives of the research are:

- Develop an understanding of loan risks concept specially term loan.
- Review literature on data mining technology with more emphasis to the decision tree and neural network techniques and their applications.
- Identify data source and collect the required data from United Bank S.C.
- Preparing and analyzing the data for model building.
- Select the data mining software that supports neural network and decision tree techniques.
- Train and build as well as test the performance of the models.
- Make comparison among the results found from decision tree and neural network and select the one which performs the best.
- Make conclusion and forward recommendation.

a great effort was made. Hence an alternative source of data was looked and United Bank S.C. was selected as a good choice based on the willingness of the bank officials in order to provide the appropriate data as well as assistance during the research period.

Second the data set for this research was collected almost from manual document as a result the researcher was limited by smaller number of records than expected.

1.7. Research Contribution

This research tried to investigate applicability of data mining technology in supporting loan risk assessment for the case of United Bank S.C.

It can support the routine, repetitive and tedious works of loan officers in processing loan request, as a result the bank can improve the quality of service, which is one of the key factors for the well being of any bank and reduce the loan risk. In addition to the above, it can be used as a source of methodological approach and support decision making in loan risk assessment for other existing and new coming commercial banks with slight modifications.

1.8. Thesis organization

This research report is organized into five chapters. The first chapter briefly discusses background to the problem area, statement of the problem, objectives and research methodology, scope and limitation and research contribution.

Chapter two discusses about data mining technology with its techniques namely decision tree and neural network. Chapter three reviews the general overview of loan and existing loan risk assessment in United Bank S.C.

Chapter four reports the experimentation part. It includes the data collection, data preparation, training and model building process, evaluation of each technique's performance and finally the deployment of the model. The final chapter provides conclusion and recommendation for future work.

Chapter Two

Data Mining

This chapter will discusses the concepts and activities of data mining technology with techniques and the application in related to banking sector.

2.1. Introduction to Data Mining

Companies have kept records of data generated even long after their life times has expired because they believe that there are valuable facts and figures coded within it [11].Even if there is still large amount of data mainly they used it to get endless facts and figure but not for knowledge discovery [2]. This is due to luck of well automated information system that can collects, stores and analysis data.

As a result most companies face with problems in identifying facts and knowledge that are crucial for the day to day activities [17]. The importance of knowledge and information in today's business can never be seen as an exogenous factor to the business. Organizations and individuals having access to the right information at the right moment will have greater chances of being successful in the epoch of globalization.

It is estimated that the amount of information in the world doubles every months and this volume expected to increase in the future [17]. Currently in response to this, huge electronic data repositories are being maintained by banks and other financial institutions across the globe. Even if valuable bits of information are embedded in these data repositories but the huge size of these data sources make it impossible for a human analyst to come up with interesting information that will help in the decision making process [9].

primarily to study data and data relationships are the foundation of most technologies on which data mining is founded upon [17].

The second largest technique for data mining is AI (artificial intelligence). This discipline which is built upon heuristics as opposed to statistics, attempts to apply human thought like processing to statistical problems. AI is mainly used to create new ways in addressing and solving very complex and math driven problems.

Researches in many aspects of intelligence aim to understand human intelligence at all levels, including reasoning, perception, language development, learning and social levels and to build useful artifacts based on intelligence [4].

The third family line of data mining is machine learning, which is the study of computational methods for improving performance by mechanizing the acquisition of knowledge from experience [24].

Machine learning attempts to let computer programs learn about the data they study, so that programs make different decisions based on the characteristics of the studied data. It uses statistics for fundamental concepts and adding more advanced AI heuristics and algorithms to achieve its goal [17].

The extracting of meaningful and unobvious information from database enables to efficiently solve business problems and give a competitive edge. There are many industries that have already highly benefited from data mining capabilities, Such as Banking and Finance, Retail industries, Healthcare and Telecommunication [18].

2.2. Data Mining and Knowledge Discovery

Data mining and knowledge discovery have been attracting great attention to many researchers. Knowledge discovery in database (KDD) was first coined at the first KDD workshop in 1989 to emphasize that knowledge is the end product of a data driven discovery [34]. It has been popularized in the AI and machine learning fields [11].

There is some uncertainty between the term data mining and KDD. It refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. In other word, KDD is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data whereas data mining is the application of specific algorithms for extracting patterns from data.

Most of the time, the term data mining has become more popular in industries, in media and in the database researches as a synonym for knowledge discovery and hence the two terms are used interchangeably [16].

KDD is an iterative sequence of steps in building and implementing a data mining solution and regularities or high level information can be extracted from the relevant sets of data in databases [1]. Normally it can be investigated from different angles in large databases and serve as rich and reliable sources for knowledge generation and verification [23].

According to Cross-industry standard process for data mining (CRISP-DM), the following steps are known; business/research understanding, data understanding, data preparation, data modeling, deploy the model and evaluation the model are the basic steps [28]. Below here each step will be summarized.

2.2.1. Business/Research Understanding

The prerequisite to data mining is to understand one data and the function that one wants it to serve. Without this understanding no algorithm is going to provide one with a result in which they should have confidence [37]. It is the first and the most crucial step in data mining activity used to have clear and unambiguous knowledge about the basic problem.

2.2.2. Data Understanding

Once the business problem clearly defined the next step is selection process since the databases are heterogeneous it requires figuring out what data are needed, which data are the most important and integrating the information in a way that is consistent with the problem to be solved [1].

2.2.3. Data Preparation

As the success of most data mining activity highly depends on the well organized dataset, hence this step involves collecting, cleaning, consolidating and amalgamating records, summarizing fields, checking for data integrity, detecting irregularities and illegal attributes, filling in for missing values, trimming outliers. In this case it is the most time consuming task that needs 70 percent or more of the total data mining effort [4].

2.2.4. Data Modeling

At this point that one invokes data mining models and tools to interrogate the data and convert it into knowledge for decision making [11]. This model building step involves selecting data mining tools, transforming the data if the tools require it, generating samples for training and building the model and finally using the tools to test and select a model [1].

Integrated: - Data warehouse usually constructed by integrating multiple heterogeneous data sources, such as from relational databases files, flat files and on-line transaction records.

Time-variant: - Data are stored to provide information from a historical perspective. Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time. Data warehouse do not contain the most current information.

Non volatile: - Data warehouse is always a physical separate store of data transformed from the application data found in the operational environment due to this separation, a data warehouse does not require transaction processing, recovery and concurrency control mechanisms it usually requires only two operations in data accessing: initial loading and access of data.

It is also a popular business trend of collecting and cleaning transactional data to make them available for online analysis and decision support [11]. Data warehousing employs an update-driven approach in which information from multiple, heterogeneous sources is integrated in advance and stored in a warehouse for direct querying and analysis.

Data warehouse supports information processing by providing historical data for analysis in different sectors, for example in financial, retail distribution and banking sector.

Since Query processing in data ware house does not interfere with the processing at local sources hence historical data analysis is maintained separately from companies' operational database. There is some real benefit if the data to be mined is already part of a data warehouse.

For example the problems of cleansing data for a data warehouse and for data mining are very similar, if the data has already been cleansed for a data warehouse then it most likely not need further cleaning in order to be mined. Further more the use of data warehouse will have addressed many of the problems in data consolidation but a data warehouse is not a requirement for data mining [37].

2.3.1. Design of Data Ware House

A data warehouse can be built using a top-down approach, bottom-up approach, or a combination of both. The top-down approach starts with the overall design and planning. It is useful in cases where the technology is mature and well known, and where the business problems that must be solved are clear and well understood [16].

The bottom-up approach starts with experiments and prototypes. This is useful in the early stage of business modeling and technology development. It allows an organization to move forward at considerably less expense and to evaluate the benefits of the technology before making significant commitments.

In the combined approach, an organization can exploit the planned and strategic nature of the top-down approach while retaining the rapid implementation and opportunistic application of the bottom-up approach. The design and construction of a data ware house may consist: the design, data integration and testing and finally deployment of data warehouse.

In general setting up a large data warehouse can resolves data integrity problems but loading the data into a query database can be an enormous task, sometimes it takes years and costing millions of dollars.

The objects are clustered or grouped based on the principle of maximizing the inter-class similarity and minimizing the intra-class similarity hence objects that are formed within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters [16].

2.5.5. Outlier Analysis

A database may contain data objects that do not comply with the general behavior or model of data called outliers. Most data mining methods discard outliers as noise or exceptions. The analysis of outlier data is referred to as outlier mining. Outliers may be detected using statistical method when it is based on distribution, using distance measure when it is based on distance and using deviation method when it is based on the main characteristics of objects in a group [16].

2.5.6. Evolution Analysis

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association, classification or clustering of time-related data, distinct features of such analysis can include time-series data analysis, sequence or periodicity pattern matching and similarity-based data analysis [16].

2.6. Data Mining Application

Data mining can give answers to business questions that traditionally were time consuming to resolve [35]. Its powerful features allow businesses to make proactive and knowledge-driven decisions.

Today it is primarily used by companies with a strong consumer focus retail, financial, communication, and marketing organizations.

Data mining enables these companies to determine relationships among internal factors such as price, product positioning, or staff skills, and external factors such as economic indicators, competition, and customer demographics. Furthermore, it enables to determine the impact on sales, customer satisfaction, corporate profits by drilling down into summary of information [6].

Different companies use data mining for different purposes. Below here there are a few areas in which companies use it achieve a strategic benefit.

Direct Marketing

The idea here is to find out who is most likely or most desirable to buy certain products. This information can be used for several marketing activities.

Trend Analysis

Trend analysis used by companies to predict trends in the marketplace. Using this information can lead to have strategic advantage because it is useful in reducing costs and timeliness to market.

Fraud Detection

Companies use data mining techniques to model which business transactions are likely to be fraudulent. As a result it can be used for insurance claims, cellular phone calls or credit card purchases.

2.6.1. Application of Data Mining Technology in Banks

Currently data mining tools address problems related to businesses that were previously impossible due to lack of processing capabilities [15].

It is only through the application of data mining techniques that a large enterprise can hope to turn the myriad records in its customer's databases into some sorts of coherent picture of its potential customers [4].

Data mining offers value across a broad spectrum of industries that fit these requirement including banking, credit customer relationship management, healthcare, human resources, insurance, marketing, retail, telecommunication and manufacturing.

According to the professional and trade literature, more industries are using data mining as the foundation for strategies that help them outsmart competitors, identify new customers and lower costs[20].

From the above banking sector is one of the industries that has experience in many changes in information technology as a result the sector can reduced cost and time of data processing, increase the profit and become competitive in the industry [18]. Since banks offer a wide variety services hence the application of data mining can make them to identify underserved populations, evaluating loan payment prediction, analyzing profitability, direct marketing and detecting credit card fraud [18].

Otherwise it will be very difficult to instantly generate a pattern analysis by experts when the volume is larger [9].

2.7.1.1. Structure of Neural Network

Artificial neural networks may be represented by different structure, but they are each designed to make use of some of the organizational principles felt to use by the brain [33].The artificial network consist of processing units called neurons that are linked to certain of its neighbors with varying coefficient of connectivity that represent the strength of the connection.

The processing units transport the incoming information on their out going connection to other units. The electrical information is simulated with specific values stored on those weights that makes these networks have the capacity to learn, memorize and create relationship amongst data.

Normally the set of neurons processing the entire neural network task and each is acting like a separate computation device by doing its own job. In addition the system is inherently parallel in that many units can carry out their computations at the same time. The most common type of artificial neural network consists three layers of units as listed below.

- **Input Layer**: It receives raw information for neural network and it is connected to hidden layer.
- **Hidden Layer**: It processes information between the input layer and output layer. The activity of each hidden unit is determined by the activities of the input units and the weight on the connections between the input and the hidden units.

- **Output Layer**: It produces neural network's output. The behavior of the output unit depends on the activity of the hidden units and the weight between the hidden and output units.

The hidden layer makes the network recognize more patterns, therefore the number of hidden nodes often increase with as the number of inputs and the complexity of the problem increase. But too many hidden nodes can lead to over fitting, and too few hidden nodes can result in models with poor accuracy.

Even if there could be a number of input, hidden and output neurons in each corresponding layer finding an appropriate number of hidden nodes is an important part of any data mining effort in neural network [7].

In the Figure 2.2 below there are three inputs, three hidden and three output neurons and the network has one input, one hidden and one output layer.

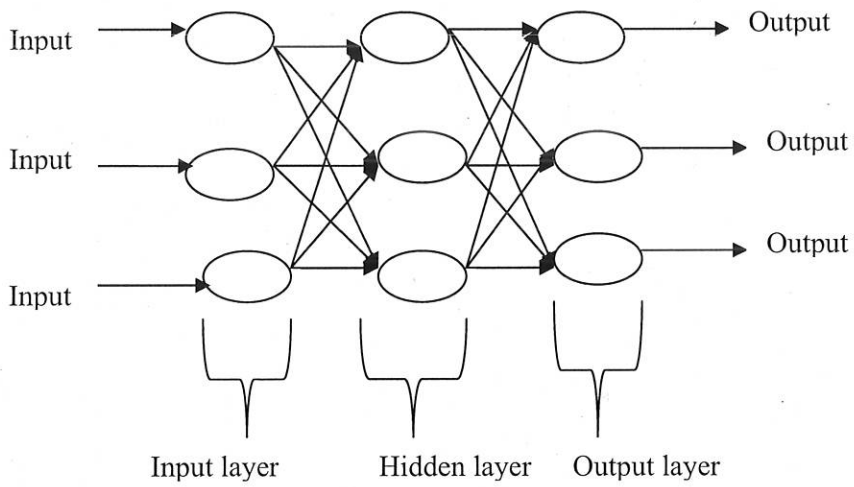


Figure 2. 2: A simple Neural Network with input, hidden and output neurons

2.7.1.2. Function of Neural Network

Like its biological counterpart, the unit in a neural network has the property that small changes in the inputs can have relatively large effects on the output.

Conversely; large changes in the inputs may have little effect on the output. This property where sometimes small changes matter and sometimes they do not is an example of nonlinear behavior. The power and complexity of neural networks arise from their nonlinear behavior, which in turn arises from the particular activation used by constituent neural units.

The neuron combines its inputs into a single value, which it then transforms to produce the output; these together are called the activation function. The most common activation functions are based on the biological model where the output remains very low until the combined inputs reach a threshold value.

When the combined inputs reach the threshold, the unit is activated and the output is high. The activation function has two parts. The first part is the combination function that merges all the inputs into a single value. Each input into the unit has its own weight.

The most common combination function is the weighted sum, where each input is multiplied by its weight and these products are added together. Although there is a lot of flexibility in the choice of combination function, the standard weighted sum works well for the combination function in many situations.

The second part of the activation function is the transfer function, which gets its name from the fact that it transfers the values of the combination function to the output of the unit.

There are three typical transfer functions namely: the logistic (sigmoid), linear and hyperbolic tangent [4]. In the figure 2.3 below the unit of artificial network is given with inputs associated to each input neuron with specified weight, combination function, transfer function and output.

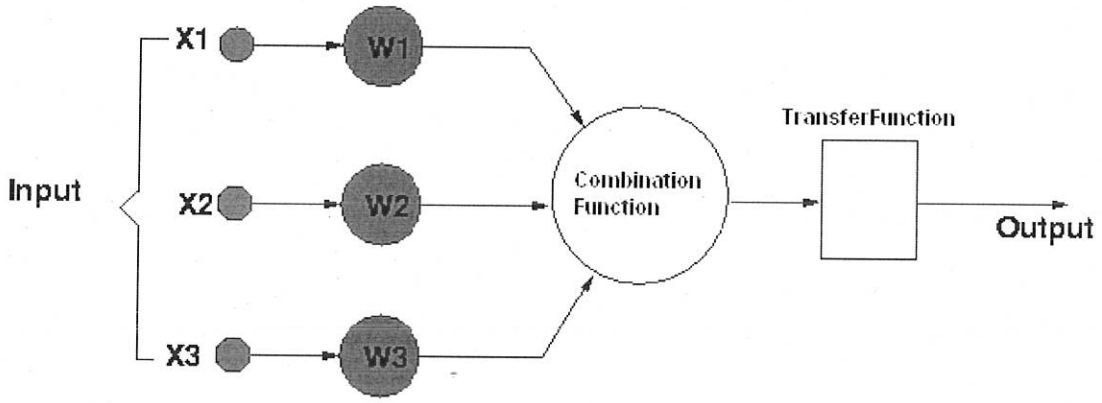


Figure 2. 3: The unit of an Artificial Neural Network

The logistic and the hyperbolic tangent are the most common neural networks. The major difference between them is the range of their outputs, between 0 and 1 for logistic and between -1 and 1 for the hyperbolic tangent.

Even though they are not linear, when the weighted sum of all the inputs is near 0, then these functions are a close approximation of a linear function and as magnitude of the weighted sum gets larger, these transfer functions gradually saturate (to 0 and 1 in the case of the logistic; to -1 and 1 in the case of the hyperbolic tangent) [4].

- The network gets a training example and using the existing weights in the network, it calculates the output or outputs. This forward pass would produce the actual or predicted output.
- Multilayer perceptron then calculates the error by taking the difference between the calculated result and the expected result (desired).
- The error is fed back through the network and the weights are adjusted to minimize the error, hence the name back propagation is given because the errors are sent back through the network. The weights are then adjusted and the neural network is said to have learned from experience [4].

The above steps are repeated iteratively until the weights on the network no longer change significantly and error no longer decreases.

2.7.1.5. Advantage and Disadvantages of Neural Network

Neural network is able to capture associations or discover regularities within a set of attributes. The application domain of neural networks very much depends on the nature of the problem being modeled. In general neural network has the following advantages and disadvantages [35].

Advantages

- Have a great applicability in the case when the number of attributes or the volume of data is very diverse.
- It used in finding the relationship among attributes when it is inherently complex and cannot easily be identified.
- It used for modeling diverse behavior by finding patterns among cases.

Disadvantages

- It is viewed as a black box and there is no explanation of the result.
- It suffers from long learning times which become worse as the volume of the data grows.
- It lacks of diagnostic help. If something goes wrong, it is difficult to pinpoint the problem from the mass of inter-related nodes and links in the network.

2.7.2. Overview of Decision Tree

Decision tree is powerful and popular for both classification and prediction. It also useful for exploring data in order to gain insight into the relationships of a large number of candidate input attributes to a target attributes. Because decision tree combines both data exploration and modeling, it is a powerful in the modeling process [4].

Decision tree usually divided into regression trees and classification tree. The regression one is used when the response attribute is continuous and the classification one is used when the response attribute is quantitative discrete or qualitative [14].

2.7.2.1. Decision Tree Structure

A decision tree is a structure that can be used to divide up a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules that we humans can understand, with each successive division, the members of the resulting sets become more and more similar to one another [4]. In another word decision tree performs many tests and then try to arrive at the best sequence for predicting the target. Each division creates branches that lead to more divisions until it terminates in a leaf node. The path from

root to the target leaf is the rule that classifies the target. The rules are expressed in If-then form [7].

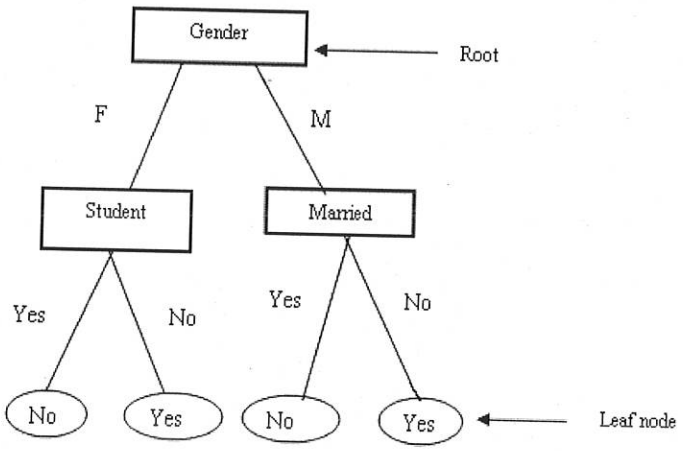


Figure 2. 6: A simple decision tree structure

Figure 2.6 shows a simple classification decision tree, for the concept of Have a Car, indicating whether or not an individual at specific company is likely to have a car, with the rectangular shapes show the attribute and the branching depict the different values of the attributes and the oval shapes show the value of the class label.

For example in the above decision tree an individual whose gender is male and is not married has a car and an individual whose gender is female and a student has not a car.

2.7.2.2. Decision Tree Algorithm

Although all the decision tree algorithms share the same basic procedure there are many variations on the core of decision tree algorithm. Decision tree algorithms commonly used for decision tree construction include, Chi-squared Automatic Interaction Detection (CHAID), Classification and Regression trees (CART), C4.5 and C5.0 [4].

The distinguishing features between tree algorithms include;

- **Target attributes**: - Most tree algorithms require the target (dependent) attribute to be categorical. Such algorithms require that continuous attributes are binned (grouped) for use with regression [7].
- **Splits**: - Many algorithms support only binary splits, that is, each parent node can be split into at most two child nodes. Others generate more than two splits and produce a branch for each value of a categorical attribute [7].
- **Split measures**: - Help select which attribute to use, to split at a particular node. Common split measures include criteria based on Information gain, Gini index, Chi-squared and Gain ratio [7].
- **Rule generation**: - Algorithms such as C4.5 and C5.0 include methods to generalize rules associated with a tree; this removes redundancies, other algorithms simply accumulate all the tests between the root node and the leaf node to produce the rules [7].

2.7.2.3. Attribute Selection in Decision Tree

The purpose of data mining is to explore the data and to ultimately discover certain relationships, rules, correlations that can give some insights about the data and also serve for prediction. Instances are evaluated and classified based on the values of their attributes .Thus an attributes that may be irrelevant to the process of classification should be excluded [4].

Relevant attributes may contain useful information directly applicable to the given task by itself, or the information may be hidden among a subset of attributes. Therefore, the selection of a subset of essential attributes is an important part of data mining.

Reducing the number of attributes from hundreds to within a few dozen, not only speeds up the learning process, but also prevents most of the learning algorithms from getting fooled into generating an inferior model by the presence of many irrelevant or redundant attributes. Since most practical learning algorithms are heuristic in nature and they often are misled by the presence of many nonessential attributes [21].

Decision trees constitute different measures in order to choose the best attribute. One of the most common measure is information gain that used in decision tree to select the attribute at each node in tree. The attribute with the highest information gain (greatest entropy reduction) is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions and reflect the least randomness or "impurity" in these partitions [16].

Let S be a set consisting of s data samples. Suppose the class label attribute has m distinct values defining m distinct classes, Ci (for i = 1… m). Let si be the number of samples of S in class Ci. The expected information needed to classify a given sample is given by:

$$I(s1\ldots sm) = -\sum_{i=1}^{m} P_i \log_2(P_i) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textbf{(2.1)}$$

Where: $P_i = \frac{s_i}{s}$ is the probability that an arbitrary sample belongs to class C_i.

s_i, is the number of samples of S in class C_i.

To select the test attribute (i.e., the best attribute for splitting), the entropy and information gain need to be calculated for each attribute. Therefore, if an attribute A has v distinct values, {a1, a2… av}, then attribute A can be used to partition S in to v subsets such as {S1, S2… Sv}.

Where Sj contains those samples in S, those have value aj which is in A. If A were selected as the test attribute (i.e., the best attribute for splitting), then S1, S2... Sv would correspond to the branch grown from the node containing the set S. The entropy, or expected information based on the partitioning into subsets by an attribute A, is given by:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} \cdot I(s_{ij}, \ldots, s_{mj}) \quad \text{(2.2)}$$

Where: s_{ij} - is the number of samples of class C_i in a subset S_j

$\frac{s_{1j} + \ldots + s_{mj}}{s}$ - acts as the weight of the j^{th} subset and is the ratio of number of samples in the subset to total samples in S

The smaller the entropy value, the greater will be the purity of the subset partitions. It should be noted that, for a given subset Sj,

$$I(s_{1j}, s_{2j} \ldots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij}) \quad \text{(2.3)}$$

Where: $p_{ij} = \frac{s_{ij}}{|S_j|}$ - is the probability that a sample in S_j belongs to class C_i

As a result, the encoding information that would be gained by branching on attribute A is:

$$\text{Gain}(A) = I(s_{1j}, s_{2j} \ldots, s_{mj}) - E(A) \quad \text{(2.4)}$$

In other words, Gain (A) is the expected reduction in entropy caused by knowing the valued of the attribute A. An attribute that yields maximum information gain will be chosen for data

set partitioning, then a node is created and labeled with the chosen attribute, branches are formed for each value of the attribute, and the samples are partitioned accordingly.

The same criteria will then be applied to each split sample. The iterative divide and conquer process executes until no further split is required.

2.7.2.4. Decision Tree Induction

The training process that creates the decision tree is called induction and requires small number of passes through the training set. Most decision tree algorithms go through two phases: a tree growing (splitting) phase followed by pruning phase [7].Each phase will be discussed below.

2.7.2.4.1. Tree Growing

The tree growing phase is an iterative process which involves splitting the data into progressively smaller subsets. One of the basic algorithm for decision tree induction is greedy algorithm that constructs tree in top-down recursive divide-and-conquer manner [16].In this case greedy means the algorithm does not look forward in the tree to see if another decision would produce a better overall result.

The first iteration considers the root node that contains all the data. Subsequent iterations, work on derivatives nodes that will contain subsets of the data. At each split, the attributes are analyzed and the best split is chosen.

Tree building algorithms usually have several stopping rules. These rules are usually based on several factors including maximum tree depth, Minimum number of element in a node considered for splitting and the minimum number of element that must be in a new node.

Some algorithms, in fact, begin by building trees to their maximum depth that can precisely predict all the instances in the training set. The problem with such a tree is that, more than likely, it has over fit the data [7].

2.7.2.4.2. Pruning

After a tree is grown, one can explore the model to find out nodes or sub trees that are undesirable because of over fitting or rules that are judged inappropriate. Hence pruning removes nodes and the sub trees created by the above conditions [7].

Pruning is the process of removing leaves and branches from a decision tree as a result the performance of the decision tree will be improved [4]. Algorithms that build trees to maximum depth will automatically invoke pruning. There are two common approaches to tree pruning namely prepruning and post pruning. In the prepruning approach, a tree is pruned by halting its construction early. Upon halting, the node becomes a leaf. The second approach, post pruning removes branches from a fully grown tree [16].

2.7.2.5. Understanding the Output of Decision Tree

Decision trees have obvious value as both predictive and descriptive models. Prediction can be done on a case-by-case basis by navigating the tree. More often, prediction is done by processing multiple new cases through the tree or rule set automatically and generating an output file with the predicted value or class label appended to the record for each case. Once trained, a tree can predict a new data instance by starting at the top of the tree and following a path down the branch until encountering a leaf node [7].

The knowledge represented in decision trees can be extracted and represented in the form of classification IF-THEN rules.

One rule is created for each path from the root to a leaf node. Each attribute value pair along a given path forms a conjunction in the rule antecedent ("IF" part).The leaf node holds the class prediction, forming the rule consequent ("THEN" part).The IF-THEN rules may be easier for humans to understand, particularly if the given tree is very large [16].

2.7.2.6. Advantage and Disadvantages of Decision Tree

Based on the Berry, M.J.A and Linoff, G [4], Witten, I.H and Frank, E [37], some of the advantage and disadvantages are listed below.

Advantages

- It allows a human expert to easily understand the solution of a problem.
- Decision tree make few passes through the data and they work well with many predicator attributes.
- Decision tree can handle raw data with little or no pre-processing.
- Using decision tree, it is possible to pick the most important attributes for predicting a particular outcome because these attributes, chosen for splitting, found in the tree.

Disadvantages

- It is sensitive to change in data and noise.
- Decision tree can never discover rules that involve a relationship between attributes. This puts a responsibility on the miner to add derived attributes to express relationships that are likely to be important.
- Decision tree is error prone when the number of training examples per class gets small.

Chapter Three

The Current Loan Approval Procedure at United Bank S.C.

The purpose of this research was intend to evaluate the possible application of data mining technology in supporting loan risk assessment the case of United Bank S.C. Hence it is important to introduce the current loan approval procedures of the bank.

Based on the interviews with the bank's loan experts and official documents, this chapter will discuss the existing loan approval procedure. First it begins with by introduction of different types of loan available at the bank to be followed by the economic sectors served by the bank, description of the loan approval process and follow-up. At the end, the chapter will discuss on the observation of the survey.

3.1. Types of Loans

Loans at United Bank shall be made in the following forms.

3.1.1. Term Loans

Term loans are credit facilities provided for specific duration. This type of lending is disbursed in lump sum and will be paid by periodic installments to be made within a specified time. The duration can be classified further as short, medium and long term.

3.1.2. Overdraft Facility

This type of loan is operated for some specified time by overdrawing current accounts. The customer is also required to deposit income from sales to this account. Accordingly, this guarantees the healthy operation of accounts.

It has to be renewed either every six month or every year and can be extended to businesses with fast business performance. The customer should request renewals in writing 45 days before the overdraft expires.

3.1.3. Letter of credit facility

This type of loan is extended to importers based on the value of import documents, certain margin of loans are delivered. It is to help importers by collateralizing the merchandise they import so that they will not face shortage of capital. Based on the letter of credit opened up on arrival of the documents, the customers must pay the remaining value of the documents and takeover the document. The margin held limit must be renewed every six month or every year.

3.1.4. Merchandise Loans

This is a type of credit facility granted against the pledge of merchandise goods for a very short duration. Merchandise loan is not a type of credit facility readily available to all types of customers and commodities due to the administrative problem involved and the risk associated with sudden price decline.

3.1.5. Pre- Shipment Credit Facility

This facility is extended to exporters based on export letter of credit for export products. It is availed to exporters who face shortage of working capital in the process of exporting their products via airlines or shipping lines. It can be settled once the letter of credit and other relevant documents are presented to the bank.

3.1.6. Advance on Export Bills

This facility is extended to exporters based on collateralization of shipment document and the volume of the document. The objective of the facility is to help exporters who face shortage of working capital. It is settled once the bill of lading or the airway bill is presented to the bank.

3.1.7. Letter of Guarantees

Letter of guarantee is an unconditional commitment given to a third party on behalf of the bank's customers'. The guarantee issued could be for a local or foreign entity.

3.2. Economic Sectors to be served

As a general policy, United Bank shall provide credit facilities to business organizations, individuals, co-operatives, public enterprise. It is also the general policy of the bank to extend loans to all viable economic sectors and activities without compromising the profit-seeking motive.

The United Bank categorizes loans to different economic sectors and the details of categories of loans currently entertained by the bank are;

- Domestic Trade
- International Trade
- Factories
- Service sector

- Transport Sector
- Construction Sector
- Hotels & Tourism

3.3. Credit Functions Structure and Approval Authority

Although the organizational structure of the lending functions of a bank varies with its size and type of business, the credit structure of United Bank will at all time ensure maximum efficiency in credit processing, clearly delineate responsibility, allow effective credit supervision and ensure efficient credit reporting. Currently United Bank's credit function is structured as shown in the figure 3.1 below.

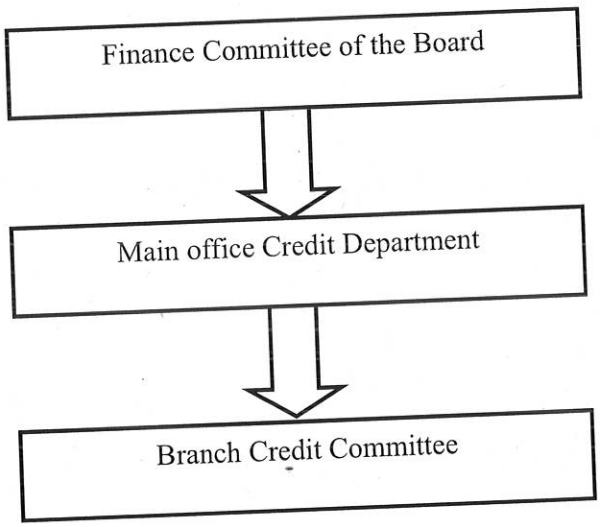


Figure 3.1: Credit function structure

Below here some of the functions of each department will be discussed.

Finance Committee of the Board

- The board retains the ultimate responsibility for ensuring the loan risk management of the bank is properly handled by the management.

- It ensures problem or deteriorating loans or advances are properly and timely identified, classified and appropriate provisions are made for bad loans based on the pertinent directives of NBE (National Bank of Ethiopia).
- The board sets loan approval limits for the Management.

Main office Credit Department

- The credit department shall be in charge of the day-to-day credit functions of the bank. The major functions include the processing of loan recommendations forwarded to it by branches.
- The department shall ensure that the bank's credit policies, procedures, guidelines are complied with by branches in the credit execution and decision making process. It also ensures compliance with the directives issued by the regulatory authority.
- The credit department shall provide interpretations of credit policies and procedures that are not clear or ambiguous to branches.

Branch credit committee

Branch credit department in connection with loan applications do the following

- Handle customer enquiries.
- Provide credit advice.
- Interview the credit applicant.
- Conduct business visits to the customer's premises and gather relevant information for credit analysis.
- Approve loan requests under their discretionary lending limit.
- Dispatch loan requests with their opinion to the appropriate credit committee.
- Collect repayments, and prepare and dispatch periodic reports to higher organs.

3.4. Requirement for Loan Request

In order to minimize and control the risk arises due to loan disbursement it is a must for one bank to obtain enough information about any prospect borrower that is related to income, capital, business sector, previous relationship with other banks and collaterals etc. Hence according to the bank's loan experts and official documents the prospective borrower should satisfy the following criteria.

- Open account in the branch where the loan would be requested.
- Fill an application form mentioning the objective and the amount.
- Supplement a renewed trade license and other relevant licenses.
- Memorandum and articles of association for business with legal personalities.
- Profile of management members if it is organization.
- Business plan.
- Business with legal personality should submit audited financial statements preferably done by external auditors.
- Financial statement showing the business and financial position.
- Balance sheet.
- Income statement.
- Cash flow statement and other relevant document.

3.5. The Loan Approval Processes

Before the loan is granted to prospect borrower the request passes through different evaluation stages. Each stage is described below here. The flow chart, which is given in Figure 3.2, describes the activities were done in the experimentation part briefly.

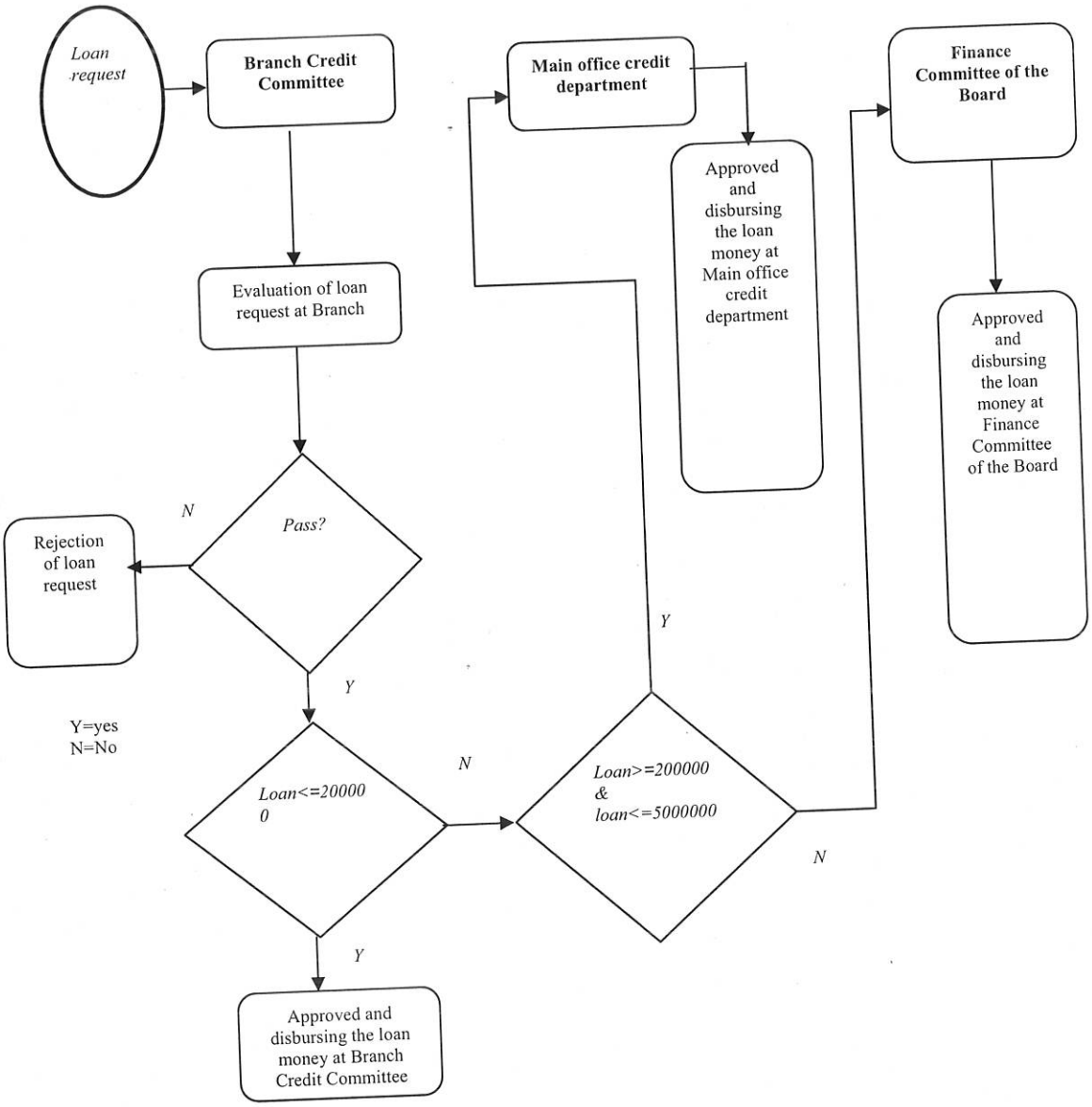


Figure 3.2: Flow chart for loan approval process.

3.5.1. Evaluation of the Document

The first stage of the loan process is to evaluate the document given by prospective borrower for loan application. This preface review of document at branch level ensures that all the required documentation is properly filled based on the given criteria. Hence the subsequent stage can proceed.

3.5.2. Business Visit

Once the document is ascertained that all required documents are properly filled the next stage is to visit the business of the prospective borrower in order to check the overall situation and the business worth based on the information provided.

3.5.3. Analysis of Financial Statement

The third stage is devoted to analyzing the financial statement. Since for any bank to grant a loan the information about financial status is crucial factor. Financial status determines the profitability and viability of the business under consideration. When the financial statement is audited, the banks takes the information as it is, otherwise the required financial statement will be prepared by using the financial credit report form.

3.5.4. Collecting Previous Credit History

Collecting previous loan history about the prospect borrower is very important task for a bank. In collecting the loan history, the bank will not only consider the records found at its own record but it also request records (it includes the previous performance) of the specified borrower from other commercial banks.

3.5.5. Evaluation of Collateral

While granting a loan any bank requires collateral to guarantee for the repayment of the loan money. United Bank share Company accepts as collaterals the following types of assets; houses, buildings ,light and heavy duty dry cargo trucks, trailers, liquid cargo trucks, automobile, public transport buses, bank guarantee, deposits as collateral, merchandise pledge, share certificate, personal guarantee, bills and bonds. But there are few exceptions where a loan may granted without collateral.

3.5.6. Recommendation and Approval of the Loan

Recommendation will be given by the bank's loan experts based on the following five measuring factors,

Account performance: - It gives clear understanding about credit facility utilization of the customer. The bank uses up to three year's information about the customer (if available) or deposit account utilization.

Financial soundness: - It is a pivotal area in assessing a risk arisen from finance point of view since it gives financial position and operational results of the borrower.

Management Quality: - It generally measure utilizing scarce resource efficiently and effectively in setting attainable and measurably objective, and in identifying opportunities.

Banking relationship: - This factor is based on the length of relationship and the total credit exposure that customer has with United Bank S.C.

Collateral strength: - There are different types of collateral to safeguard the possible risk of default. Hence it is very important to identify the risk associated with collaterals. For example vehicles are prone to high level of risks as they are fast depreciable asset and due to mobility character it may be difficult to trace during foreclosure. On the other hand collaterals are not equally realizable that it may differ from one another based on the type of collateral, location, year of make etc.

Based on the above criteria the branch credit committee mainly consist branch manger, assistant branch manager and loan officer put their recommendation in a form known as loan approval form. If the loan request is up to 200000 birr, the loan will be approved at branch level. If it is between 200000 and 5000000 birr, it will be approved by credit committee at

head office mainly consists vice president-operation, manager-credit and risk management department, manager-special branch, head credit analysis division and presenting loan officer. If it is above 5000000 birr, it will be approved by finance committee of the board of the United Bank S.C. The finance committee of the board contains board members, president, vice president-operations, and managers of credit and risk management department.

3.5.7. Disbursing the Loan Money

The disbursement of the loan money is the final stage of the loan process but before disbursement the following condition must be satisfied. The collaterals must be registered and insured by appropriate official bodies so that the bank and the borrower will sign a contract that states the right and obligation of each side.

3.6. Credit Follow-Up

Loan must be followed on a continuous basis so as to check the borrower's financial status. If the loan shows signs of deteriorating financial status, the bank would be able to react during early warning stage. The credit follow-up is one of the main tasks of the managers and the loan officers as well as the credit and risk management department. Follow up should start right after the loan disbursement has been effected and to be exercised until final settlement of the loan in order to be preventive, value saving and remedial for outstanding problem loans. Only with such a close follow-up can a bank avoid ending up with sick loans.

- Major activities in credit –follow up at United Bank S.C. are;
  - ➢ Ensure that loans repayments are made in accordance with the terms and conditions of the loan contract.
  - ➢ Keep regular communication with borrowers through written reminders, telephone and other convenient communication to accelerate repayment.

- Check for early warning signals of each loan.
- Prepare periodic reports on collection status, arrears, overdue loans and non performing loans.
- Finding solution to problems loans.

3.7. Observation of the Survey

Loan is the core business for most banks. This is mainly because most banks generate substantial proportion of their income from interest earned by advancing loan facility to their customers. It is also true that loan constitutes the largest proportion of banks asset.

According to the interview held with the bank's loan experts, currently the bank's loan portfolio is increasing at a faster rate in terms of its volume and number of accounts and borrowers particularly in term loans category. In addition, on the report at June 30, 2008 United Bank S.C. has successfully maintained the lowest non-performing loan ratio in the country so far, which is only 4%. Although majority of the loan customers are non defaulter but as the number of loan customers' increase it is obvious fact that the probability of borrower prone to default will also increase. Hence the bank has to give more emphasis and control the level of defaulter and reconsider the loan controlling system otherwise.

On the other hand, the loan recommendation and approval is the most important stage in loan approval procedure. At this stage, concerning the evaluation of the loan request more emphasis is given to Capital, Current Liability, Collateral Value, Performance of Prior loans, Collateral Type and Credit Relationship with other bank of prospective borrower. In addition performance of prior loans is done based on the information stored in manual format; therefore there may be some problems in getting well organized information about the customers.

Finally when the number of new customers increased, it is obvious fact that work load also increased at different level of the department; as a result it can affect the overall performance of the bank's loan experts. Hence it is necessary to adopt a technology that can give great help.

In response to the above application of data mining technology can help in finding facts that are important for decisions making, shorten the length of the loan approval process, make all important information available in electronic format and decrease the work load.

Chapter Four

Experimentation

In this chapter the researcher describes the different data mining steps that were carried out during experimentation, such as data collection, preprocessing, training and model building. In addition it presents the findings and interpretations. The flow diagram, which is given in Figure 4.1, describes the activities that were done in the experimentation part briefly.

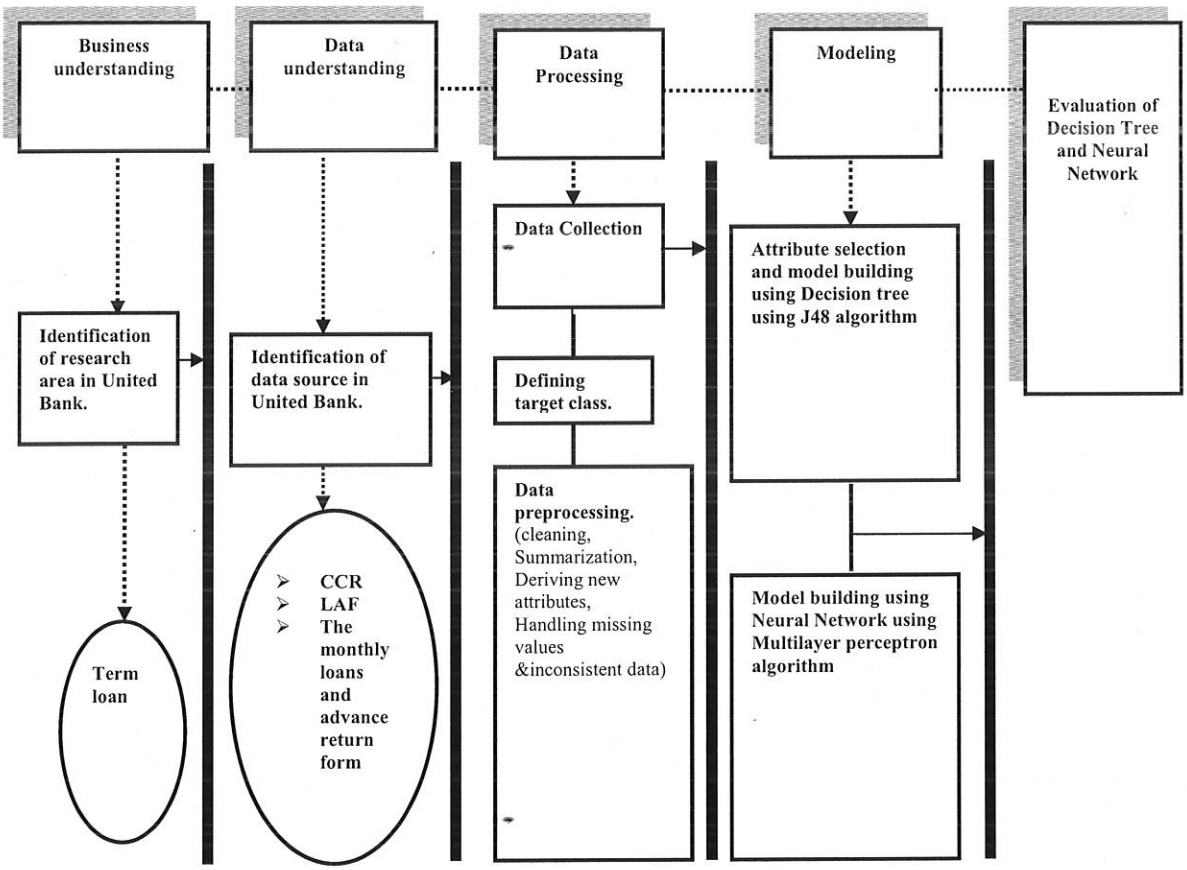


Figure 4.1: The flow diagram of the experimentation

4.1. Business understanding

Based on the survey conducted on chapter three, the researcher focused on loan risk assessment particularly in term loan, since the bank has large number of customer on it and it is the most type of loan that is liable to risk.

4.2. Data understanding

It is obvious fact that relevant data is the main component and crucial factor for successful result of every data mining activities. Therefore the data must be identified and collected from the sources. Corporate data warehouse is a good source of data since the data is collected from different source and integrated in one location with common format [4].

Even if currently United Bank starts to collect the information in electronic format about customers information during in loan approval and follow up stages, but there is no a corporate data warehouse.

Since the previous loan customers data was the most important asset for this research, the researcher with the bank's loan experts used three manual documents namely loan approval form, the monthly loans and advances return form and commercial credit report in order to collect the necessary information.

4.3. Data Preparation

In this stage the following activates were done in order to ready the data for model building as discussed below.

4.3.1. Data collection

Currently United Bank has been giving its services using 36 branches, among them 22 branches are found in Addis Ababa and the rest are found in Dessie, Mekele, BaherDar, Awassa, Adama, Hossana, Gonder, Harar .

Since it was very difficult to access all branch due to time and geographical constraint, the researcher tried to access nineteen branches from Addis Ababa and one from Adama based on the numbers of service year and customer.

Before collecting the data from each branches the researcher considered the period of the bank started its operation, as stated on chapter one it was established in 1998, hence the researcher tried to obtain customer loan records whose life time was between 1999 and 2008.

But the researcher was able to get the loan records from 2002 till end of 2008.The initial number of records collected from the twenty branches is summarized in table 4.1.

No	Branch name	Record size	No	Branch name	Record size
1	Ayer Tena	223	11	Leghar	63
2	Adama	422	12	Lideta	466
3	Bekelobet	1739	13	Mehal Arada	446
4	Birramba	376	14	Mesalemia	57
5	Bole	367	15	Misrak	550
6	Bole Medehanialem	59	16	Shiro Meda	47
7	Bomb Tera	59	17	Tana	27
8	CMC	22	18	Teklehaimanot.	253
9	Hilton	5	19	Wollosefer	182
10	Kailti	178	20	Yereber	93
Total		**5634**			

Table 4. 1: Collected records from twenty branches

Data collection from each branch was take considerable amount of time, since all the important information was found in manual format. The identified attributes were based on those identified by Askale Worku [3], Mertework Shawel [26] and the banks loan experts' opinion. The following attributes were collected from the three sources.

From the Commercial credit report (CCR) (Attached as annex 3).

- Total asset
- Total liability
- Current asset
- Current liability
- Capital
- Business establishment year
- Number of prior term loans settled
- Performance of prior term loans

From the loan approval form (LAF) (Attached as annex4).

- Trade sector
- Credit relationship with other banks
- Performance in other types of loans

From the monthly loans and advance return form (Attached as annex5).

- Granted amount
- Booking date
- Collateral type
- Collateral value
- Term of payment

- Maturity date
- Branch
- Classification.

Basically these are not the only attributes that are found from the sources, but only taking into consideration their importance in loan approval procedure and in driving new calculated attributes. All the above listed attributes are called independent attributes.

4.3.2. Defining the Target Class

A data mining training set has to be pre-classified in order the data mining algorithm know what are looking for [4]. All loans of the bank for the purpose of provisioning are classified according to the standard set by National Bank of Ethiopia in its directive NO.SBB/32/2002 by five types of Classification that are given in table 4.2.

Classification	Description
Pass	Loans and advances in this category are fully protected by the current financial and payment capacity of the borrower and are not subject to criticism.
Special mention	Any loan or advance past due 30 days or more, but less than 90 days.
Substandard	None performing loans and advances past due 90 days or more but less than 180 days.
Doubtful	None performing loans and advances past due 180 days or more but less than 360 days.
Loss	None performing loans & advances past due 360 days or more.

Table 4. 2: Loan customer classification in United Bank S.C.

But the classification adopted in this research has small modification from the above. Based on discussion with the bank's loan experts new attribute called payment performance was created that contains excellent, satisfactory and poor category.

The excellent category contains customers whose loans was under pass, the satisfactory category contains customers whose loans under special mention and finally the poor category contains substandard, doubtful and loss since these types of loans currently called non performing in the bank context. The payment performance attribute is dependent attribute or target class for this research.

4.3.3 Data preprocessing

Usually in real world database contains incomplete, noisy and inconsistent data. Such unclean data may cause confusion for the data mining process. Hence data preprocessing is a must in order to increase the quality of the data and to put it into a form that is suitable for use in subsequent mining process [16]. It includes the following activities as listed below.

4.3.3.1. Handling Missing Values

During data entry missing data is common. Hence in order to handle these kinds of problems, the first solution was to recheck the data from manual format. If not possible to get the missing values, there are solutions as suggested below.

The first one is in the case of continuous attributes; the missing values can be replaced with the mean value for that attributes [36]. In this research this approach was applied for attributes number of prior term loans settled and years in business.

This was done based on trade sector based on the bank loan experts' recommendation, first the dataset was grouped and the mean values of the two attributes were taken and used in the missing values.

But based on the discussion with the bank's loan experts it was found that it is very difficult to assign the average value for records whose missing attributes related to finance attributes.

Hence the researcher simply removed these kinds of records (628 in number) from the dataset. The second one is, in the case of categorical attributes. These attributes can be grouped into ordinal, whose values can be meaningfully ordered and nominal whose values were unordered. During handling missing values of the median for ordinal attributes and the modal value of nominal attributes will be taken [36].The attributes whose missing values were handled based on the nominal case were Trade sector, Collateral type and Term of payment as shown in table 4.3.

Attribute	Modal value
Trade sector	Import
Collateral type	Building
Term of payment	Monthly

Table 4. 3: Attributes treated by nominal case for missing values

The attribute whose missing value was handled based on ordinal case was payment performance. As stated the possible values of payment performance were, excellent, satisfactory and poor, so the median value is satisfactory which was taken as replacement of the missing value. The distribution of records with respect to payment performance after handling missing values is given in the table 4.4.The sample dataset is attached as annex 6.

Payment performance	Number of records
Excellent	2929
Satisfactory	1074
Poor	1003
Total	5006

Table 4. 4: Distribution of records with respect to payment performance

4.3.3.2. Data Summarization

Data summarization needed when an attribute has certain values which are different in expression but have the same meaning. In this research this condition was observed in the case of Trade sector and collateral type. Hence by discussion with the bank's loan experts the following summarized categories were taken for the two attributes.

Trade sector

- Building&Construction
- Domestic trade service
- Export
- Import
- Health service
- Hotel&tourism
- Manufacturing
- Project
- Transportation

Collateral type

- Vehicle
- Share certificate
- Machine
- Building&vehicle
- Building
- Cash

- Cash&Cash substitutes
- Local bank guarantee
- Foreign bank guarantee
- Vehicle&machine
- Personal guarantee

4.3.3.3. Inconsistent Data Handling

When information on the same topic is collected from different sources, the various sources often represent the same data in different ways [4]. Some branch of the bank uses different kinds of data encoding mechanism during registration. This condition was observed in this research on two attributes i.e. term of payment and collateral type.

For instance some branches represent the value of the term of payment attribute as bimonthly and others as B. But the two values have the same meaning in the bank context. Considering the above cases the following encoding mechanism was adopted as shown in table 4.5.

Category	Description
B	Bimonthly
M	Monthly
Q	Quarterly
H	Half year
Y	Yearly

Table 4. 5: Inconsistent data handling for term of payment

For the case of collateral type, some branches for example encode loan against vehicle as truck, heavy truck etc. Hence the following encoding mechanism was adopted as shown in table 4.6.

Category	Description	Category	Description
BLD	Loan secured against Building	FB	Loan secured against Foreign bank
PG	Loan secured against Personal guarantee	SH	Loan secured against Share certificate
VH	Loan secured against Vehicle	VM	Loan secured against Vehicle Machine
MAC	Loan secured against Machine	LB	Loan secured against Local bank guarantee
BV	Loan secured against Building & Vehicle	CC	Loan secured against Cash&Cash substitute
CA	Loan secured against Cash		

Table 4. 6: Inconsistent data handling for collateral type

4.3.3.4. Deriving New Attributes from Existing Attributes

Adding new attributes that represent relationships in the data are likely to be important in increasing the chance of knowledge discovery process and will yield useful result [4]. Hence according to the findings of Askale Worku [3], Mertework Shawel [26] and the bank's loan experts the following attributes were derived.

- Duration of the term loan in number of days= (Maturity date)-(Booking date)
- Yearly payment=(Granted amount*365)/(Duration of the loan in number of days)
- Networking capital= (Current Asset)-(Current Liability)
- Current Ratio=(Current asset)/(Current liability)
- TL/TA=(Total liability)/(Total asset)
- TL/C =(Total liability)/(Capital)
- CV/ GA=(Collateral value)/(Granted amount)
- Years in business=[2008-(Business establishment year)]

Finally the total attributes became twenty three as it is given in table 4.7.The full descriptions of the attributes attached as annex 2.

No	Attribute Name	Source
1	Trade sector	Loan Approval form
2	Number of prior term loans Settled	Loan Approval form
3	Duration of the term loan in number of days	Derived
4	Term of payment	The Monthly Loans and Advances form
5	Granted amount	The Monthly Loans and Advances form
6	Yearly payment	Derived
7	Total Asset	Commercial Credit Report
8	Capital	Commercial Credit Report
9	Current Asset	Commercial Credit Report
10	Current Liability	Commercial Credit Report
11	Networking capital	Derived
12	Total liability	Commercial Credit Report
13	Current Ratio	Derived
14	TL/TA	Derived
15	TL/C	Derived
16	Collateral type	The Monthly Loans and Advances form
17	Collateral Value	The Monthly Loans and Advances form
18	CV/ GA	Derived
19	Years in business	Derived
20	Performance in other types of loans	Loan Approval Form
21	Performance of prior term loans	Commercial Credit Report
22	Credit Relationship with other bank	Loan Approval Form
23	Payment performance	The Monthly Loans and Advances form

Table 4. 7: The initial total attributes

4.4. Modeling

In this stage first by selecting data mining software different experimentations were done by using decision tree and neural network. Each will be discussed below.

4.4.1. Data Mining Software Selection

Even if there are many commercial data mining system software in the market to choose data mining software is appropriate for a specific task. In addition it is important to have a multi dimensional view of the software this includes data types, scalability, functionality, usability and performance [16]. Hence from available data mining software, Weka-3-5 version 3.5 was selected since,

- It can support the two algorithms that were used in this research namely decision tree and neural network.
- It runs on almost any plat form, for this research on MS windows XP operating system.
- It can handle a variety of data source (for this research Ms excel).
- It provides extensive preprocessing methods.
- It is open source.
- The software is familiar with the researcher.

Weka-3-5 software

Weka was developed at the University of Waikato in New Zealand, and the name stands for Waikato Environment for Knowledge Analysis. It is designed so that users can quickly try out existing machine learning methods on new datasets in very flexible ways.

It provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing both the input data and the result of learning.

This has been accomplished by including a wide variety of algorithms for learning different types of concepts, as well as a wide range of preprocessing methods. This diverse and comprehensive set of tools can be invoked through a common interface, making it possible for users to compare different methods and identify those that are most appropriate for the problem at hand.

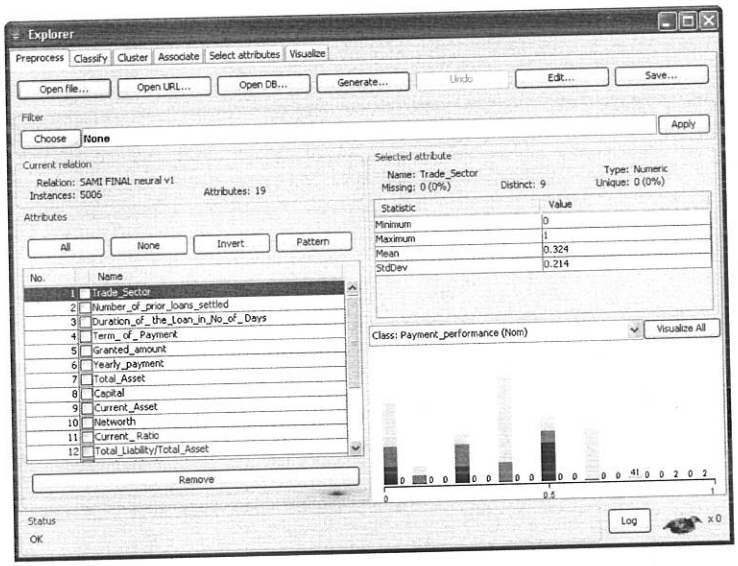


Figure 4.2: Weka's preprocessing interface

The easiest way to use Weka is through a graphical user interface called the Explorer that is shown in figure 4.2. It has six different panels, accessed by the tabs at the top, which correspond to the various data mining tasks supported.

In the "Preprocess" panel, data can be loaded from a file or extracted from a database using an SQL query. The file can be in CSV format, or in the system's native ARFF file format.

Once a dataset has been read, various data preprocessing tools called "filters" can be applied such as Discretization, Normalization, Resampling, Attribute selection, Transforming and Combining attributes.

Through the Explorer's second panel, called "Classify," classification and regression algorithms like J48, Multilayer perceptron can be applied to the preprocessed data. This panel also enables users to evaluate the resulting models; both numerically through statistical estimation and graphically through visualization of the data and examination of the model if the model structure is amenable to visualization. Users can also load and save the models. The third, the fourth and the fifth panel are used for clustering, association and visualizing activities respectively.

Explorer interface helps by presenting choices as menus, by forcing to work in an appropriate order by graying out options until they are applicable, and by presenting options as forms to be filled out by helpful tool tips pop up as the mouse passes over items on the screen to explain what they do. One way of using Weka is to apply a learning method to a dataset and analyze its output to learn more about the data. Another is to use learned models to generate predictions on new instances.

The learning methods are called classifiers, and in the interactive Weka interface you select the one you want from a menu. Many classifiers have tunable parameters, which you access through a property sheet or object editor. Sensible default values ensure that you can obtain results with a minimum of effort but you will have to think about what you are doing to understand what the results mean [30].

4.4.2. Attribute Selection and Model Building Using Decision Tree

Attribute selection is an important step in order to effectively develop a model. Neural network is not efficient at attribute selection since it runs on each possible set of attributes to determine the perfect subset as a result decision tree algorithm would help in identifying the most important attribute [4]. As a result the following processes were done.

4.4.2.1. Data Organization and Preparation

The final processed records for this research is stored in Microsoft Excel file so in order to present these records to Weka software , the file type must be changed into the format that is acceptable for Weka software, as a result the file was saved as a comma delimited (CSV) file type.

4.4.2.2. Attribute Selection and Decision Tree Model Building

In order to select the best attributes and build a decision tree model, the following experiments were done by 22 independent attributes. Here more emphasis was given for combination of attributes that can give the most conceivable rule with accuracy.

Experiment One

The first experiment was done to select the best attributes by using the selection feature of Weka software. Taking the default values of attribute evaluator InfoGainAttributeEval and search method ranker-T-1.7976931348623157E308-N-1, the following attributes were selected as shown in table 4.8.

No	Selected Attribute
1	Duration of the term loan in number of days
2	Term of payment
3	Granted amount
4	Collateral value
5	Years in business

Table 4. 8: Attributes selected by attribute selection feature of Decision Tree

Based on the above selected attributes a decision tree was constructed. The decision tree algorithm used was an implementation of C4.5 decision tree algorithm in weka called J48. Due to the following advantages;

- Choose an attribute that best differentiates the output attribute values.
- Create a separate tree branch for each value of the chosen attribute. Each can originate easily understandable rules.
- It can handle numeric and nominal attributes.
- It can handle training data with missing attribute values [37].

Further, Weka uses the four test mode but the researcher used the following two test modes as stated below that partition the dataset into training and test data in order to test the accuracy.

- **Cross-validation**: - The classifier is evaluated by cross-validation, using the number of folds that are entered in the folds text field.
- **Percentage split**: - The classifier is evaluated on how well it predicts a certain percentage of the data which is held out for testing.

The amount of data held out depends on the value entered in the percentage field. The first one performs independent tests without requiring separate test datasets and without reducing the data used to build the tree. The learning dataset is partitioned into some number of groups called "folds".

It then involves using a single fold from the original sample as the validation data, and the remaining folds as the training data. This is repeated such that each fold in the sample is used once as the validation data and finally the average accuracy will be taken. Experience on a large number of datasets has shown that the number of fold equal to 10 has achieved good test result hence 10 is the recommended [16].

The second one is the method that splits the training data using the percentage provided by user, for example 66%. The 66% of data will be used as training data and the rest 34% of data will be used as validation data to re-train the model and generate evaluation results [16].

In order to classify the records based on their payment performance the model was trained. First the Weka file format saved in CSV file format was opened, the default value of all parameters and a ten fold cross validation mode was taken. Then the J48 program was run and the following result was found as it summarized in the table 4.9.

Actual	**Predicted**			**Total**	**Accuracy**
	Poor	Satisfactory	Excellent		
Poor	926	44	33	1003	92.32%
Satisfactory	9	1053	12	1074	98.04%
Excellent	52	49	2828	2929	96.55%
Total	987	1146	2873	5006	96.03%

Table 4. 9: The confusion matrix of 5 attributes

Even if the model perform well in each category by the discussion with the bank's loan experts the researcher found that the attributes used by the constructed tree were not the only

important attributes that are considered in loan approval procedure and also the rules found from decision tree were not much conceivable.

Experiment Two

After the above experiment the researcher using all 22 independent attributes constructed a decision tree taking all the parameters with the default values (minnumobj=2, confidence factor=0.25) and taking a ten fold cross validation mode. The result of the decision tree is summarized in table 4.10.

Actual	**Predicted**			**Total**	**Accuracy**
	Poor	Satisfactory	Excellent		
Poor	910	59	34	1003	90.73%
Satisfactory	4	1044	26	1074	97.21%
Excellent	63	126	2740	2929	93.58%
Total	977	1229	2800	5006	93.79%

Table 4. 10: The confusion matrix of 22 attributes

As it shown in the table, the model correctly classify the satisfactory with less error than the others category. Although 22 independent attributes were used in the input data for this case, the decision tree selected 16 attributes without including the target class during decision tree construction.

Normally decision tree building algorithms put the attributes that does the best jobs of splitting at the root node of the tree [4]. This is an indication that the attribute not found in the decision tree are not relevant for classification.

Based on the discussion with the bank's loan experts, the researcher found that even if the accuracy was encouraging but the rules generated from this tree were not much conceivable.

Even though some data mining algorithms will automatically ignore irrelevant attributes and properly account for related columns it is advised to avoid depend on a tool because often the knowledge of the problem domain experts may help in making the selection of the attributes correctly [26].

Even if the selected attributes are important in loan approval stage there are still some attributes left , hence by discussion with the bank's loan experts, the researcher add three more attribute which were not used by the decision tree namely "credit relation ship with other banks", "capital" and "performance of prior term loans". The selected attributes are presented in the table 4.11.

No	Attribute	No	Attribute
1	Duration of the term loan in number of days	11	Performance in other types of loans
2	Term of payment	12	TL/TA
3	Granted amount	13	Trade sector
4	Collateral Value	14	TL/C
5	Years in business	15	Credit relation ship with other banks
6	Current Ratio	16	Networking capital
7	Current Asset	17	Capital
8	Collateral type	18	Performance of prior term loans
9	Number of prior term loans Settled	19	Current liability
10	Total Asset		

Table 4. 11: The selected attributes by J48 decision tree

In the work of Mertework Shawel [26] two terms were described, the first one is credit risk in which a bad debtor is categorized as a good one and the second is commercial risk in which a good debtor is categorized as a bad. From the two, the credit risk can put banks at great risk.

Therefore the researcher taking into consideration the credit risk, further experimentation were done by varying the number and combination of the attributes in order to get a tree with conceivable rules and less error in categorizing poor payment performance in this case the 10 fold cross validation was used with other parameters were in default values (minnumobj=2, confidence factor=0.25).

In table 4.12 the results of some of decision tree constructed by using different attributes combination are presented.

Number of attributes used	Accuracy of poor category	Accuracy of satisfactory category	Accuracy of excellent category	Total accuracy
8	86.54%	97.85%	77.80%	83.86%
10	89.33%	94.88%	97.91%	95.54%
14	90.32%	95.81%	97.47%	95.68%
16	86.94%	97.40%	89.24%	90.53%
18	89.23%	97.30%	91.63%	92.37%

Table 4. 12: The results found by using different attributes combination

4.4.2.3. Attribute Selection Result

From the above table by discussion with the bank's loan experts the decision tree constructed using 14 attributes was selected based on the conceivability of the rules that it generates and the performance in poor category i.e. 90.32 %(attached as annex 1).

Actual	Predicted			Total	Accuracy
	Poor	Satisfactory	Excellent		
Poor	906	55	42	1003	90.32%
Satisfactory	10	1029	35	1074	95.81%
Excellent	44	30	2855	2929	97.47%
Total	960	1114	2932	5006	95.68%

Table 4. 13: The confusion matrix for the 14 attributes.

As shown in table 4.13, the overall accuracy of this learning scheme was 95.68% which indicates that out of 5006 records supplied, 4790 records were classified correctly while the remaining 216 records were classified incorrectly.

Based on the result of the experiment 97.47% of the records in the excellent category were correctly classified while 95.81% and 90.32% of records in the satisfactory and poor category respectively were classified correctly. Further the performance of the model in poor and excellent categories were best than the others.

In order to improve the accuracy of the selected decision tree further experiments were done by changing the mode and adjusting some parameters. For examples by using the percentage split mode with 66%, 80%, and 70%, three models were constructed taking the default values of other parameters (minnumobj=2, confidence factor=0.25).

The main intention was to check the performance of the models by varying the proportion of training and testing set. The result of the models is given in table 4.14.

percentage split	Accuracy of poor category	Accuracy of satisfactory category	Accuracy of excellent category	Total accuracy
66%	87.57%	95.91%	95.29%	93.89%
70%	90%	96.69%	95.222%	94.47%
80%	89.16%	94.74%	96.6%	94.71%

Table 4. 14: The results found by using different percentage split value

As shown in the table 4.14, even if the over all accuracy for each models were encouraging particularly for 70%, but the generated rules were not much conceivable. The researcher also constructed different decision tree by adjusting the confidence value.

As observed when the confidence factor value increases there is little increment in accuracy, but the size of the tree and number of leaves increase as a result it became very difficult to analyze but when confidence factor decreased the overall accuracy and the size also decreased.

The researcher also tried to make the rule generated by decision tree more understandable by adjusting the parameter minnumobj (minimum number of instance in a leaf) and it was observed that as the minnumobj parameter increased from the default value 2, the size of the tree and number of lives decreased and also the total accuracy decreased. Further the researcher tried to construct a tree by discretize of the values of some attributes but the overall performance was decreased.

Even though the mode and parameters adjustment was done to get best decision tree it was not possible to get a tree with a conceivable rule and accuracy of more than 95.68%, with 10

fold cross validation, minnumobj=2, confidence factor=0.25. Hence this model with its attributes was selected as a best model.

4.4.2.4. Generating rules from Decision tree

From the decision tree constructed it is possible to find out a set of rules for each path from the root to a leaf node. The following are some of the rules extracted from the decision tree as it shown in the figure below.

Rule1) If Networking capital <= -127027.5521: then Poor (182.0).

Rule2) If Networking capital > -127027.5521 and Networking capital <= -49266.158 and Current_ Ratio <= 0.280233 and Current_ Ratio > 0.117483: then Poor (54.0).

Rule3) If Networking capital<= -49266.158 and Networking capital > -127027.5521 and Current_ Ratio > 0.572 and Total_Asset > 565384.71: then Satisfactory (33.0/2.0).

Rule4) If Networking capital > -49266.158 and Number_of_prior_term loans_settled <= 4 and Current_ Ratio <= 0.237949 and TL/TA <= 1.860128 and TL/TA > 0.614545: then Excellent (32.0).

Rule5) If Networking capital > -49266.158 and Number_of_prior_term loans_settled <= 4 and Current_ Ratio <= 0.488802 and Current_ Ratio > 0.237949 and Total_Asset > 16692595.15 and Current_Liability <= 3939456.608: then Poor (25.0/1.0).

Rule6) If Networking capital > -49266.158 and Number_of_prior_term loans_settled <= 4 and Current_ Ratio > 0.488802 and Current_ Ratio <= 2.333333 and Collateral_Value <= 466618.5 and YearsinBusiness <= 3: then Excellent (67.0).

Rule7) If Networking capital > -49266.158 and Number_of_prior_term loans_settled <= 4 and Current_ Ratio > 0.488802 and Current_ Ratio <= 2.333333 and Current_ Ratio > 0.488802 and Collateral_Value <= 466618.5 and YearsinBusiness > 3 and Number_of_prior_term loans_settled <= 1.926452: then Poor (25.0).

Rule8) If Networking capital > -49266.158 and Number_of_prior_term loans_settled > 4 and YearsinBusiness > 12 and Performance_ of _Prior term loans = excellent: then Excellent (137.0/2.0).

Rule9) If Networking capital > -49266.158 and Number_of_prior_term loans_settled <= 4 Current_ Ratio > 0.488802 and Current_ Ratio <= 2.333333 and Collateral_Value > 635814 and Credit_ Relationship_with_other_bank = Excellent and Trade_Sector = Domestic trade service and Collateral_Value <= 989925: then Excellent (20.0/1.0)

Rule10) If Networking capital > -49266.158 and Number_of_prior_term loans_settled <= 4 Current_ Ratio > 0.488802 and Collateral_Value<=2452660 and Collateral_Value>466618.5 and Collateral_Type=BLD and Credit_Relationship with_other_bank=Excellent and Current_Asset<=120879: then Excellent (8.0).

Rule11) If Networking capital > -49266.158 and Number_of_prior_term loans_settled <= 4 Current_ Ratio > 0.488802 and Collateral_Value>5484778.638 and Collateral_Value<=8685242.56 and Credit_Relationship with_other_bank=Excellent and Performance_in other types_of_term loans=none: then Satisfactory (139.0/19.0).

The rules indicated above used the possible condition in which a specific loan customer classified in each class. The amount in the "if" part indicates the values of each attributes and the "then" part indicates the class label of the customer and the number in a bracket indicates the numbers of records classified correctly and wrongly in each class.

For instance in the rule 3,

If Networking capital greater than -127027.5521 and less than or equal to –49266.158 and if current ratio>0.572 and if total asset > 565384.71 then the payment performance is satisfactory .The model classify correctly 33 records and classify wrongly 2 records with the given criteria.

Moreover the rules generated by decision tree indicated that the attributes such as Networking capital, Current Ratio, Total Asset, TL/TA, Current Liability, Collateral Value, Years in Business, Number of prior term loans settled, Performance of Prior term loans, Collateral Type, Credit Relationship with other bank, Trade Sector, Performance in other types of Loans and Current Asset are the basis for classfication of payment performance.

4.4.3. Neural Network Model Building

In this stage data organization for neural network and then experimentation were done. Each will be discussed below.

4.4.3.1. Data Organization and Preparation

The Neural Network was employed in this research based on the attributes selected by the best decision tree model. The researcher used weka implementation of back propagation algorithm called Multilayer perceptron as stated earlier.

Before start to build the model it was need to organize the data into a form suitable for model building. Neural networks accept values only when the values of attributes are numeric within the range of 0 to 1 but the target class value could be nominal [37].But weka software version 3.5 has a feature that can normalize the input attributes, therefore no need to normalize manually.

The data set in a CSV file format was prepared similar to the case of decision tree algorithm. Before start the training, the researcher considered the network architectures i.e. the number of node in the hidden layer, the learning rate and the number of training time (epochs).

For the number of nodes in the hidden layer, learning rate and the number of training time, the developer of Weka software suggest that default number of nodes in the hidden layer is the average of the number of input and output attributes but it can possible to create a better trained network by adjusting up and down the number of nodes in the hidden layer.

For learning rate, lower learning rate require more training iterations and higher learning rate allows the network to converge more rapidly, however the chance of non-optimal solutions are greater.

Concerning the number of training time, it is the total number of times the entire set of training data will pass through the network structure. Therefore increasing this number will likely improve the accuracy of the model, but at the cost of time, and decreasing this number will likely decrease the accuracy, but take less time.

Hence based on the above suggestions first the researcher did three experiments using 10 fold cross validation and then one experiment by varying the percentage split mode.

Experiment One

The first experiment was done by varying the learning rate of the model keeping other parameters in their default values (10 fold cross validation mode, training time=500 and number of node in the hidden layer=9). Various models were built and some of the best results obtained from the experiments are shown in the Table 4.15.

Learning rate	Accuracy of poor category	Aecuracy of satisfactory category	Accuracy of excellent category	Total accuracy
0.1	73.88%	78.40%	93.65%	86.42%
0.2	88.14%	88.45%	94.71%	92.05%
0.3	88.14%	87.15%	94.78%	91.81%
0.4	87.54%	84.73%	94.84%	91.21%
0.5	87.14%	86.78%	94.5%	91.37%

Table 4. 15: The results found by varying learning rate

As shown in the table the overall performance of the models is good particularly the model with learning rate 0.2, since the overall performance and the performance in the poor and excellent category were best. But the performance in satisfactory was moderate performance. The next experiment was conducted based on the numbers of neurons.

Experiment two

This experiment was done by varying the number of the neurons in the hidden layer keeping the other parameters in their default values to see the contribution of number of neurons in the model performance (10 fold cross validation mode, learning rate=0.3 and training time=500). Various models were built and some of the best results obtained from the experiments are shown in table 4.16.

Number of neurons in hidden layer	Accuracy of poor category	Accuracy of satisfactory category	Accuracy of excellent category	Total accuracy
6	79.46%	78.50%	93.82%	87.65%
9	88.14%	87.15%	94.78%	91.81%
12	85.14%	85.30%	94.33%	90.55%
14	86.44%	84.92%	94.57%	90.87%
16	88.53%	86.50%	95.00%	91.75%

Table 4. 16: The results found by varying the number of neurons

As shown in the table the overall performance of the models is good specially the models with number of neurons 16 and 9.

The model with 16 neurons is better in poor and excellent category, but in the case for the overall performance and satisfactory category, the model with 9 neurons is better. The next experiment was conducted based on number of epochs.

Experiment three

The third experiment was done by varying the number of training time keeping the other parameters in their default values (10 fold cross validation mode, learning rate=0.3, and number of node in the hidden layer=9). At this case also various models were built and some of the best results obtained from the experiments are shown in table 4.17.

Number of training time	Accuracy of poor category	Accuracy of satisfactory category	Accuracy of excellent category	Total accuracy
200	87.14%	85.50%	94.64%	91.17%
300	87.04%	86.22%	94.74%	91.81%
400	87.94%	85.85%	94.84%	91.53%
500	88.14%	87.15%	94.78%	91.81%
650	88.73%	87.52%	95.05%	92.20%
900	88.04%	87.43%	95.40%	92.20%
1500	89.23%	88.83%	95.52%	92.83%

Table 4. 17: The results found by varying number of training time

As it shown in the table 417, the model with number of training time 1500 has shown best performance than the other. The next experiment was conducted based on percentage split mode.

Experiment four

Keeping the default value of the other parameters (learning rate=0.3, training time=500 and number of node in the hidden layer=9) the researcher tried to construct models by varying the percentage split mode. By varying the proportion of training and testing set the following results were found as it shown in table 4.18.

percentage split	Accuracy of poor category	Accuracy of satisfactory category	Accuracy of excellent category	Total accuracy
66%	85.21%	86.88%	94.41%	91.07%
70%	87.7%	80.10%	90.11%	87.75%
80%	85.71%	89.95%	91.34%	89.91%

Table 4. 18: The results found by varying the percentage split mode

Here the constructed models perform best once in each category. For example in the case of 66% percentage split, it has best performance at excellent category, for the case of 70% percentage split, it has best performance at poor category and for the case of 80% percentage split, it has best performance at satisfactory. But the overall performance of 66% percentage split was best of all constructed models in percentage split mood in the case of neural network.

4.4.3.2. Selection and Interpretation of Neural Network

From the above four tables, one can see that over all performance of the models was in general encouraging. Most of the models have best accuracy for excellent category and next to poor category. Three models were selected from experiment one whose learning rate was 0.2 and one from experiment two whose number of neurons was 16 and one from experiment

three whose number of training time was 1500 respectively based on their performance in poor category. Below here the confusion matrixes of the selected models are given.

Actual	Predicted			Total	Accuracy
	Poor	Satisfactory	Excellent		
Poor	884	71	48	1003	88.14%
Satisfactory	59	950	65	1074	88.45%
Excellent	53	102	2774	2929	94.71%
Total	996	1123	2887	5006	92.05%

Table 4. 19: The confusion matrix of a model with learning rate 0.2

Actual	Predicted			Total	Accuracy
	Poor	Satisfactory	Excellent		
Poor	880	69	46	1003	88.73%
Satisfactory	61	940	84	1074	87.52%
Excellent	45	101	2784	2929	95.05%
Total	986	1110	2914	5006	92.20%

Table 4. 20: The confusion matrix of a model with number of neurons 16

Actual	Predicted			Total	Accuracy
	Poor	Satisfactory	Excellent		
Poor	895	65	43	1003	89.23%
Satisfactory	58	954	62	1074	88.83%
Excellent	54	77	2798	2929	95.52%
Total	1007	1096	2903	5006	92.83%

Table 4. 21: The confusion matrix of a model with number of training time 1500

Even if, accuracy can be used to select the best model, it is not usually describe every detail. Here by discussion with the bank's loan experts, the researcher applied the credit risk condition in order to select the best model.

The model from experiment three whose training time was 1500 was selected as best model of all the constructed neural network models. Since it maximize the accuracy in predicting poor payment performance and minimizing in predicting poor as satisfactory or as excellent from others as it shown in its confusion matrix.

4.5. Evaluation of Decision Tree and Neural Network

The first modeling used was decision tree. In this part attributes were selected from the decision tree constructed and some more attributes were added by the discussion with the bank's loan experts. After that different experiment were done in order to select the best attributes combination that can give conceivable rules and perform well.

Hence the model was selected as best with conceivable rules and fourteen attributes. Then the attributes selected by this decision tree were given to the neural network and trained iteratively in order to build different models on the same datasets. The model that showed the best performance of all the neural network models was selected.

When the performance of each technique for each category considered, the performance of the neural network was encouraging but less than the decision tree, it was 89.23% to 90.32% for poor category,88.83% to 95.81% for satisfactory,95.52% to 97.47% for excellent and 92.83% to 95.65% for overall.

The neural network showed best performance at excellent category than the other categories also the same was true for decision tree. On the other hand decision tree produced a set of rules that can understand by any user in order to make a decision but in the case of neural network it did not show any rule or reason about the results.

4.6. Deploy the model

The creation of the model is not the end of the work; hence the trained model can be used to perform classification and prediction on data in real time using the acquired knowledge. In order to deploy the model it must be organized and presented by discussion with the bank's loan experts in a way that it can be manageable.

Even if there is a need to have further research to get more accurate model, the result of this research can be used to support decision making in loan approval procedure at the bank. No matter how well the model is designed and tested, it is just a model that was built from a set of sample dataset. Therefore there is a possibility to failure; hence the performance of the model needs to be evaluated in regular basis.

Chapter Five

Conclusion and Recommendation

5.1. Conclusion

The granting of loans by a bank is one of the most important activities that require delicate care. The institution usually employs loan experts to make credit decisions or recommendations. These experts are given some hard rules to guide them in evaluating the worthiness of loan applications.

After some period of time, the experts gain their own observed knowledge or intuition in deciding whether an application is loan worthy or not. But as the volume of the data to be examined is increased, the nature of the relationships themselves becomes hidden and complicated. Therefore it is difficult to evaluate.

On the other hand the data routinely collected in bank's myriad master files in normal course of business, represents available asset and it is useful for assisting the decision making process. But the lack of technology that can manage the huge amount of data banks faced with intense computations and rising in loan default rates.

Currently most banks are looking for a way to effectively manage and leverage these data asset to achieve a competitive advantage. Hence the rapidly emerging technology called data mining can be used to unlock the intelligence hidden in this huge collection of data.

In this research an attempt was made to assess the possible application of data mining technology in support of term loan risk assessment at United Bank Share Company.

This research was done based on the recommendation of the previous researchers Askale Worku [3] and Mertework Shawel [26].

The researcher followed basically the following methodologies; literature review, business understanding, data collection, data preparation, training and building model and performance evaluation. In the data collection and it was so difficult to collect the data from different branch since the data was found in manual format and the target classes of some customers are not found.

In data preparation stage, even if majority of the data was pre classified based on the National Bank of Ethiopia but later by discussion with the bank's loan experts some modification was applied on target class. Concerning the distribution of the target class, due to luck of enough data there was not proportional distribution on the values of target class.

The two basic tasks done in model building were, attribute selection then model building using decision tree and then based on the selected attributes different neural network models were built. In the attribute selection part, a decision tree was selected that gave meaningful rules; this decision tree used 14 attributes from 22 independent attributes with accuracy of 95.65%.

The selected attributes by decision tree were; Networking capital, Current Ratio, Total Asset, TL/TA, Current Liability, Collateral Value, Years in Business, Number of prior term loans settled, Performance Of Prior Term Loans, Collateral Type, Credit Relationship With other Bank, Trade Sector, Performance in Other Types of Loans and Current Asset.

Previously from the survey it is found out that the bank give more emphasis to Capital, Current Liability, Collateral Value, Performance of prior term loans, Collateral type and Credit Relationship with other Bank. But from this research it was found that additional attributes such as Networking capital, Current ratio, Total asset, TL/TA, Years In Business, Number Of Prior Term Loans Settled, Trade Sector, Performance In Other Types Of Loans and Current Asset must be given more emphasis to grant or reject the term loan request.

Then the selected attributes were given to neural network for further model building. Different neural network models were generated and the model with accuracy of 92.83% was selected. When comparison was done, even if both techniques did well but the overall performance of decision tree is better than neural network.

Even though both decision tree and neural network can model data that has nonlinear relationships between attributes and both can handle interactions between attributes but neural network is more of "black box" that delivers results without an explanation of how the results were derived. But in the case of decision tree it is easy to get information on how decisions were made so that it is useful for decision making agent in order to understand the basic nature of the data being analyzed.

The possible misclassification by each model in each category can be the luck of enough training sample for each class label category. As observed, the proportion of excellent category is the largest of all the categories and each model perform well in this category compared to the other.

To sum up even if further research is needed, the result from this research showed that based on the selected attributes loan risk assessment particularly term loan can be greatly supported by application of data mining techniques specially using decision tree.

5.2. Recommendation

This research showed the possible application of data mining technology in support of the term loan risk assessment in the case of United Bank Share Company.

As result the bank can control the rate of defaulter. Below here the researcher would like to suggest future works based on the findings of this research in supporting loan risk assessment in United Bank S.C. or some other bank.

- A data mining task could be efficient if the data warehouse is available since the history of credit data can easily found in electronic format.
- Data mining techniques could contribute a lot in identifying payment performance of the customer therefore it could be more important to apply a system that uses data mining techniques as a tool for loan approval procedure.
- At this research only a limited number of attributes were considered but attention must be given to attributes specially that are more related to financial information but not tested before.
- Further investigation on banks like Commercial Bank of Ethiopia is needed since it can create an opportunity to get large data set so that more accurate data mining result can be found.
- This research was done making the target class to have three categories but there is a need to take more detailed classification of loan customer, such as pass, special mention, substandard, doubtful and loss in order to exactly to classify the prospective loan customer and to take appropriate action.
- If possible use different data mining soft wares at the same time on the same dataset and compare which software is more appropriate in mining more hidden pattern.

References

1. Abrahams, A. Hathout,F. Staubli,A. and Padmanabhan,B.(1999), Profit-Optimal Model and Target Size Selection with Variable Marginal Costs, Department of Operations and Information Management, The Wharton School, University of Pennsylvania and University of Cambridge Computer Laboratory, Cambridge, United Kingdom. Available at URL: http://www.cl.cam.ac.uk/research/srg/opera/publications/papers/abrahams-wharton-working-paper-feb04.pdf. Visited on November, 14, 2008.
2. AL-Attar, A. (1999), Data Mining-Beyond Algorithms, Xpertrule Software Ltd, UK. Available at URL:http://www.attar.com/tutor/mining.html. Visited on August, 6, 2008.
3. Askale Worku(2001), *Data mining Application in Support of Loans Disbursement Activity at Dashen Bank S.C.*,Unpublished Master's thesis, Addis Ababa University, Department Of Information Science.
4. Berry, M.J.A and Linoff, G. (2004), *Mastering Data Mining: the Art and Science of Customer Relation Management*, 2nd edition, John Wiley & Sons, Inc, Indianapolis, Indiana.
5. Bigus, J.P. (1996), *Data mining With Neural Network: Solving Business Problems from Application to Development to Decision Support*, McGraw-Hill. New York.
6. Bill, P. (2005), Data mining. Available at URL:http://www.anderson.ucla.edu/faculty/jason.frand:teacher/technologies/palace/index.htm. Visited on August, 14, 2008.
7. Bounsaythip, C. and Rinta-Runsala, E. (2001), Overview of Data mining for customer Behavior Modeling, VTT information Technology, Finland. Available at URL:Http://www.Vtt.fi/tte. Visited on August, 14, 2008.
8. Cheng, B.and D.Titterington, "*Neural Networks: A Review from a Statistical Perspective*", Statistical Science, Vol.9, PP 2-54, 1994.
9. Dass, R. (2006), Data Mining in Banking and Finance, Indian Institute of Management Ahmedabad, India. Available at URL:http://www.iimahd.ernet.in/publications/data/Note on Data Mining & BI in Banking Sector.pdf, Visited on October, 13, 2008.

10. Fabris, P. Advanced navigation: Marketing secrets from the financial sector show how data mining charts a profitable course to customer management, CIO magazine, 11, No.15, 1998, p. 50--55. Goetghebeur, France.

11. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P (1996),"From Data mining to knowledge Discovery: An overview." *In Advance in knowledge discover and data mining*, Fayyad Et al (Eds) MIT press. Available at URL:http://citeseer.nj.nec.com:fayyad96from.html. Visited on September, 17, 2008.

12. Federal Reserve Bank of Chicago (2002), Risk Management, Chicago, Illinois 60604-1413, USA. Available at URL http://www.chicagofed.org/banking_information/risk_management.cfm. Visited on September, 14, 2008.

13. Forchilich, H. (1999), Neural Net Overview, Eberhard-Karls-Universität Tübingen ,Wilhelm-Schickard-Institut für Informatik ,Lehrstuhl Rechnerarchitektur, Sand 1, D - 72076 Tübingen, Germany .Available at URL:http://rfhs8012.fh-regenburg.de/~saj39122/jfroehl/diplom/e-text.html. Visited on October, 14, 2008.

14. Giudici, P. (2005), *Applied Data mining, Statistical methods for business and industry*, John Wily & Sons LTD, faculty of economics, university of Pavia, Italy.

15. Goebel, M.and Gruenwald, L. (1998), A Survey of Data mining and Knowledge Discovery and Data mining Tools, Technical Report, University of Oklahoma, School of Computer Science, USA. Available at URL:http://www.acm.org/sigkdd/exploration/issue1-1/survey.pdf. Visited on August, 14, 2008.

16. Han, J. and Kamber, M. (2004), *Data mining Concepts and Techniques*, Morgan Kufman Publishers, San Francisco.

17. Heyuguo,Y,H.(1997), Knowledge Reduction and Discovery Based On Demarcation Information, Department of Computer Science, Beijing Institute of Technology, Beijing, P.R.China. Available at URL: http:// www.citebase.org/fulltext?format=application/pdf&identifier=oaiarXiv.org:cs/0405104. Visited on October, 10, 2008.

18. Juneja, m. and Phull, N. (2008), Data mining and its scope, swami vivekanand institute of engineering & technology, India. Available at URL: http://www.rimtengg.com/coit2007/proceedings/pdfs/109.pdf. Visited on June 15, 2008.

19. Koh Hian,C .Gerry,L and Kim,C.(2002), Data mining and customer relationship marketing in the banking industry, Goliath Business, Singapore. Available at URL:http://findarticles.com/p/articles/mi_qa5321/is_200201/ai_n21318270/pg_5. Visited on August, 24, 2008.
20. Koh Hian,C .Gerry,L and Kim,C.(2002), Data mining and customer relationship marketing in the banking industry, Goliath Business, Singapore. Available at URL http://goliath.ecnext.com/coms2/gi_0199-1871148/Data-mining-and-customer-relationship.html.Visited on September, 14, 2008.
21. Kononenko, I.and Hong, S.J. (1997),Attribute Selection for Modeling, University of Ljubljana, Faculty of computer and information science, Slovenia. Available at URL:http:/www.research.ibm.com/dar:papers/pdf/gcshong_with_cover.pdf. Visited on October, 4, 2008.
22. Kramer, B. (1994),The evaluation of Dutch Non-life Insurance Companies: A comparison of an Order Logit and a Neural Network Model, Faculty of Management and Organization, University of Groningen, The Netherlands. Available at URL: http://www.Ub.rug.nl/eldoc/som/95A20/95A20.pdf. Visited on September, 24, 2008.
23. Krzysztof, H., Mieczyslaw, L.and Maciej, P. (2003), Building Data Mining Models in the Oracle 9i Environment, Wroclaw University of Economics, Poland. Available at URL Http://informingscience.org/proceedings/IS2003Proceedings/docs/146Hauke.pdf. Visited on September, 14, 2008.
24. Langley, P. and Simon, H.A. (1995), Applications of Machine Learning and Rule Induction, Langely applications, Commun.Acm, kEG, Tsinghua. Available at URL:http://citeseer.nj.nec.com/langely95applications.html. Visited on November, 4, 2008.
25. Lawrence, J. (1994), *Introduction to Neural Networks, Design, Theory and Application,* California Scientific Software Press, Nevada City:
26. Mertework Shawel (2001), *Possible Application of Data Mining Technology in Supporting Credit Risk Assessment: The Case of NIB International Bank S.C.,* Unpublished Master's thesis, Addis Ababa University, Department Of Information Science.
27. Mitchell, M. (1997), *Machine learning*, The McGraw-Hill Companies Inc, New York.

28. Peter, C, Julian, C, Randy, K, Thomas, K, Thomas, R, Colin, S, and Rudiger, W. (2000), CRISP–DM Step-by-Step Data Mining Guide, CRISP–DM consortium, NCR system engineering Copenhagen (USA and Denmark). Available at URL:http://www.spss.fi:pdf/crisp-dm.pdf.Visited on September, 14, 2008.(substitute for no 10).
29. Piatetsky-Shapiro, G. (1996), The Data Mining Industry Coming of Age, IEEE Intelligent System, university of Minnesota,USA. Available at URL:www.kdnuggets.com/gpspubs/ieee-intelligent-dec-1999-x6032.pdf. Visited on July 24, 2008.
30. Pirooznia, M., Y Yang, J., Qu Yang, M. and Deng, Y. (2008), A comparative study of different machine learning methods on micro array gene expression data. Available at URL:http://www.pubmedcentral.nih.gov/articlerender.fcgi. Visited on July 24, 2008.
31. Plate, T, Bert, J Grace, J and Band, P (1997), "A comparison between neural networks and other statistical techniques for modeling the relationship between tobacco and alcohol and cancer", in *Advances in Neural Information Processing 9 (NIPS*96)*, M.C. Mozer, M.I. Jordan and T. Petsche, Eds, pp967-973, MIT Press.
32. Stergiou, C. (1996), Neural Network, by Christos Stergiou and Dimitrios Siganos, USA. Available at URL:http://www.doc.ic.uk/~nd/surprise_96/journal/vol1/csl1/article.html. Visited on August 29, 2008.
33. Tesfaye Hintsay (2002), *Predictive Modeling Using Data Mining Techniques In Support Of Insurance Risk Assessment,* published Master's thesis, Addis Ababa University, Department Of Information Science.
34. Thearling, K (2003), An introduction to data mining, Kurt Thearling, USA. Available at URL: http:/ www3.shore.net/~kht/text/dmwhite.htm.visited on June 15, 2008.
35. Thearling, K. (2005), An introduction to data mining; Discovery hidden value in your data warehouse, Kurt Thearling, USA. Available at URL:http://database.about.com:gi:dynamic:off site.htm. Visited on November, 1, 2008.
36. Two Crows Corporation (1999), Introduction to Data mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, U.S.A. Available at URL:http://www.twocrows.com. Visited on August 2, 2008.

37. Witten,I.H and Frank,E(2005),*Data mining practical machine learning tools and techniques*, 2nd ed, Morgan Kaufmann series in data management system, Elsevier Inc, San Francisco.

Glossary of Terms

Asset: Anything of value that is owned.

Balance Sheet: The statement showing what is owned, what is owed and what the business is worth on a specific date.

Booking Date: The date, that, the loan is granted.

Capital: What the business is worth. It is the total asset minus total 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.

Collateral: Asset that is pledged or mortgaged to secure a loan, thereby reducing risk to the lender.

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

Liabilities: Amounts that are owed to creditors.

Maturity date: The expiry date of the loan's period.

Networking capital: Networking capital is a measurement of an enterprise to meet its short-term debt with its current asset. It is Current assets minus current liabilities.

Appendices

Annex 1: Decision Tree constructed using 14 attributes

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

Networking capital <= -49266.158
|   Networking capital <= -127027.5521: Poor (182.0)
|   Networking capital > -127027.5521
|   |   Current_ Ratio <= 0.280233
|   |   |   Current_ Ratio <= 0.117483: Satisfactory (23.0/2.0)
|   |   |   Current_ Ratio > 0.117483: Poor (54.0)
|   |   Current_ Ratio > 0.280233
|   |   |   Trade_Sector = Domestic trade service: Satisfactory (25.0/4.0)
|   |   |   Trade_Sector = Manufacturing: Satisfactory (22.0/1.0)
|   |   |   Trade_Sector = Transportation: Satisfactory (6.0)
|   |   |   Trade_Sector = Import: Satisfactory (49.0/1.0)
|   |   |   Trade_Sector = Building&Construction
|   |   |   |   Total_Asset <= 528933.7486: Poor (9.0)
|   |   |   |   Total_Asset > 528933.7486: Satisfactory (9.0)
|   |   |   Trade_Sector = Export: Satisfactory (0.0)
|   |   |   Trade_Sector = Hotel&tourism: Satisfactory (0.0)
|   |   |   Trade_Sector = Health service: Satisfactory (0.0)
|   |   |   Trade_Sector = Project: Satisfactory (0.0)
Networking capital > -49266.158
|   Number_of_prior_termloans settled <= 4
|   |   Current_ Ratio <= 0.488802
|   |   |   TL/TA <= 1.860128
|   |   |   |   Current_ Ratio <= 0.237949
|   |   |   |   |   TL/TA <= 0.614545
|   |   |   |   |   |   Total_Asset <= 13750000: Satisfactory (23.0)
|   |   |   |   |   |   Total_Asset > 13750000: Poor (2.0)
|   |   |   |   |   TL/TA > 0.614545: Excellent (32.0)
|   |   |   |   Current_ Ratio > 0.237949
|   |   |   |   |   TL/TA <= 1.587172
|   |   |   |   |   |   Total_Asset <= 16692595.15
|   |   |   |   |   |   |   TL/TA <= 1.3
|   |   |   |   |   |   |   |   Trade_Sector = Domestic trade service
|   |   |   |   |   |   |   |   |   Total_Asset <= 199000: Poor (6.0)
|   |   |   |   |   |   |   |   |   Total_Asset > 199000: Excellent (17.0)
|   |   |   |   |   |   |   |   Trade_Sector = Manufacturing: Excellent
(5.0)
|   |   |   |   |   |   |   |   Trade_Sector = Transportation: Excellent
(8.0/1.0)
|   |   |   |   |   |   |   |   Trade_Sector = Import: Excellent (9.0)
|   |   |   |   |   |   |   |   Trade_Sector = Building&Construction
|   |   |   |   |   |   |   |   |   Current_ Ratio <= 0.325472: Excellent
(4.0)
|   |   |   |   |   |   |   |   |   Current_ Ratio > 0.325472: Poor (11.0)
|   |   |   |   |   |   |   |   Trade_Sector = Export: Excellent (0.0)
|   |   |   |   |   |   |   |   Trade_Sector = Hotel&tourism: Excellent
(0.0)
|   |   |   |   |   |   |   |   Trade_Sector = Health service: Excellent
(0.0)
|   |   |   |   |   |   |   |   Trade_Sector = Project: Excellent (0.0)
|   |   |   |   |   |   |   TL/TA > 1.3: Excellent (59.0)
```

```
|   |   |   |   |   |   Total_Asset > 16692595.15
|   |   |   |   |   |   |   Current_ Ratio <= 0.360564
|   |   |   |   |   |   |   |   Current_Asset <= 10952271.55: Poor
(36.0/3.0)
|   |   |   |   |   |   |   |   Current_Asset > 10952271.55: Excellent
(4.0)
|   |   |   |   |   |   |   Current_ Ratio > 0.360564
|   |   |   |   |   |   |   |   Trade_Sector = Domestic trade service:
Excellent (4.0)
|   |   |   |   |   |   |   |   Trade_Sector = Manufacturing: Excellent
(0.0)
|   |   |   |   |   |   |   |   Trade_Sector = Transportation: Satisfactory
(2.0)
|   |   |   |   |   |   |   |   Trade_Sector = Import: Excellent (0.0)
|   |   |   |   |   |   |   |   Trade_Sector = Building&Construction:
Excellent (0.0)
|   |   |   |   |   |   |   |   Trade_Sector = Export: Excellent (0.0)
|   |   |   |   |   |   |   |   Trade_Sector = Hotel&tourism: Excellent
(0.0)
|   |   |   |   |   |   |   |   Trade_Sector = Health service: Excellent
(0.0)
|   |   |   |   |   |   |   |   Trade_Sector = Project: Excellent (0.0)
|   |   |   |   |   TL/TA > 1.587172
|   |   |   |   |   |   Performance_ of _Prior_termLoans = Poor: Poor (0.0)
|   |   |   |   |   |   Performance_ of _Prior_term Loans = Excellent:
Excellent (3.0/1.0)
|   |   |   |   |   |   Performance_ of _Prior_term Loans = None: Poor
(36.0)
|   |   |   |   |   |   Performance_ of _Prior_term Loans = Satisfactory:
Poor (0.0)
|   |   |   TL/TA > 1.860128
|   |   |   |   TL/TA <= 2.154756
|   |   |   |   |   TL/TA <= 2.122796: Excellent (32.0)
|   |   |   |   |   TL/TA > 2.122796
|   |   |   |   |   |   Collateral_Value <= 9803405.142: Satisfactory (6.0)
|   |   |   |   |   |   Collateral_Value > 9803405.142: Excellent (2.0)
|   |   |   |   TL/TA > 2.154756: Excellent (182.0)
|   |   Current_ Ratio > 0.488802
|   |   |   Collateral_Value <= 5484778.638
|   |   |   |   Current_ Ratio <= 2.333333
|   |   |   |   |   Collateral_Value <= 2452660
|   |   |   |   |   |   Collateral_Value <= 466618.5
|   |   |   |   |   |   |   YearsinBusiness <= 3: Excellent (67.0)
|   |   |   |   |   |   |   YearsinBusiness > 3
|   |   |   |   |   |   |   |   Number_of_prior_termloans_settled <=
1.926452: Poor (25.0)
|   |   |   |   |   |   |   |   Number_of_prior_termloans_settled >
1.926452
|   |   |   |   |   |   |   |   |   Number_of_prior_termloans_settled <= 3
|   |   |   |   |   |   |   |   |   |   Trade_Sector = Domestic trade
service: Excellent (32.0)
|   |   |   |   |   |   |   |   |   |   Trade_Sector = Manufacturing:
Excellent (4.0)
|   |   |   |   |   |   |   |   |   |   Trade_Sector = Transportation:
Excellent (4.0)
|   |   |   |   |   |   |   |   |   |   Trade_Sector = Import: Excellent
(20.0/2.0)
|   |   |   |   |   |   |   |   |   |   Trade_Sector =
Building&Construction
|   |   |   |   |   |   |   |   |   |   |   Total_Asset <= 393480:
Excellent (4.0)
```

```
|   |   |   |   |   |   |   |   |   |   |   Total_Asset > 393480: Poor
(8.0)
|   |   |   |   |   |   |   |   |   |   Trade_Sector = Export: Poor
(2.0/1.0)
|   |   |   |   |   |   |   |   |   |   Trade_Sector = Hotel&tourism:
Excellent (0.0)
|   |   |   |   |   |   |   |   |   |   Trade_Sector = Health service:
Excellent (0.0)
|   |   |   |   |   |   |   |   |   |   Trade_Sector = Project: Excellent
(0.0)
|   |   |   |   |   |   |   |   |   |   Number_of_prior_termloans_settled > 3
|   |   |   |   |   |   |   |   |   |   |   Current_ Ratio <= 1.4: Poor (9.0)
|   |   |   |   |   |   |   |   |   |   |   Current_ Ratio > 1.4: Excellent
(4.0/1.0)
|   |   |   |   |   |   Collateral_Value > 466618.5
|   |   |   |   |   |   |   Collateral_Value <= 2024401.875
|   |   |   |   |   |   |   |   Collateral_Value <= 1080808.081
|   |   |   |   |   |   |   |   |   Collateral_Value <= 1070000
|   |   |   |   |   |   |   |   |   |   Collateral_Value <= 635814
|   |   |   |   |   |   |   |   |   |   |   Collateral_Type = BLD
|   |   |   |   |   |   |   |   |   |   |   |   Collateral_Value <= 619026
|   |   |   |   |   |   |   |   |   |   |   |   |   Credit_
Relationship_with_other_bank = Excellent
|   |   |   |   |   |   |   |   |   |   |   |   |   Current_Asset
<=120879: Excellent (8.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   Current_Asset >
120879: Poor (13.0)
```

Annex 2: List of initial attributes

No	Attribute Name	Attribute type	Description
1	Trade sector	Nominal	The kind of business the customer is involved
2	Number of prior term loans Settled	Numeric	The number of loan a given customer is settled
3	Duration of the term loan in number of days	Numeric	Duration of the loan in number of days
4	Term of payment	Nominal	The term in which the bank collect amount from the given customer such as monthly, Quarterly or yearly
5	Granted amount	Numeric	The approved amount of money
6	Yearly payment	Numeric	The estimated amount of money in one year by borrower
7	Total Asset	Numeric	Total asset of the customer
8	Capital	Numeric	The total capital of the customer
9	Current Asset	Numeric	The current asset of the customer
10	Current Liability	Numeric	The current liability of the customer
11	Net worth	Numeric	Current asset-Current liability
12	Total liability	Numeric	The total liability of the customer
13	Current Ratio	Numeric	Current asset divided by current liability
14	TL/TA	Numeric	Total liability divided by total asset
15	TL/C	Numeric	Total liability divided by capital
16	Collateral type	Nominal	The type of collateral secured against the loan by the customer
17	Collateral Value	Numeric	The money value of the collateral type
18	CV/ GA	Numeric	Collateral value divide by granted amount
19	Years in business	Numeric	Number of years the borrower stayed in business
20	Performance in other types of loans	Nominal	Performance in other types of loans if any.
21	Performance of prior term loans	Nominal	Previous performance in term loan
22	Credit Relationship with other bank	Nominal	Credit relation ship with other that is different from United Bank in the country
23	Payment performance	Nominal	The payment status of the borrower

Annex 3: Commercial Credit Report

ሕብረት ባንክ አ.ማ.
UNITED BANK S.C.

የብድር መጠየቂያ ቅጽ
COMMERCIAL CREDIT REPORT

ቅርንጫፍ ብርአምባ ቀን_____
Branch Birr Amba Date

1. የአመልካች ስም/የዋስ ስም
Name of Applicant/Guarantor_____ ዕድሜ Age_____

2. የሚስት/ባል ስም
Name of wife/Husband_____

3. አድራሻ Address	ከተማ Town	ክፍተኛ Keftegna	ቀበሌ Kebele	የቤት ቁ. House No.	የስልክ ቁ. Phone No.	የመ.ሣ.ቁ P.O.Box
የሥራ Business	A.A	—	—	—	—	—
የመኖሪያ ቤት Residence	A.A	—	—	—	—	

4. የፈቃድ ቁጥር፣ ያገር ውስጥ ንግድ ሚኒስቴር
Licence No. Ministry of Domestic Trade_____ የውጭ ንግድ ሚኒስቴር Ministry of Foreign Trade_____

የማዘጋጃ ቤት
Municipal_____ ሌሎች/ዘርዝር/ Others (Specify)_____

5. የተፈቀደላቸው የሥራ ዓይነት/ዝርዝር/
What are the types of business the applicant is licensed to operate

6. የተቋቋመበት ዘመን Date Established _____ መነሻ ካፒታል በብር Original Investment Br._____00 አሁን ያለው ካፒታል በብር Present Capital Br._____

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. የቀስትናው ዓይነትና ግምት
Type and value of security offered

11. የብድሩ አከፋፈል /እቅድ/
Proposed mode of repayment:

12. አመልካች ከባንኩ ጋር ያለው፣ ግንኙነት፣ በአስቀማጭነት፣ የቅርንጫፍ ስምማ የሚሣብ ቁጥር
Relation with Bank: as Depositor_____ Name of Branch A/No

በዋስትና
as Guarantor_____ ብድሩን የሰጠው ቅርንጫፍ ስም
Name of lending Br .Same Branch

13. የቀድሞ ብድሮች ሁኔታ/በቅርንጫፉ ኃላፊ የሚሞላ/
Record of previous loan (to be filled by Branch)

Loan No.	Amount in Birr	Date Granted	Due Date	Remarks
1				
2				

የሂሳብ ሠንጠረዥ
BALANCESHEET

AS AT_____

LINE NO.	ሐብት ASSETS	ያስመዘገቡት ንብረት ልክ DECLARED ቀን Date_____	የታየው ወይም የተረጋገጠው ንብረት ልክ CHECKED ቀን Date 29/11/2005	KEY
1	ጥሬ ገንዘብ በባንክ ያለ Cash: a) in bank			
	በእጅ b) on hand			
2	ወደፊት የሚሰበሰብ Receivables ከእቃ ሽያጭ/እቁብ Accounts/Iqub			
	ከተስፋ ሠነዶች Notes			
3	በሱቅ ወይም በመጋዘን ያለ የሸቀጥ መጠን Goods in Stock			
4	ለእቃ ግዢ የተደረገ የቅድሚያ ክፍያ Prepayment on Merchandise			
5	ተንቀሳቃሽ ሐብት Current Assets			
6	የፋብሪካ እቃዎችና መሣሪያዎች Equipment and machinery			
7	ተሽከርካሪ Motor vehicles			
8	የቤትና የቢሮ እቃዎች Furniture and Fittings			
9	ቤቶች Buildings			
10	ሌላ ተጨማሪ ሐብት Other Assets			
11	ቋሚ ሐብት/የተጣራ/ Fixed Assets (Net)			
12	ጠቅላላ ሐብት Total Assets Birr			
	እዳ LIABILITIES			
13	የሚከፈል እዳ Payable: ለእቃ ግዢ/ለእቁብ Accounts/Iqub			
	የተስፋ ሰነዶች Notes			
14	ለግብር የሚከፈል Tax Payable			
15	የባንክ ብድር Bank Loans			
16	ለረዥም ጊዜ እዳዎች በዚህ ዓመት የሚከፈል Current Portion, Long Term Debts			
17	ሌላ እዳ Others			
18	በቅርብ የሚከፈል እዳ/ጊዜያዊ እዳ/ Current Liability			
19	በረጅም ጊዜ የሚከፈሉ እዳዎች Long Term Debts			
20	ካፒታልና መጠባበቂያዎች Capital and Reserves			
21	የእዳና ካፒታል ድምር Total Liabilities and Capital Birr			

የገቢና የወጭ ሠንጠረዥ
STATEMENT OF INCOME AND EXPENSES

LINE NO.	ገ ቢ INCOME	AMOUNT	KEY
22	ሽያጭ Sales 3,750 x 275 days		
23	የተሸጠ እቃ ዋጋ Cost of Goods Sold		
	መነሻ እቃዎች Beginning inventory, as at Birr		
	ሲደመር የተገዙ እቃዎች Add-Purchases for the period "		
	ሲቀነስ መጨረሻ የቀሩ እቃዎች Less ending inventory, as at "		
24	ሌሎች Others		
25	ያልተጣራ ትርፍ Gross Profit		
	ወጭዎች EXPENSES		
26	ደመወዝ Wages and Salaries		
27	የንግድ ቤት ኪራይ Business Premise Rent 874 X 12		
28	ስልክ፣ መብራትና ውሃ Utilities 200 x 12		
29	ማደሻና ጥገና Maintenance and Repair		
30	መድን Insurance		
31	የአገልግሎት ተቀናሽ Depreciation		
32	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 እዳ Guarantee Liability (in total)		

የአመልካች ፊርማ
Applicant's Guarantor's Signature _____

	የትንተና ሚዛኖች ANALYTICAL AND COMPARATIVE RATIOS	በዚህ ዓመት This year	ባለፈው ዓመት Last Year	
39	የተጣራ መንቀሳቀሻ ካፒታል Net Working Capital			
40	ተንቀሳቃሽ ሀብት ከጊዜያዊ እዳ ማነጻጸሪያ Current Ratio			
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 እዳ ከተጣራ ሀብት ማነጻጸሪያ Total Debt to Worth Ratio			

አዘጋጁው
Processed by_____

የ ሂ ሣ ብ ሠ ን ጠ ረ ዥ ማ ብ ራ ሪ ያ
EXPLANATORY NOTES ON FINANCIAL STATEMENTS

በቅርንጫፉ ኃላፊ የሚሞላ
(to be filled by the Branch)

KEY	DESCRIPTION	AMOUNT
1		
2		
5		
3		
6		
7		
8		
9		

Annex 4: Loan Approval Form (LAF)

UNITED BANK S.C.
BIRR AMBA BRANCH

CODE NO
CATEGORY
TRADE LICENCE NO
TIN No

LOAN APPROVAL FORM UB/BAB/LAF/	**BRANCH: BIRRAMBA**	**DATE**
1. NAME OF APPLICANT:		
2. FACILITY APPLIED FOR (other request):		
3. PURPOSE OF FACILITY:		
4. APPLICANT'S BUSINESS:		

5.	**PRESENT FACILITIES**	**LIMIT APPROVED**	**PRESENT BALANCE**	**EXPIRY DATE**	**INTEREST RATE P.A**
	Term Loan at Birr				
	Term Loan at Birr				
	Merchandise @ %				
	Merchandise @ %				
	SUB TOTAL				
	Letters of Credit at-- margin (sight)				
	Letters of Credit at _____ (Acceptance)				
	Guarantees Issued by the Bank (Foreign)				
	Guarantee Issued by the Bank (Local)				
	Performance guarantee				
	GRAND TOTAL				

6	**SECURITY DESCRIPTION**	**EVIDENCED BY**	**VALUE**
1.			
2.			
3.			
4.			
5.			
		TOTAL	

HIGHEST DEBIT	HIGHEST CREDIT	HIGHEST DEBIT	HIGHEST CREDIT
1	1	1	1
2	2	2	2
3	3	3	3
LOWEST DEBIT	**LOWEST CREDIT**	**LOWEST DEBIT**	**LOWEST CREDIT**
1	1	1	1
2	2	2	2
3	3	3	3

8. Applicant Net Yearly Income br

9.OTHER LIABILITIES TO THE BANK AND PARTIES (INDICATED ANY OVERDUE)

BORROWER	GUARANTOR
N I L	N I L

11. BREAK DOWN OF THE FINANCIAL STATEMENTS OF BORROWER AND GUARANTOR

	(BORROWER IN 000'S BIRR)		(GUARANTOR IN 000'S BIRR)	
Provisional Particulars:	This Year	Last Year	This Year	Last Year
Net working capital	Birr	Birr	Birr	Birr
Capital & Reserves	Birr	Birr	Birr	Birr
Profit	Birr	Birr	Birr	Birr

10. FACILITIES WITH OTHER BANKS (Give details)

GENERAL REMARKS AND SUMMARY OF PAST RECORDS

11. RECOMMENDATION OF THE LOAN OFFICER

12. RECOMMENDATION OR DFICION OF THE BRANCH CREDIT COMMITTEE
(Give reasons for declining or for deviation from the recommendation)

13. RECOMMENDATION OR DECISION OF HEAD OFFICE CREDIT COMMITTEE
(Give reasons for declining or for deviation from the recommendation)

14. RECOMMENDATION OR DECISION OF MANAGEMENT CREDIT COMMITTEE
(Give reasons for declining or for deviation from the recommendation)

15. DECISION OF FINANCE COMMITTEE OF THE BOARD
(Give reasons for declining or for deviation from the recommendation)

Annex 5: Monthly Loans and Advance Return Form

NO.	Name of customer	Booking Date	Granted Amount	Maturity Date	O/Standing Amt	Repayment	Arrears	Classification	Period prepared as at	Total Asset	Net Worth

Gross Income	Non Taxable Income	Gross Profit	Tax Liability	Audited	Risk Free Amt	Collateral type

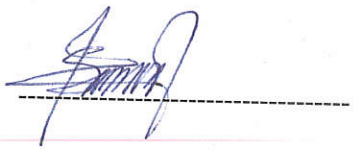
Annex 6: Sample data for Loan Customers

Customer	Trade_Sector	Number_of_prior_termloans_settled	Duration_of_ the_termloan_in_No_of_ Days	Term_ of_ Payment	Granted_amount
1	Domestic trade service	4	622	M	350433.24
2	Manufacturing	2	416	M	490506.4
3	Import	5	359	Q	234901.98
4	Building&Construction	1	596	M	1724731.57
5	Transportation	1	688	M	396438.48
6	Hotel&tourism	4	558	M	1276306.74
7	Export	93	63	M	1678377.72
Customer	Yearly_payment	Total_Asset	Capital	Current_Asset	Current_Liability
1	205640.0846	3525831.479	928648.086	1283636.958	665823.156
2	430372.2019	9427900.888	1800158.488	3254509.964	2825316.864
3	238827.9184	422823.564	281882.376	44819.29778	140941.188
4	1056253.394	4311828.925	2914796.353	1422903.545	667816.0639
5	210319.8331	1500618.756	475726.176	572853.6036	401988.6187
6	834860.1435	15680704.61	1148676.066	7217514.615	5194568.432
7	9723934.41	12648254.5	1258783.29	9491226.007	507289.6659
Customer	Networking capital	Total_Liability	Current_ Ratio	TL/TA	TL/C
1	617813.8021	1296602.988	1.927895	0.367743891	1.396226415
2	429193.1	3254509.964	1.15191	0.345199849	1.807901907
3	-96121.89022	140941.188	0.318	0.333333333	0.5
4	755087.4813	1517763.782	2.130682	0.352	0.520710059

5	170864.9849	572853.6036	1.425049	0.381744931	1.204166667
6	2022946.183	24887981.43	1.389435	1.587172391	21.66666667
7	8983936.341	32728365.54	18.70968	2.587579618	26
Customer	**Collateral_Type**	**Collateral_Value**	**CV/ GA**	**YearsinBusiness**	**Performance_ of _Prior_termLoans**
1	BLD	6248070.08	17.82955886	12	Poor
2	BLD	2452660	5.000260955	5	Excellent
3	SH	493900	2.102579127	12	Satisfactory
4	VM	7185120	4.165935224	3	None
5	VM	5590408	14.10157763	3	None
6	VM	18608347	14.57983917	12	Excellent
7	VM	19254583.31	11.47213948	12	Excellent
Customer	**Performance_in_otherTypes_of_Loans**	**Credit_ Relationship_with_other_Bank**	**Payment_performance**		
1	None	Excellent	Satisfactory		
2	Excellent	Excellent	Satisfactory		
3	Excellent	None	Poor		
4	Excellent	Excellent	Excellent		
5	Excellent	Excellent	Satisfactory		
6	None	Excellent	Excellent		
7	None	Excellent	Excellent		

Declaration

The thesis is my original work and has not been presented for a degree in any other university and all the sources of material used for the thesis have been duly acknowledged.



Samson Tadesse
January, 2009

This thesis has been submitted with my approval as a university advisor

---

DR. Manoj VNV