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

APPLICATION OF DATA MINING TECHNIQUES FOR EFFECTIVE CUSTOMER RELATIONSHIP MANAGEMENT OF MICROFINANCES: THE CASE OF WISDOM MICROFINANCE

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

WAKGARI DIBABA

AUGUST, 2009

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

APPLICATION OF DATA MINING TECHNIQUES FOR EFFECTIVE CUSTOMER RELATIONSHIP MANAGEMENT OF MICROFINANCES: THE CASE OF WISDOM MICROFINANCE

BY

WAKGARI DIBABA

AUGUST, 2009

NAME AND SIGNATURE OF MEMBERS OF THE EXAMINING BOARD

1. Million Meshesha (Ph.D.),Advisor_____

2. Getachew H/Mariam (M.I.Sc.),Co-Advisor_____

3._____

4._____

5._____

6._____

AKNOWLEDGEMENT

First of all, I would like to extend my special gratitude to my advisor Dr. Million Meshesha for his unreserved support and assistance throughout my research work. His constructive comments and suggestions, in my study and in writing the thesis, were highly valuable. In general, his helpful personality, cooperativeness and dedication are remarkably worth mentioning.

My special thanks also go to Ato Getachew Hailemariam, my co-advisor, whose comments and suggestions were also highly crucial in the research. My special appreciation and gratitude is not only for the comments and suggestions he provided me, but also for his willingness to cooperate and assist me whenever I requested him.

I would like to thank the entire staff of the WISDOM microfinance Head office, especially Ato Worku Tsegaye (General Manager), Ato Dereje Tadesse (Human Resource Manager), and Ato Tigistu and Ato Biniam who are IT officers, for their cooperation and willingness to provide me with all valuable information as and when I needed.

Last, but not least, I am very much grateful to my family for their constant assistance and encouragement throughout my study.

Table of Contents **Page**

LIST OF FIGURES

Page

LIST OF TABLES

Page

LIST OF ABBREVIATIONS

ADP	Area Development Program
ARFF	Attribute-Relation File Format
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSV	Comma Separated Value
DM	Data Mining
GNU	General Public License
IT	Information Technology
KDD	Knowledge Discovery in Data base
MFI	Microfinance Institutions
ML	Machine Learning
OLAP	On-Line Analytical Processing
SEWA	Self Employed Women Association
TMFS	Total Microfinance solution
WEKA	Wekato Environment for Knowledge Analysis
WMF	Wisdom Microfinance
WMFI	Wisdom Microfinance Institution

ABSTRACT

The proliferation of information and communication technologies enabled companies to deal with large quantities of data. Microfinances are one of such institutions that collect, process and store huge amounts of records from time to time and therefore deal with voluminous amount of data. On the other hand, Microfinances are facing problems in customer handling; the proportion of customers staying intact with the same microfinance as a customer is very less compared to potential customers. The WISDOM microfinance is facing such problem where most customers are churning/ shifting to other competitors after using the loan service once or few times only. The existing past and historic data could be actionable and usable for decision making process that improves customer relationship management with the help of data mining techniques. One of the various applications of data mining is in support of customer relationship management through pattern mining and uncovering regularities.

This paper reports the study of application of data mining in microfinance that helps build a classification model which supports in prediction of a new borrowers status (highly privileged, moderately privileged or less privileged) during the loan decision making in the organization.

A classification model is built based on the borrowers' corpus data obtained from the WISDOM microfinance. Essential preprocessing activities have been applied to clean and make it ready for the Experimentation. Then experiments using J48 decision tree classifier of the WEKA 3.7.0 software have been conducted using the preprocessed dataset with different attributes and parameters setting in order to arrive at the optimal model. The classification model with the best accuracy level (78.502%) and relatively less number of leaves and tree size is constructed to predict the new customer class label (highly privileged, moderately privileged or less privileged).

CHAPTER ONE

INTRODUCTION

1.1 Background

Microfinances are one of the critically important sectors in any part of the world for its socio-economic value in the society and in the country in general. Different literatures define Microfinance in slightly different ways.
Microfinance is defined as banking the unbankable, bringing credit, savings and other essential financial services within the reach of millions of people who are too poor to be served by regular banks, in most cases, because they are unable to offer sufficient collateral or the banking policy (Dows, 2008).

Microfinance is the supply of loans, savings and other basic financial services to the poor (Kiva, 2005). Since the targeted customer for the microfinance institutions are the poor, the financial services usually involve small amounts of money- small loans and small savings that differentiate the microfinance from formal banks.

By providing small loans and savings facilities to people who are excluded from commercial financial services, microfinance has become a strategy for reducing poverty. Access to credit and deposit services is a way to provide the poor with opportunities to take an active role in their respective economies through entrepreneurship, building income, bargaining power and social empowerment among poor women and men (Men,2006).

The financial services offered by microfinance institutions include micro credit, micro saving, money transfer vehicles, and micro insurance. However, the most popular are micro credit and micro savings especially for microfinance of developing countries. Microcredit is a service for poor entrepreneur or farms that are not bankable for reasons such as lack of collaterals, steady employment, income and verifiable credit history, but still possess entrepreneurial capability and possibility.

Therefore microfinance plays an important role for both developing and developed countries, the degree becoming greater for the developing countries where the poor prevails.

For countries like Ethiopia where above 75 percent of the society is poor (World Bank, 2002), Microfinance institutions play a key role in the welfare of the society and the economic and social development of the country in general. U.S. Agency (2005) states that Ethiopia is one of the least developed countries in the world, ranking 168th out of 173 countries in the 2002 United Nations Development program, Human Development Index. The per capita GDP was $668 in 2000; 76.4 percent of the population lives on less than $2 per day (World Bank, 2002). Hence it is unquestionable that such Microfinance institutions play key role in Ethiopia.

Microfinance has evolved as an economic development approach intended to benefit low-income groups; the provision of financial services to low income clients including the self employed. Sometimes it is said to be “Banking the poor” (Valarie, et.al, 1976) that has been proven to empower very poor people around the world to pull themselves out of poverty. Relying on their traditional skills and entrepreneurial instincts, very poor people, mostly women, use small loans, other financial services and support from organization called Microfinance institutions to start, establish, or expand very small, self supporting businesses.

One of the fundamental challenges in the micro financing sector is the capacity to design and implement an effective loan disbursement mechanism that ensures high customer attraction and retention. The micro financing institutions can utilize their past data using powerful technologies such as data mining technologies in order to help them devise new strategies that enable them to attract more customers and retain existing ones.

Data mining is defined as a process of extracting valid, previously unknown, comprehensible and actionable information and potentially useful knowledge from large data base and using it to make crucial business decision (Connally, et.al, 1999; Han and Kamber, 2001).

The area of data mining has got much attention of industry due to the existence of large collection of data and the increasing need of data analysis and comprehension. Today, data mining is being used by several industries including banking and finance, retail, insurance, telecommunications, etc (Madhan, 2006).

In banking and finance institutions, data mining has been applied for various purposes, which include customer segmentation and profitability, credit scoring and approval, predicting payment default, marketing, detecting fraudulent transactions, cash management and forecasting operations, optimizing stock portfolios, and ranking investments mainly for credit risk assessment and customer scaling (Madhan, 2006).

1.2 Statement of the problem and Justification

Customer attraction and retention issues have become indispensable factors for such business sectors as Microfinances in this information age where every company is striving to win a competitive advantage. Competitive pressure is becoming very strong in microfinance enterprises that call for continuous assessment and analysis of customers' service use behavior.

A preliminary investigation on the microfinance activity of Wisdom Microfinance (WMF) shows that the customer attraction and satisfaction is not as much as pre-envisaged. This is mainly revealed from the number of fewer customers coming to the company for loan service repeatedly compared to the number of customer appearing for the first few one to three instances(Loan cycles).

Based on customer data of the WMF, Table 1.1 shows the number of cycles customers use the loan service from Wisdom Microfinance

Cycle	Number of Customer	Relative Percentage
1	3043	31.86
2	2446	25.61
3	1223	12.80
4	1284	13.45
5	352	3.69
6	323	3.38
7	146	1.53
8	322	3.37
9	46	0.48
10	33	0.34
11	33	0.34
12	6	0.06
13	13	0.13

Table 1.1 Customer distribution based on Loan cycle,

From Table 1.1, one can observe that more than 80 percent used the loan service for only few cycles (less than 4) while less than 1 percent of the customers used the loan service repeatedly for more than 10 cycles.

This is an indication of the fact that only few customers are staying for long period with the company as customers while large percent of customers are churning/abandoning after their first few (one to four cycle) appearance for loan service.

From the interviews and discussions made with senior managers of the organization the existing system requires significant improvement in customer relationship management. There are, of course, attempts made by the company for attracting and retaining more customers. In the current system, as the officials say, there is a kind of incentive when borrowers use the loan service for large number of cycles (repeated uses). Borrowers coming for increased number of cycles for the loan service are

allowed relatively larger loan amounts with relatively higher frequency than those borrowers with less number of cycles for loan service request.

For example, according to the rule of the organization when a borrower is coming for loan cycles of 4-6, he/she can be granted a 50% increment loan amount on the previous loan amount. When the loan cycle becomes 7 and above he/she is a gain granted a 100% increment on the loan amount that he/she was offered earlier. This is done because an increase in number of service cycles is an indication for proving customers' loyalty to the company ,as the senior officials say, since there is no way to predict customers' loyalty (repeated usage).

However, in the current system ,there is no way or means of predicting whether a new customer may stay borrowing loans for prolonged time or not in order to grant him/her with a large loan amounts and frequencies he/she needs. It would be highly beneficial if a system or method were available that would classify the customer with respect to their loan usage (loan cycle) that would help to predict new customers' class label.

Business organizations gather transaction data through their day to day activities. That is why it is stated that (http://intellinova.com) most businesses own more data than they can deal with –prospect and customer list, sales data, market research, and complaints-and yet their staffs do not use this data to effectively manage customer relationship.

Likewise, Wisdom Microfinance Institution(WMFI) currently consists of huge amount of data (over 20,000) of their customers with relevant attributes such as employee size, Number of children, Area, type of engagement (sector), loan cycle, loan type, customer address, etc. The dataset consists about 16 attributes and over 20,000 records. They have been keeping track of such huge customer data with spread sheet programs on each of the computer systems they have at their various branches, a copy of which exists at the head office as well. Hence the existence of huge data with potentially relevant features would be used to support the decision making process if the organization has to achieve its objectives and aspirations. The data stored over times may help to generate the borrowers classification model that

would enable to devise and implement appropriate loan service strategy in support of customer attraction and retention.

Therefore the present research work is initiated to come up with a data mining technique that helps to predict customers' loyalty utilizing the customers' existing data, so that the company can pass proper loan decisions on the provision of loan services. This has a significant impact in improving customer relationship management of the company.

Research Questions:

The research attempts to answer the following major questions:

- What are the best features to consider in passing loan decisions by the company?
- Is data mining technique (like classification) suitable for suggesting best features for loan decisions?
- Is it possible to characterize groups of customers with similar patterns?
- Can the patterns be useful for devising new strategies in customer relationship to strengthen existing customer loyalty and to attract new customers?

1.3 Objective of the Study

The general and specific objectives of the research are described below.

1.3.1 General Objective

The objective of this study is to explore the potential applicability of data mining techniques to build customer classification model for better customer relationship management of Wisdom Micro finance.

1.3.2 Specific Objectives

In order to achieve the general objective, the specific objectives identified are the following:

- To asses related documents and previous works in the area so as to get an insight in to the area and to find out related works and their contributions to the study at hand.
- To collect relevant dataset required for the mining, analysis and performance evaluation.
- To prepare the data for pattern mining by selecting, cleaning, reducing, summarizing and integrating.
- To design a classification scheme in order to find potential patterns for better customer relationship management.
- To evaluate the performance of the classification model in characterizing the customers of the company.
- To report on the results and make recommendations for further researches.

1.4 Application of the research

Besides its being an academic exercise, the findings of the research can be used in various areas. The intended model would be used to know deeper about the existing data and contribution of some attributes for customer classifications. This would help the management of the organization to devise different strategies like different level of loan disbursement based on various factors such as number of employee, family size, level of income, type of engagement (sector), etc. with the help of the patterns mined.

Other institutions such as banks and other financial institutions, which render similar loan services with similar policies and procedures, could also benefit from the results obtained in order to make appropriate decisions during their loan disbursements.

Both rural and urban economically active poor individuals, groups, households, and the community at large would be the beneficiaries from the final result of the experimental research.

1.5 Research Methodology

The general approach of the research is a quantitative analysis in that the major processes include collecting and organizing the transaction (customer) data. Social data (Social data, in the context of the organization and throughout this study as well, is used to refer the customer data that contain demographic and social related information such as the customer identification, type of group, type of sector engagement, the loan size requested/offered, age, sex, number of children, loan type, etc) which was organized and made available for report and administration purposes will be used as an input data for the data mining purpose.

However it has also included a qualitative aspect in that it requires an input from domain experts and business operations to draw meaningful conclusions from the patterns mined following data mining procedures. To this end, interviews, formal and informal discussions were made with some senior officials and clerks as well.

The following sections are therefore about the methods that were used to undertake the research; literature Review and Business understanding, data collection methodology, and the data mining methodology.

1.5.1 Literature Review & Business Understanding

The researcher has made review of various literatures such as books, journals, articles, conference pages, etc pertaining to the subject matter of data mining and customer relationship management in order to get an insight in to the area.
For relevant and feasible decisions to be drawn from the data mining result and for the business understanding as well, interviews, observations, and document reviews have been made.

Interview is made with purposefully selected staff including management and clerk worker in loan related activities for the business understanding and to get supportive information in order to interpret the results of the data mining.

Three senior officials (General Manager, Human Resource Manager and Chief Information Officer) are interviewed for the purpose (The interview guide is attached as Appendix II).
Frequent formal and informal discussions are also made with 2 clerk workers which are accountants in the organization and with an Information Technology (IT) officer. The discussions were essential for the business understanding, to get insight how the business rules are implemented and for understanding the data in general (The interview guide for the formal interviews made with the clerk workers is annexed as Appendix III).

Observation: is made at branch Microfinances while the loan service is being offered to the customers. This enabled to reveal the procedures for approving and entertaining a new customer for the loan service.

Document analysis: documents pertaining to the organization's policies and procedures have also been reviewed in the process of conducting of the research.

This gave an insight in to the business policies and procedures regarding loan services.

1.5.2 Dataset Identification

The main sources of data for such research employing data mining technology is one or more repositories within the identified organization. For the sake of security, and to safeguard confidential issues, the analysis made in this research is entirely based on social data. Accordingly, social data available in spreadsheet programs are found to be potential sources of data. Thousands of customer related data are being recorded in each month at the head office, which gives a large number of datasets (which amounts to over 20,000 records) captured over the past years. From the existing data only the social data was separated from financial statements and confidential data, and it is finally copied to different CD and accessed for the data mining purpose. The total dataset identified for this research work amounts to 9715 (which is purely social data).

1.5.3 Data mining methods

There are various tools available for data mining, such as Knowledge Studio, WEKA (Wekato Environment for Knowledge Analysis), and others (Han and Kamber, 2001). Among those tools, WEKA is selected and used for data mining tasks since it provides sufficient facilities and since it is easily accessible as well.

WEKA is a ccollection of machine-learning algorithms with an open-source Java package that supports numeric, nominal, string, and date format files for processing data on several methods (Palous, N.D.). WEKA software is issued under the GNU General Public License. It incorporates an association rule learner. In addition to the learning schemes, WEKA also comprises several tools that can be used for datasets preprocessing (Palous, N.D.).

Among the various data mining tasks, as also stated in the scope, classification is given emphasis and classification model building is applied for the mining process since it provides method for predicting the value of a (categorical) attribute (the class) based on the values of other attributes (the predicting attributes). Moreover,

according to Romerio, et.al (2007), classification is one of the most frequently studied problems by Data Mining (DM) and Machine Learning (ML) researchers.

In data mining, there are various methods or algorithms used for classification including Neural network, Bayesian Network, Decision tree, Regression, etc (Zemke,2003).The decision tree method is used for the current research for its better visualization capability, very good generalization capability, and ease of interpretations compared to the other methods (Bakir, 2006). The decision tree method generates tree shaped structures in which construction of trees is simple. Unlike the other methods such as regression models, the decision trees can easily be understood and interpreted by the users (Bakir, 2006).

Important preprocessing tasks are applied for the pattern mining task. CRoss Industry Standard Process for Data Mining (CRISP-DM) is followed in the present research work. CRISP-DM involves the data mining processes, data mining goal, data understanding, data preparation, model building, model evaluation and deployment (Han& Kamber, 2001).

Han& Kamber (2001) further describes the CRISP-DM as a data mining process addressing the following issues:

- Mapping from business idea to data mining problem
- Capturing and understanding data
- Identifying and solving problems within the data
- Applying data mining techniques
- Interpreting data mining results within the business context
- Deploying and maintaining the data mining results
- Capturing and transferring expertise to ensure future benefits from experience

Hence most of the issues described by Han & Kamber (2001) are addressed in the same steps stated above except that the last two issues are not incorporated for the same reason specified under the scope.

1.5.4 Testing/Experimentation Mechanism

After all the necessary data preprocessing activities such as data selection, cleaning, summarization/aggregations are done and after the data is put in a format that WEKA software can process, that is Attribute Relation File Format(ARFF) file, it is then used to generate patterns and models. Training the decision tree model helped to get a pattern on borrowers status that are classified as “highly privileged”, ”moderately privileged” or “less privileged” based on certain variable inputs.

A testing mechanism known as stratified tenfold cross validation method is used for evaluating the model built. Ten fold cross validation is defined as testing mechanism where data is divided randomly into 10 parts in which each part is held out for testing in turn and the learning scheme trained on the remaining nine-tenths and the leaning is executed 10 times as a result of which the average is taken as the overall error rate. The Tenfold cross validation is called stratified type if random sampling is done in such a way as to guarantee that each class is properly represented in both training and test sets(Witten and Frank, 2005).

The stratified tenfold cross validation method is found more appropriate and feasible than other methods like partitioning the data in to training and test sets. Because this gives the opportunity to make use of all the available dataset for the training while still testing for the accuracy of the model is possible. The dataset is randomly split in to 10 parts and the class is represented in approximately the same proportion as in the full dataset. Testing will be then made 10 times in each turn 1 of the 10 partitions are used for testing and the remainder is used for training. The method repeats the procedure 10 times so that, in the end, every instance has been used exactly once for testing after which the average of the 10 times test results is taken as the overall error rates.

In evaluating the model, its accuracy, the number of leaves generated and the tree size are taken into consideration. The one with higher number of accuracy and least number of leaves and tree size is selected to be the best model.

1.6 Scope and limitation of the Study

The scope of the research is limited to investigating the potential applicability of data mining in identifying determinant factors for customer classification in Wisdom Microfinance so as to create an effective customer relationship management. There are about six common tasks that a data mining can accomplish: namely, classification, estimation, prediction, association, clustering and prediction. However the current research is mainly concerned with classification for its suitability to build predictive model to achieve the purpose, and for its popularity.

Hence the components included are data understanding, data preparation, model building and testing. It doesn't include some of the CRISP-DM process such as deployment.

The dataset was obtained from the head office where data about customers of the company are available on a computer system. However the input data was limited to those dataset which are purely social data. This is done because, as the officials in the organization say, other data such as collection data, financial data, etc are confidential for the organization.

Some of the limitations in the research were accessibility of clerks/Accountants at the area/branch microfinances in the parts of the country. Even though the borrowers' data of all branches could be accessible at the head office it was difficult to get access to those clerks and accountants working at different branches for interview, because of time and other constraints, that would have helped the researcher get more insight in to the data and potential factors that could have contributed to come up with more improved accuracy. It was also requiring much effort and time to separate some of the data from those top secrete and confidential ones, hence some of the data with high integration or interrelation with such confidential and secrete data were disregarded from the dataset used for mining which in other way, could have helped to improve the classifier accuracy.

The research is limited to develop a classification predictive model based on the predefined target class values (categorizations). This is so because it was requiring

much time and efforts, and frequent and exhaustive discussions with the experts on the area to modify existing ones and to come up with more categorizations. On top of constraints in time and other resources, the fact that the officers have been highly occupied with overlapping duties during the conduct of the research contributes to the limitation of the research. This is so because it was difficult to contact the officers as frequently as it was necessary that could have helped to illicit more information, do more exhaustive analysis on both predictor and target variables of the dataset for an improved, more accurate and more efficient model building.

1.7 Thesis Organization

The thesis is organized in to five chapters. The first chapter is about general overview of the research that includes background study, statement of the problem, application, objective, methodology and scope of the present research work.

Chapter two is review of literatures on related areas. In the first part, overview of microfinance; definition and related concepts, the historical overview of microfinances in Ethiopia and the WISDOM microfinance, are discussed. The second part of this chapter discusses about data mining, related technologies such as Online Analytical Processing (OLAP), data warehouses, and the data mining processes. Review of related works on the application of data mining for customer relationship management and application of data mining in financial institutions are also presented in this chapter.

Chapter three deals with the discussion of the decision tree classification algorithm used for the model development, testing mechanisms and the experimental design in brief.

Chapter four presents the experimentation of the data mining process and performance of the classification model. Finally, the conclusion and recommendation part of this research are detailed in chapter five.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter presents review of literature on background of the domain area/organization, data mining concepts, and related research works. The purpose of the current study is to experiment on application of the data mining techniques for improving customer relationship management of Microfinances taking the case of Wisdom microfinance. To this end, it is essential to see an overview about the microfinance institution, conceptual discussions regarding the data mining technology and review of related works.

2.2 Overview of Microfinance

This section presents the historical overview of microfinance, microfinance in Ethiopia, and finally the Wisdom Microfinance including its overview, major services and automation efforts.

2.2.1 Historical overview of Microfinance

The history of today's microfinance takes us to the history of early 1970's where the concept of credit union was developed by Fredrick Wilhelm Raiffeisn and his supporters (Mercy Corps, 2006). However formal credit and saving institutions for the poor have been around for decades providing customers who were traditionally neglected by commercial banks a way to obtain financial services through cooperatives and development finance institutions (Mercy Corps, 2006).

According to Littlefield and Rosenberg (2004), Microfinance institutions have emerged over the past three decades providing financial services to low income clients. They further state that most of the early pioneer organizations in the microfinance movement operated as non profit, socially motivated non governmental organizations.

Over the last 30 years, the microfinance industry has proven that the extreme poor are bankable. Not only do they repay loans, but they also do so with very low defaults and relatively high interest rates. Microfinance Institutions (MFIs) can, and have, become commercially viable enterprises (McKinsey & Company, 2005).

Between the 1950 and 1960's governments and donors focused on providing subsidized agricultural credit to small and marginal farmers, in hopes of raising productivity and incomes (Kiva, 2005). In the 1990's many of the institutions transformed themselves into formal financial institutions in order to access and on-lend client saving, thus enhancing their outreach (Kiva, 2005).

SEWA (Self Employed Women Association) registered as trade union in Gurjarat, India in 1973;Garameen-Bank established in Bangladesh in 1976 to address the banking problem; Bank Rakyat Indonesia, largest microfinance in developing countries during the mid 1980's were some or the early pioneer formal microfinances (Mercy Corps, 2006).

According to (Mercy Corps, 2006), it was not until the mid 1990's that the term micro credit began to be replaced by a new term that included not only credit, but also savings and other financial services. Since then, microfinance has been used as the term or choice to refer to a range of financial services to the poor including credit, services such as insurance and money transfer.

The microfinance sector has been expanded in many countries as a strategy for poverty alleviation since the mid of 1990's (Mercy Corps, 2006). As stated in Kiva (2005), the World Bank estimates that there are now over 7000 microfinance institutions, serving some 16 million poor people in developing countries. The total cash turnover of MFIs worldwide estimated at US$ 2.5 billion and the potential for new growth is outstanding.

2.2.2 Microfinance in Ethiopia

In case of Ethiopia, lack of finance is one of the fundamental problems impeding production, productivity and income of rural and urban households. Hence Microfinance Institutions (MFIs) in Ethiopia are recognized/and being used/ as the

key institutions to channel microfinance to the poor including funds under the food security program.

The development of microfinance industry in Ethiopia can be traced back to the early 1970's when NGOs in Ethiopia were delivering relief and development services such as emergency food, education, water and medicine to the underprivileged; the NGO's were directly funding micro credit services as part and parcel of their relief program (Mercy Corps, 2006).

Today, practitioners and donors are increasingly focusing on expanded financial services to the poor in frontier markets and on the integration of microfinance in financial system development (Mercy Corps, 2006). Owing to the same fact, various microfinance institutions have been emerging in Ethiopia since a decade.

Establishment of sustainable microfinance institutions serving large number of the poor has become one of the key components of Ethiopia's development strategy. The government instituted a legal and policy framework for microfinance institutions in 1996 through proclamation 40/1996; however Non Governmental Organization (NG0), Credit schemes and informal sources of finance have existed in Ethiopia for many years before (Gebrehiwot, 2002).

Sebstad (2003) states that since the institutionalization of the legal and policy framework for MFI about 20 MFIs have registered with National bank of Ethiopia and operate under the auspices of this proclamation. These MFIs in Ethiopia focuses on group- based lending and promote compulsory and voluntary savings.

As the microfinance institutions do not require as such collaterals unlike banks and other financial institutions, the microfinance institutions so far established in Ethiopia use joint liability, social pressure, and compulsory savings as alternatives to conventional forms of collateral. In fact, stated in Sebstad (2003), currently the proclamation requires the MFIs to provide credit through group based lending methodologies which is said to mobilize savings by restricting the size of loan up to max of birr 5000 and repayment term restricted to no more than one year.

In the concept of microfinancing even though the aim is to provide the poor with small loans or other financial services in order to help them start their own business generating, or sustain an income and often begin to build up wealth and exit poverty, the MFIs provide the small loans with relatively high interest rates.

There are three basic reasons why the interest rate becomes higher than that of banks and other finance institutions, as stated in (Mercy Corps, 2006). There are three types of costs a MFI has to cover when they make micro loans. These are the cost of the money it lends, cost of loan defaults, and transaction cost. The cost of the money is just some amount or some percent of the money amount lent, which is attributed to the money value for its use; cost of loan default is some amount of money or percent for defaults that is determined from experience. Finally, the cost of transaction is just the cost of processing the loan disbursement and payments and follow up monitoring and this is attributed to the staff, time and effort spent. This cost of transaction doesn't depend on the amount of money lent unlike the first two costs, once it is determined by the microfinance institution.

Based on the general principle stated above the MFIs in Ethiopia have relatively higher rates than the banks and other formal finance intuitions. However there is no fixed rate for all organizations to use. The interest rates vary across organizations.

MFIs in Ethiopia provide both non agricultural and agricultural loans. Both types of loans are provided through group lending methodologies. The agricultural loans generally require a one time or balloon payment at the end of the loan term while other loans typically are paid on weekly or monthly basis.

According to the report of Sebstad (2003), a few microfinance institutions also provide financial services beyond saving and credit. For example two governments supported MFIs manage remittance for about 100,000 pensioners each month. And some MFIs have initiated money transfer services on pilot basis. However, the credit loan and saving service remains the basic and major services, especially loan service being dominant, which is the focus of this data miming research as collection of data that is the major ingredient of data mining research, is available on the loan services.

Among the 20 registered MFIs, six are supported by the regional governments (Amhara, Oromia, Tigay, SSNP, Addis Ababa and Benishangul), while 14 are Non Government based Organizations (NGO), stated in the Sebstad (2003). All the NGO based MFIs are registered as share companies and they are linked to the activities of national and international non-governmental organizations including, for example, world vision, catholic Relief service and Christian Relief and development association. The Wisdom Micro Finance Institution (WMFI) is one of such MFIs, which is highly linked with the activities of world vision.

2.2.3 Wisdom Microfinance Institution (WMFI)

As the case under study is the Wisdom microfinance institution, this section presents an overview of the institution, relevant services rendered at the institution and the automation effort taking place in the organization.

2.2.3.1 Overview of WMFI

WMF is a Microfinance Institution registered as a business entity under the Ethiopian commercial law and proclamation No. 40/1996 to undertake delivery of financial and non financial services to the able poor who are willing, capable and ready to engage in to productive economic activities.

The institution started operations in 1998 as spin-off World Vision in areas where World Vision Ethiopia currently undertakes Area Development Programs (ADPs). It focuses upon delivery of basic services and creation of enabling environment for the cultivation, development and expansion of micro enterprises that create productive employment and income generation for the urban and rural poor.

The institution, WMFI, supports income earning opportunities and helps its clients to achieve food security, and has become very important in supporting the society especially the poor. Products/or services of the WMFIs are central to the well being of the society, to women and children in particular.

As the officials of the institution state, the microfinance institution has shown a rapid growth and expansion that is reflected in terms of an increased number of its

area/branch microfinances and its capital over the last few years since its establishment. As of December 31/ 2004 WMFI was serving 19,912 people, 42 percent of whom were women. In 2004 the institution's active clients grew 57 percent. Wisdoms outstanding 6.757 loans were worth $2.162 million with an average loan size of $ 109 (U.S. Agency, 2005).This significant rate in its outreach and capital indicates that the institution is playing a great role to bring changes to the lives of the society in general. Further the prosperity and increased rate of customers with large collection of data, transaction data, makes it a potential area for application of data mining research. Accordingly, the current research focuses on the customer data available in the organization.

2.2.3.2 Major services/function of the WMFI

Loan and saving services are at the core of any microfinance institution. Likewise WMFIs currently renders significant level of loan and saving services to the ever increasing of its customers. The loan services are based on the different sector areas where the borrowers are to invest. Accordingly, the services WMFI delivers are categorized under the following financial and non-financial service types:

Loan Products

1. Business Loans
2. Enterprise Loans
3. Individual Loans
4. Agri-business Loans
5. Agricultural Loans
6. Consumption Loans

Saving Products

1. Compulsory Saving
2. Voluntary Saving (under pilot)

In line with the objective of the data mining, the research focused on social data of the loan services that contains the product type the customer was rendered in the loan type column.

According to the organizational policy obtained from document analysis and interviews, there is a rule that determine the maximum loan size for each new borrower depending on the group type. However the maximum loan size will increase with some percent as the number of loan cycle increases regardless of the group type. For the first three loan cycles, 1-3 (where he/she is considered as less privileged) the borrower is just allowed to get up to the maximum of fixed amounts similar to the initial service. When the borrower is requesting loan for loan cycles 4^{th}, 5^{th} and 6^{th} times (where he/she is considered as moderately privileged), he/she could be granted loan amount to the maximum of 50% increment on the loan size he/she was given in the earlier cycles (1-3). For the loan cycles greater or equal to 7 (in which case he/she is considered as highly privileged), the borrower is allowed a loan amount to the maximum of 100% increment on the loan amounts he/she was given during 4^{th}-6^{th} cycles of his/ her loan usage.

2.2.3.3 Automation effort at WMFI

WMFI can be regarded as one of the pioneer microfinance institutions in its effort to put in place the Information Technology (IT) solutions for the enhancement of all its operational and managerial activities. Discussions made with IT officers regarding how the operations are handled, how the data and reporting activities are handled reveals this. According to them all the transaction data handlings and report generations made in each department at all area branches are supported by powerful systems. Currently, a Transaction Processing System called Total Micro Financing Solution (TMFS) is used to handle the data such as financial (including disbursement and collection data), balance sheet, income statements and reports. The TMFS is currently being replaced by a more powerful and sophisticated system/software known as Global 1, produced by Indian company Infra soft. The software is on the implementation phase, and as the IT professional say, by the time it is completely put in place it makes every operational and managerial tasks more efficient.

This current movement to implement the most powerful system that highly improves the transaction and reporting process in the organization is on the pilot basis with a financial sponsor obtained from an international fund known as Vision Fund. This all

in all, say the IT professionals in the organization, indicate the remarkable initiation and readiness of the institution to adopt IT solutions for its operations.

However up to the date of the acquisition of the data for the research at hand, the data was available on the TMFS.

2.3 Data Mining Technology (DM)

This section presents the conceptual study made regarding the data mining technology. Definition and overview of data mining, related concepts where data mining is important (such as knowledge discovery, Online Analytical Processing, and data warehousing), and the data mining process are discussed in this section.

2.3.1Definition and overview of DM

Databases today can range in size into more than terabytes -1,000,000,000,000bytes (Two Crows Corporation, 2005). This indicates the existence of massive amount of data collection that has strategic importance (Two Crows Corporation, 2005). Computerization of many businesses, scientific and governmental transactions, advances in data collection tools ranging from scanned text and image platforms to satellite remote sensing systems, popular use of the World Wide Web (WWW) as a global information system, are some of the contributing factors for the availability of large collection of data in different organizations.

The fast growing, tremendous amount of data, collected and stored in large and numerous databases, has far exceeded our human capability for comprehension that requires a power full tool in order to process the data and gain the advantage. To get benefit from the collected data, there should be a way to identify relevant and useful information (Han and Kamber, 2001).

There are a number of definitions for data mining in various literatures each with slight variations.

Witten and Frank(2005) define data mining as the process of discovering patterns in data where the process must be automatic or (more usually) semiautomatic resulting

in patterns that must be meaningful in that they lead to some advantage, usually an economic advantage.

According to Hand, et.al (2001), and Larose (2005), Data mining is the analysis of (often large) observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

Berry and Linoff (2004) define Data mining as the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules.

The one which is essentially relevant for the research at hand is taken from Giudici (2003).

Data mining is the process of selection, exploration and modeling of large quantities of data to discover regularities and relations that are at first unknown, with the aim of obtaining clear and useful results for the owner of the database (Giudici, 2003).

Data mining employs much of the tools and techniques of statistics. However, the data mining is more powerful in that it can do more than the statistical analysis (Berry and Linoff, 2004). For instance some problems may demand learning from experience which cannot be addressed using statistical methods. Moreover statistics usually employs sample data (part of population data thought to be representative) to build statistic models and this methods can miss large body of information about the population while data mining essentially requires larger data (Thearling, 1999).

Data mining typically deals with data that have already been collected for some purpose other than the data mining analysis (for example, they may have been collected in order to maintain an up-to-date record of all the transactions in a bank). This is one way in which data mining differs from much of statistics, in which data are often collected by using efficient strategies to answer specific questions.

Data mining is becoming an essential technology in science and business areas where there is large collection of data. Next section discusses the data mining technology with respect to related concepts.

2.3.2 Data mining and Knowledge Discovery

In most cases, data mining is treated as synonym for knowledge Discovery in Data base (KDD). However, according to Han and Kamber (2001), Geobel and Le Gruenwald (1999), Fayyad, et.al (1996), Pal and Jain (2005), data mining is viewed as an essential step in the process of knowledge discovery in database. According to them the knowledge discovery refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process.

In the view of the above distinction, DM is concerned with the actual extraction of knowledge from data, while the KDD process is concerned with many other activities. Therefore the data mining process contributes to the knowledge discovery in data base in such a way that the results obtained from the actual data mining would be an input for the remaining steps in the knowledge discovery process according to these distinctions.

Due to the popularity of the term data mining than the longer term Knowledge Discovery in Databases, Han and Kamber (2001) favored to adapt and use the term data mining with the broader view of data mining functionality. In this study, however, the terms data mining and knowledge discovery process are both used to refer to the entire process from data collection through pattern identification and deployment and usage of the results. That is just to be consistent with major data mining projects, to use the corresponding experiences, and avoid any confusion between the two phrases, ‘data mining’ and ‘knowledge discovery in databases’.

2.3.3 Data Mining and Data Warehousing

The ability to automate every operational system in a business through wider applicability of computer and Information Technology (IT) resulted in enormous data available in dozens of separate systems. This has brought about the need of integrated system and powerful technology such as data warehousing to deal with the ever increasing data.

Data warehousing is the process of bringing diverse data together from throughout an organization for decision support purpose (Berry and Linoff, 2004).

Data warehouse is an enterprise database from which the data to be mined is extracted. Han and Kamber (2001) also define data warehouse as a subject oriented, integrated, time variant and non volatile collection of data in support of management decision making process. The data warehousing, hence, refers to the process of constructing and using data warehouse.

If the data warehouse already exists in an enterprise, it will be beneficial for the data mining process. A data mining endeavors includes the effort to identify, acquire and cleanse the data. But if the data is already put in terms of data warehouses, most likely, there will be no need of repeating some of the activities such as the data cleaning and integration, for the mining purpose since these are also the essential tasks in the data warehousing (Berry and Linoff, 2004).

According to Berry and Linoff (2004), the data mining converts the essentially inert source of data in to actionable information. Therefore, even if it is not prerequisite to have a data warehouse before the data mining, it would make efforts much easier if the data warehouse already exists as it address the issues of consolidating data from multiple sources, data integrity problems, etc which would be some of the tasks in the mining process also. But this is not much practical in the real world, especially when it comes to developing countries like Ethiopia. This means that having data warehouses is really advantageous for efficient handling of the data and for applicability of such data analysis as data mining as well, however data warehouses remained uncommon because of some constraints. Putting large database up becomes enormous task, taking over a number of years and costing over millions of dollars that makes availability of the data warehouses less common in enterprises.

In fact it is possible to apply the data mining on one or more operational or transactional databases after simply extracting it into read-only database, where the new database functions as data mart. Similarly, this investigation uses a separately created database to which the social data are extracted to make it ready for the mining purpose.

Even though the data mining task may depend on the data warehouse, where it exists prior to the mining task, the data mining has, in other way, brought great influence/initiation for the developments of the data warehouses. Researches (Mento

and Rapel, 2003) indicate that a number of data warehouses have been developed for institutions with the initiations of data mining for diverse information.

2.3.3 Data mining and Online Analytical processing (OLAP)

Online Analytical Processing (OLAP) is the dynamic synthesis, analysis and consolidation of large volumes of multidimensional data (Fayyad, et.al, 1996). According to Pal and Jain (2005), OLAP gives fast, consistent, interactive access to a variety of views of any information and involves many short, update-intensive commands including day to day operations like purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.

OLAP is a popular approach for data analysis; hence it is a primary task of data warehouse system. As to Fayyad, et.al (1996); Han and Kamber (2003), OLAP and data mining are very different tools that complement each other. OLAP is used to answer why certain things are true as hypothesis verification. In this case, the user gives a hypothesis about a relationship and verifies or disproves it with series of queries against the data.

In contrast, a data mining, instead of verifying hypothetical patterns, it uses the data itself to uncover such patterns. In short, the OLAP tools are targeted towards simplifying and supporting interactive data analysis while data mining tools enhance the processes through automating as much of the tasks as possible; it allows the automated discovery of implicit patterns and interesting knowledge hidden in large amounts of data while OLAP is data summarization/aggregation tool that helps simplify data analysis.

Even though data mining and OLAP have different objectives as stated earlier, there are ways in which the data mining process contributes for the OLAP activities. For example, those tasks accomplished in the early stages of the data mining, such as exploring the data, identifying important variables, and understanding the data in general, would make the OLAP more effective and efficient (Two Crows Corporation, 2005).

2.3.4 Data mining and Customer Relationship Management (CRM)

According to Two Crows Corporation (2005), many organizations are using data mining to help manage all phases of the customer life cycle, including acquiring new customers, and retaining good customers.

Microfinances, like any business organizations, have to take in to consideration the customers' behavior and interests so as to be competent in the market. Knowing customers' behavior and interest will help the microfinance institutions target loyal customers and adjust the service policies in order to attract and retain more customers. This has to do with customer relationship management (CRM).

CRM is the business practice that is intended to improve service delivery, build social bonds with customers, and secure customer loyalty by predicting customer behavior and selecting actions to influence that behavior (Prakashi and Kumar, 2000).

According to Edelstein (2002), Customer Relationship Management (CRM) helps companies do a better job of matching products and service campaigns to customers and prospects for improved company's profitability and prosperity. Hence relationship building and customer oriented management are key factors to which company's success or failure is closely linked. Customer management requires the collection of significant amount of data and set up of procedures for interpreting the data.

Hence data mining has tremendous role in improving CRM in any sector wherever huge collection of data related to customer and customer behavior exists. The data mining can help for customer profiling. By determining characteristics of customers (profiling), a company can target prospects with similar characteristics. For example by profiling customers who have bought a particular product, the company can focus attention on similar customers who have not bought that product. Similarly, by profiling customers who have shown up for long period of time (for more loan cycle) the organization can devise a strategy and refine its policy in which case such customers can be prioritized and given due attention in the financial service for banking and microfinance.

On the other way, profiling customers who have left, a company can act to retain customers who are at risk for leaving; because it is far less expensive to retain a customer than a acquire new one.

According to Ruey-Shun, et.al (2005), the systematic application of data mining techniques reinforces the knowledge management process and allows marketing personnel to know their customer well to provide better services. The data mining approach helps users to identify valuable patterns contained in diverse data and their relation so as to help the major decisions.

Many researches have been conducted on the role of CRM in various sectors, especially in banking and finance. For example, a Research on the Application of Customer Relationship Management in Chinese Banking by Guangshi and Ning (2005) reports that CRM will help banks to grasp customer demands deeply thus helps to provide exact financial service products to customers. According to their conclusion, CRM system is helpful to optimize market value chain. It will enable commercial banks to timely catch up with market demands, and attract new customers on the basis of retaining old ones by continuously improving customers' satisfaction and loyalty.

2.3.5 Data Mining (DM) process

Data mining is a process that involves series of steps to preprocess the data prior to mining and post processing steps to evaluate and interpret the modeling result (Han and Kamber, 2001). In 1996, a standardization effort resulted in Cross-Industry Standard Process for Data Mining (CRISP-DM), (Larose, 2005). The Cross-Industry Standard Process for Data Mining (CRISP–DM) that was developed in 1996 by analysts representing DaimlerChrysler, SPSS, and NCR. CRISP provides a nonproprietary and freely available standard process for fitting data mining into the general problem-solving strategy of a business or research unit (Larose, 2005). According to Han & Kamber (2001), the CRISP-DM is another standardization effort related to data mining. Its aim is to define and validate a data mining process that is generally applicable in diverse industry sectors.

According to the Cross-Industry Process for Data Mining(CRISP-DM), the process takes an iterative form consisting of six main phases as shown in figure 2.1 (TWO Crows Corporation, 2005; Berry and Linoff ,2004; and Larose,2005).

.

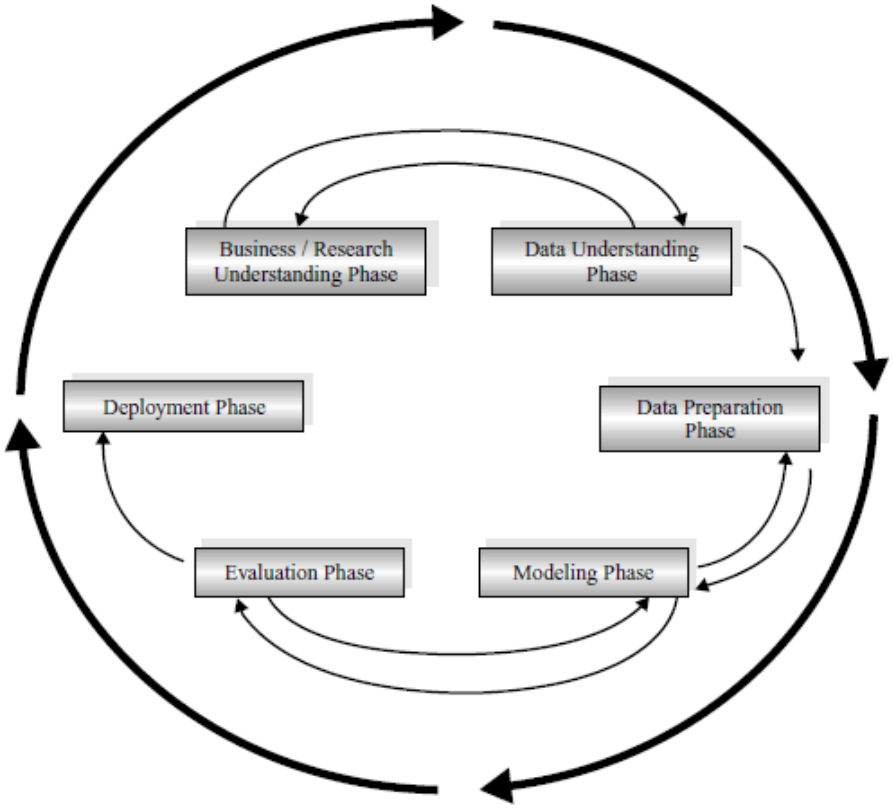


Figure 2.1 Phase of the CRISP-DM process cycle (Larose, 2005).

In figure 2.1, the outer bigger arrow rounding the different phases of the model indicates the cyclic notion of the processes in general. The main phases are interrelated with lines between them. Once the data mining solution is deployed, lessons learned during the process and from the deployed solution calls for new, more focused business question that could be solved through repeated data mining problem. Hence the data mining will be iterative in the business world if the business should gain maximum benefits from the experiences of previous ones. The arrows pointing both left and right directions indicate the most significant dependencies between phases. For example, suppose that we are in the modeling phase. Depending on the behavior and characteristics of the model, we may have to return to the data

preparation phase for further refinement before moving forward to the model evaluation phase.

As the research at hand follows the CRISP-DM model depicted in the figure 2.1, the following sections give the summary of the phases according to Han and Kamber (2001).

2.3.5.1 Business Understanding

This is the first phase in the CRISP–DM standard process which is also termed as the research understanding phase. In this step one works closely with domain experts to define the problem and determine the project goals, identifies key people, and learns about current solutions to the problem. It involves learning domain-specific terminology. This step may also include initial selection of potential DM tools.
Business understanding requires undertaking the following activities:

- Articulate the project objectives and requirements clearly in terms of the business or research unit as a whole.
- Translate these goals and restrictions into the formulation of a data mining problem definition.
- Prepare a preliminary strategy for achieving these objectives.

2.3.5.2 Data understanding

After understanding the business environment data understanding comes next in the CRISP-DM. This step includes collection of sample data and deciding which data will be needed, including its format and size. If background knowledge exists, some attributes may be ranked as more important. Next, we need to verify usefulness of the data in respect to the Data Mining and Knowledge Discovery (DMKD) goals. At this phase data need to be checked for completeness, redundancy, missing values, plausibility of attribute values, and the like.

During data understanding the following tasks are performed:

- Collect the data.
- Use exploratory data analysis to familiarize yourself with the data and discover initial insights.
- Evaluate the quality of the data.
- If desired, select interesting subsets that may contain actionable patterns.

2.3.5.3Data preparation

This step involves preparation from the initial raw data to come up with the final dataset that is to be used for all subsequent phases. This phase is very labor intensive. After the data is acquired and the final dataset is identified then next tasks will be done as part of the data preparation process. This includes the data preprocessing activities such as data cleaning, data integration, data selection and transformation (Han and Kamber, 2001; Berry and Linoff, 2004).

Data cleaning is where the noisy, incomplete and inconsistent data, which are very common in large real world data, are resolved to come up with a complete and consistent data with the noises removed. The original dataset may contain incomplete data for one or more reasons. Attributes of interest may not always be available or other data may not be included simply because it was not considered important at the time of entry. Relevant data may not also be recorded due to misunderstanding or because of equipment malfunctioning or the recording of the history or modification to the data may have been over looked.

Noisy data may arise from faulty data collection instrument used or human or computer errors at the time of data entry, or inconsistencies in naming conventions or data code used. Hence the first step in the data preparation is data cleaning that cleanse the data by filling in missing values, smoothing nosy data, identifying or removing outliers and resolving inconsistencies.

2.3.5.4 Modeling

This step involves usage of the planned data mining tools, and selection of the new ones if needed. Although it is the data mining tools that discover new information, their application usually takes relatively lesser time than data preparation. As stated in Pal and Jain (2005), about 20% of the total effort is spent on the data mining (model building) and analysis of knowledge and knowledge assimilation, 60% is spent on the data preparation, and the rest 20 % of the total effort is spent on business and data understanding.

Two of the most commonly applicable data mining (modeling) techniques are clustering and classification. Classification is a data mining task that helps to build a Predictive model that is trained using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning, because calculated or estimated values are compared with the known results. By contrast, descriptive techniques, such as clustering, are sometimes referred to as unsupervised learning because there is no already-known result to guide the algorithms (Two Crows Corporation, 2005).

Clustering is the task of segmenting a diverse group (records) into a number of similar subgroups or clusters; related records are grouped together on the basis of having similar values for attributes (Two Crows Corporation, 2005). Clustering is said to be unsupervised learning method that will help to identify segments without prior knowledge about the number of segments and the class labels to which the records may be grouped. Hence according to Witten and Frank (2005), the clustering tasks could be applicable where there is no specified class in advance of the data mining, in order to group the items that seem to fall naturally together and also in

search of the clustering indexes that would serve as class labels in any further data mining tasks such as classification.

Classification assumes that there is a set of objects characterized by some attribute which belong to different classes. The organization that is used as the case of this study has a customer data (each record) that is essentially labeled as highly privileged, moderately privileged and less privileged, corresponding to the values of loan cycle attribute. Since there is an attribute (loan cycle) that is used to characterize the customer records in to different segment in advance, the current study is aimed at building a classification model that would potentially help the company to predict the likely status of a new borrower. Therefore brief discussion about the classification based data mining technique is made below.

Classification based Data Mining Technique

Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown (Han and Kamber, 2001). In most cases, the learnt model is represented in the form of classification rules, decision trees, or mathematical formulae that can be utilized to categorize future data samples. There are various data mining techniques for classification task- Decision tree, k-nearest neighbor, Bayesian Network, and Neural network are some of the most commonly used techniques (Han and Kamber, 2001; Zemeke, 2003; Larose, 2005). Since the study at hand employs decision tree technique, decision tee classification is discussed with more emphasis below.

Classification using decision tree

A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions where the topmost node in a tree is the root node (Han and Kamber, 2001).

In order to classify an unknown sample, the attribute values are tested against the decision tree. A path is traced from the root to leaf node that holds a class prediction for that sample. At each level of the tree, the appropriate value would be compared and a decision on which direction to go would be made. This means by navigating the decision tee one can assign a value or a class to a case (sample) by deciding which branch to take, starting at the root node and moving to each subsequent node until leaf node is reached.

Decision tree models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make prediction. The set of rules produced constitutes the predictive learning of the response class/value of new objects, where only measurements of the predictors are known. As an example, a new client of a bank is classified as a good client or a bad one by dropping it down the tree according to the set of splits of a tree path, until a terminal node labeled by a specific response-class is reached (Aurelian and Adrian, 2007).

A number of different algorithms may be used for building decision trees including CHAID (CHi-squired Automatic Interaction Detection), CART (Classification And Regression Trees), C5.0 and C4.5 (Two Crows Corporation, 2005).

CHAID: Chi-Square Automatic Interaction Detection developed in 1980 creates an n way splits and used for categorical variables (Thearling, 1999). The CHAID algorithm uses chi-squared testing to produce an implicit stopping criterion based on testing the significance of the homogeneity hypothesis; the hypothesis is rejected for large values of chi-square value. If homogeneity is rejected for a certain node, then splitting continues, otherwise the node becomes terminal. Unlike the CART algorithm, CHAID prefers to stop the growth of the tree through a stopping criterion

based on the significance of the chi-squared test, rather than through a pruning mechanism (Guidici, 2003).

CART: Classification and Regression Trees creates binary splits and is used for continuous variables (Thearling, 1999). The CART tree model consists of a hierarchy of univariate binary decisions. Each internal node in the tree specifies a binary test on a single variable, using thresholds on real and integer-valued variables and subset membership for categorical variables (Hand, et.al, 2001).

C4.5: C4.5 developed by Quinlan in 1993 is also used for rule induction (Thearling, 1999). According to Witten and Frank (2005), C4.5 is the popular decision tree algorithm, which, with its commercial successor C5.0, has emerged as the industry workhorse for off-the-shelf machine learning.

Depending on the algorithm used, different trees may be built with each node having two or more branches. For example, CART generates trees with only two branches at each node, while others may generate more than two branches at each node. Where there are only two branches at each node the tree is called binary tree, otherwise, if there are more than two branches at each node, the tree is called multi-way tree (Two Crows Corporation, 2005). A classifier algorithm known as J48, which is an implementation of C4.5, is used for the decision tree modeling. J48 is popularly used for its ease of understandability and its being relatively fast and being suitable for categorical data according to Romerio et.al (2003).

Attribute Selection Mechanism

Not all attributes are equally important in classifying given dataset with respect to some target class (Han and Kamber, 2001). Therefore, the issue of which attribute should be considered first, second, third, etc is the basic thing to be addressed in the tree construction step. According to Han and Kamber (2001), most decision tree induction tools adopt attribute selection measure known as information gain to

determine the precedence of the test attributes among the candidate attributes given in the dataset.

The attribute with the highest information gain 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 selects the least randomness or impurity in these partitions. Such an information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that a simple tree is found.

According to Han and Kamber (2001) information gain of an attribute A is computed as:

$$\text{Gain(S, A)} = \mathit{Entropy}(S) - \sum_{v \in \mathit{Values}(A)} \frac{|S_v|}{|S|} \mathit{Entropy}(S_v)$$

Where Values (A) is the set of all possible values for attribute A, and S_v *is* the subset of S for which the attribute A has value *v*.

Entropy(S) is a measure in the information theory that characterizes impurity of a collection S. If the target attribute takes on *c* different values, then the entropy S relative to this c-wise classification is defined as:

$$\text{Entropy(S)} = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Where p_i is the proportion/probability of S belonging to class *i*. Logarithm is base 2 because entropy is a measure of the expected encoding length measured in bits.

Pruning Decision Tree

The decision tree algorithms operate in two phases; the construction and pruning phases. The construction phase of decision tree usually results in a complex tree that often 'over fits' the data, reducing its accuracy. A tree T 'over fits' if there is another

tree T that gives higher error on the training data, yet gives lower error on test data. This will especially happen if the training data has a few erroneous instances in it. Hence the construction phase must be followed by another task that turns the tree to more understandable, usable reduced form. This is achieved through what is known as tree pruning (Ahmed, 2007; Gidole and Vydiswaran, 2003).

The Pruning phase of decision tree is the process of removing some non- promising branches to improve the accuracy and performance of the decision tree. There are two approaches in tree pruning: pre-pruning and post-pruning. In pre-pruning approach, a tree is pruned by stopping its construction by deciding not to further partition the subset of training data at a given node. Consequently, a node becomes a leaf that holds a class value with the most frequent class among the subset of samples. Pre-pruning criteria are based on statistical significance such as information gain (Ahmed, 2007).

On the other hand, as stated by Ahmed (2007), Post-pruning removes branches from the completely grown tree, by traversing the constructed tree and uses the estimated error or confidence threshold to decide whether some undesired branches should be replaced by a leaf node or not.

This pruning will result in a tree that is more understandable, easy to be interpreted, with less over fitting.

2.3.5.5 Evaluation

The classification model created in the modeling phase is evaluated for quality and effectiveness before deploying it for use in the field. Here it is decided whether the model in fact achieves the objectives set for it and therefore come to a decision regarding the use of the data mining result or the need to undertake further research work to enhance its efficiency.

Error rate, false positives and false negatives are some of the evaluation methods of predictive classification model in the context of the decision tree algorithms such as J48 (Larose, 2005). WEKA provides us with a matrix of the correct and incorrect classifications made by the algorithm, termed as the confusion matrix. The negative

classifications that were made in error are said to be false negatives. That is, a false negative represents a record that is classified as negative but is actually positive. Where as a false positive represents a record that is classified as positive but is actually negative. The overall error rate, or simply error rate, is the sum of the false negatives and false positives, divided by the total number of records. Hence, according to Larose (2005), using error rate, false positive rate, and false negative rate, analysts may compare the accuracy of various models.

2.3.5.6 Deployment

This is the last step in the process of data mining tasks, and it is in the hands of the database owner. It consists of planning where and how the discovered knowledge will be used. The application area in the current domain should be extended to other domains within an organization. A plan to monitor the implementation of the discovered knowledge should be created and the entire project needs to be documented.

2.4 Review of Related works

Kim (2000) states that data mining can be applied in any organization that has large collection of data with keen interest to explore the possibility of hidden knowledge that resides in the data.

Owing to existence of large data and advancements in technology adoption, however, some industries have already made significant progress in the application of data mining, more widely than the others. Examples include banking and finance, airline industries, marketing, medical and insurance. These are some of the most common sectors where data mining is widely applied for different purposes such as customer analysis or customer relationship management (CRM), Risk analysis, predictions and fraud analysis.

As stated by Singhal (2007), data mining has been effectively applied for financial data analysis where the data mining techniques are used to undertake loan payment and customer credit policy analysis in the banking service (such as credit and investment services).

Seifert (2004) states that several instances of the application of data mining techniques are also emerging in the banking and finance industry and in the Microfinance industries as well which is the focus of the research at hand. In banking and finance, the data mining techniques are frequently applied to assign a score to a particular customer or prospect indicating the likelihood that the individual will behave in a particular way (Dass, 2008). Similarly, Microfinance institutions have got a wide application of data mining for the various purposes. According to Dass (2008), credit scoring and customer classification and segmentations are the most common data mining application areas. The following sections discuss summary of the most related works in relation to the customer relationship management and financial areas.

2.4.1 Review of Related works for Customer Relationship Management

A number of researches have been conducted using different data mining tools and techniques on various areas within the country and abroad. In this section the most importantly related works on application of data mining are reviewed.

Gashaw (2004) has conducted data mining research in order to support customer Insolvency Prediction at Ethiopian Telecommunication Corporation. His objective was to build a classification model that can classify customers as potentially solvent or insolvent to support decision making process on the area. On his study he used the neural network back propagation algorithm, and MATLAB 6.5 neural network toolbox. The model accuracy registered on the test data set shows 90.74%. The researcher recommends that further researches can be done for similar area with different variable sets or using other techniques like decision tree to investigate on

whether better results may be achieved for customer insolvency/churning related problems.

Denekew (2003) has conducted research on assessing the application of data mining techniques to support Customer Relationship Management activities at Ethiopian Airlines. His objective was to assess the application of data mining techniques to support CRM activities at Ethiopian Airlines. The clustering technique was applied on the collected dataset consisting of mainly the customers' demographic data and trip related data that was available at the Ethiopian Airline's database (ShebaMiles). He used K-means clustering algorithm that segmented individual customer records into different clusters based on the features pertaining to the demographic and trip information. He then classified those customers using J48 and PART algorithms to develop a model that assigns new customer records into the corresponding segments. His model resulted in an accuracy of 96.575%.

In his recommendation part, Denekew has stated that data mining techniques such as classifications and clustering should continuously be employed in areas where customer records extremely increases and gets complex through time, to come up with improved , refined and hence more meaningful and useful models that better support the CRM of the organization.

In addition to the above areas, the data mining technology has got many potential applications in other areas such as banking and finance for customer relationship management, for risk analysis, fraud detection, etc. As far as the banking and finance areas are concerned data mining has gained widespread attention and increasing popularity. Since the research at hand is exploring the potential application of data

mining in the microfinance, the next sub-section reviews the possible applications of data mining in this area.

2.4.2 Related works on Application of Data mining in financial Institutions

Forecasting stock market, currency exchange rate, bank bankruptcies, understanding and managing financial risks, trading futures, credit ranking, loan management, bank customer profiling, and money laundering analysis are core financial tasks for data mining (Boris, N.D.). Some of these tasks are similar with data mining for other fields.

Financial institutions produce huge datasets that build a foundation for approaching these enormously complex and dynamic problems with data mining tools.
According to Boris (N.D.), most of the data mining methods and techniques can effectively be applied in financial modeling. Some of these data mining techniques commonly applied in the financial modeling task includes k-means and hierarchical clustering and classification techniques such as k nearest neighbor, decision tree analysis, regression, Bayesian learning and multi layer neural networks.

In financial institutions such as banks and microfinances, marketing management is one of the best application areas (Kim, 2000; Doming, N.D.). The data mining allows identifying and predicting each individual customer behavior-his or her response, sensitivity or elasticity to different marketing stimuli so that one can carry out a finite market segmentation analysis and provide customized offering to target customer (Kim, 2000). According to Kim (2000), specific marketing applications of data mining include Customer Relationship Management (CRM), Market basket analysis, and cross selling campaigns.

Kim (2000) further states that the application of data mining in financial institutions can be used to predict the value of financial future in real time basis, which is an essential contract that allows someone to buy a commodity at certain price on the

certain data in the future. In addition, he states that data mining can be used to deal with the problem of attrition, loss of customer to competitors.

To exploit such benefits of data mining, various researches have been conducted in the application of data mining in financial institutions such as banking and finance.

Askale (2001) has conducted a research on the possible application of data mining technology in supporting loan disbursement activity at Dashen bank S.C., Ethiopia. The objective of the research conducted by Askale was to develop a model that supports the loan decision making process that would contribute in alleviating the high default rate in the company. The research focused on developing a classification model for the customer's repayment behaviors (Regular, Loss or Default) that in turn could support the credit risk assessment. The research came up with a predictive model that could help predict whether a potential borrower would default or not, which further guides in assessing the credit decision processes. Hence it was mainly based on collection data (credit report) and loan approval data pertaining to the individual borrower of the bank for extracting predictor features for the dependent class called repayment behaviors. The features mainly considered include security types (building, vehicle, personal guarantee, etc) and collection related attributes such as term of repayment, month/date of loan disbursement, as inputs for the model building.

Neural network back propagation algorithm was employed for the modeling and the researcher concludes that the model, with 88% accuracy proved potential applicability of the data mining in the area of financial institutions.

The researcher recommends that other data mining techniques such as decision tree classifications would also be applied in financial areas to effectively utilize the borrowers' data for supporting loan decisions with respect to different purposes such as customer relationship management in addition to credit risk assessments.

Koyuncugil and Ozgulbas (2008), from Turkey, have explored the possibility of designing data mining techniques for financial institutions using data obtained by means of financial analyses of balance sheets and income statements of companies under Turkish Central Bank. They came up with a model for detecting financial and

operational risk indicators of Small and Medium Enterprises (SMEs).The study focused on creating segmentation model using decision tree algorithm for customer profiling where the method of CHAID (CHi-Square Automatic Interaction) is applied as defined in the scope of their study. They reported that the decision tree approach using the CHAID method helped to construct best model with acceptable accuracy that could be used to detect financial and operational risk indicators of SMEs. And their recommendation states that further data mining researches can be conducted utilizing the potentially useful customers' data through application of different classification techniques such as decision tree and neural network for different mining objectives.

Aggelis (2005), at University of Patras, Department of Computer Engineering and Informatics, Greece, has conducted a research on Predictive Model in Electronic Banking Data. He employed a stepwise linear regression data mining method that models a response value (daily logins) with Predictor variables such as daily Athens Stock Market Rate (ASMR) and daily Bank Share Value (BSV). The resulting model was aimed to help in prediction, decision making and design of the Bank policy. He concludes that the model is an acceptable one since it registers an accuracy of 78%.

Aggelis has recommended that further researches could be made to come up with enhanced predictive models using features that he did not consider; external sources like inflation rate, oil price and others. Furthermore, he has recommended that various data mining researches could be conducted to develop useful models including other features of the banking and finance data such as number of transactions, number of active users, and others.

Sirikulvadhana (2002), from Swedish School of Economics and Business Administration, has conducted research on data mining as financial auditing tool with the objective to determine if data mining tools can directly improve audit performance. The data mining employed clustering analysis to develop cluster mode. The model is aimed to assist auditors to select the samples from some representatives of the groups categorized in the way that they have not distinguished before and the obviously different transactions from normal ones or outliers. An accounting transaction archive with vast amount of datasets was used for the data mining process. In his conclusion he states that there are some interesting patterns

discovered automatically by data mining technique, clustering, that cannot be found by generalized audit software this proving the vast significant application of the technology in the area namely banking and finance and micro finances as well.

The research has shown an encouraging result of an application of data mining techniques to assist auditors. Further, the researcher recommends that other techniques such as classification mining for risk assessment and customer churning problems could be further research areas worth conducting to come up with models useful to the owner of the data.

.

CHAPTER THREE
THE DATA MINING TECHNIQUE

3.1 Introduction

The aim of this study is to build a predictive model that could help the organization in predicting new customers' status that supports loan decisions. To this end, CRISP-DM (Cross-Industry-Standard-Process for Data Mining) process cycle (CRISP-DM, 2000) is followed. Therefore this chapter discusses the data mining technique, particularly decision tree technique, used in the model building process.

3.2 Model building technique-Decision tree

A classification modeling technique with a decision tree algorithm is used to build a model that classifies customer records in to a defined target class. As described earlier, based on discussions made with domain experts, the target class is identified to be Customer status and the model was built based on 9,550 instances.

The decision tree is used for a number of advantages. It is easy to explain and to interpret, operations are completely interactive and also they have powerful visualization features. Moreover, the decision tree can handle large number of variables with either continuous or discrete dependant variables.

A J48 decision tree, which is an implementation of C4.5 decision tree algorithm, is used in the modeling experiment. Hence below gives description on the decision tree algorithm that takes a processed data known as model set to produce patterns. Once the pattern (model) is produced it can be used to classify a previously unseen record taking the selected part of its features, with respect to the target class identified in the early stages of the data mining, data preparation.

The J48 decision tree algorithm works in two different phases: a training phase, where the decision tree is built from the available instances, and the testing, or performing phase, where new instances are classified using the constructed model.

3.2.1 The decision tree algorithm

Definition:

We assume there is a database D that contains tuples t_1, t_2, … ,t_n (n is the number of dataset) with attributes a_1,a_2,…,a_m (m is number of features) and target class C={C_1,C_2,C_3} where C_1 is less privileged, C_2 is moderately privileged and C_3 is highly privileged class values. Then the decision tree is a tree associated with the customer data D, each internal node labeled with either of attributes a_1,a_2,…,a_m, each edge labeled with a predicate that is applied to the attribute associated with parent node, and each leaf node is labeled with either C_1, C_2 or C_3.

The decision tree algorithm, therefore, generates a tree using the training tuple of data partition from D and selected attributes by applying attribute selection method on the set of candidate attributes. The algorithm functions as follows:

DT (Instances; Decision attribute; Attributes)
Start by creating a root node N,

- **If** all tuples in the dataset have the same class value C, then return the root node N as leaf node with class C
- **If** attribute list is empty-there is no candidate attribute selected, then return N as leaf node with the majority class value in the dataset D.

Else **BEGIN**

- Compute Information gain ratio for each attribute
- Take the attribute with maximum information gain ratio as the ***decision attribute*** for the root.
- For each possible value V_i in the list of ***Attributes*** of A,
    - If A is Nominal, Add a new tree branch below *Root*, corresponding to the test $A = v_i$

- Else If A is numerical Add two new branches corresponding to the tests $A \leq tA$[1] or $A > tA$,
- Let Instance V_i be subset of ***Instances*** that have values V_i for A

if Instances(v_i) is empty **then**

Below this new branch, add a leaf node with label = most Common value of Decision attribute in Instances

else

Below this new branch add the sub tree

DT (Instances (v_i); Decision attribute; Attributes-{A})[2]

End

End

End

End

End BEGIN

As indicated in the decision tree algorithm, as to which attribute should come first, second, etc is determined by the computation result of the information gain ratio obtained at each level. At each node, attributes are evaluated on the basis of separating the classes of the training examples. Information gain ratio function is used for the evaluation purpose.

The information gain ratio is just a modification of information gain that reduces the effect of attributes with a large number of values. According to Yang, et.al (N.D.) Information gain is biased towards choosing attributes with large values which creates over fitting (selection of attribute that is non optimal for prediction). For example, if id of a customer is considered, the id could have larger number of different values than any other attribute in the dataset as a result of which the entropy of split becomes zero and information gain becomes maximal. Therefore according

[1] tA is a randomly selected value between the lower and upper bound of a numerical attribute A

[2] Repeat the tree generation process (with the data set D, Information gain ratio computation for each attributes, with the early selected attributes removed).The recursive partitioning stops when Samples for a given node belong to same class, or there are no remaining attributes or there are no samples for the branch

to Yang, et.al (N.D) the information gain value the id would be the selected splitting attribute. However the information gain ratio corrects the information gain by taking the intrinsic information of a split in to account (how much information we need to tell which branch an instance belongs to).

At each node, the information gain ratio is calculated in terms of the information gain and entropy following the following three steps.

Step 1: Compute Entropy of S, note that, according to Bahety (2009), Entropy(S) is a characteristic measure of impurity of an arbitrary collection of examples

$$\text{Entropy}(S)=\sum -p(I)\log_2 P(I),$$

Where, P (I) is the proportion of S belonging to class I.

Step 2: Compute Information gain

Given example set S on attribute A, the information gain is defined as:

$$Gain(S,A) = Entroyp(S)- \sum |S_V| \diagup |S_V| \; Entropy$$

Where, S_V =subset of *S* for which attribute *A* has value *v*

Step 3: The information gain ratio of sample S on attribute A is calculated as:

$$GainRatio\ (S,A)= \frac{Gain(S,A)}{SplitInfo(S,A)}$$

Where

$$\text{SplitInfo(S,A)} = I[\frac{S_1}{S},\frac{S_2}{S},...\frac{S_m}{S}]$$

and $S_1, S_2, \ldots, S_m$ are the partitions induced by attribute A in S

The attribute with the highest information gain is chosen to test attribute for the current node. Hence the algorithm iteratively computes the information gain ratio at each node to determine the best attribute for splitting criteria. Once the tree is fully constructed considering all the training data set and all attributes exhaustively, then tree pruning is applied to improve the tree.

The decision tree algorithm applies a type of post pruning (Berry and Linoff, 2004). The algorithm generates a fully grown tree and then carries out pruning as post processing step. That means, after the decision tree is produced by the divide and conquer[3] algorithm, it prunes the tree in a bottom up pass.

As also described in Witten and Frank (2005), the J48 algorithm employs either of the two methods in its pruning process; sub tree replacement or sub tree raising. It decides which of the methods to perform, or not to perform any, at each of the nodes in the constructed tree. In sub tree replacement, the method selects some sub trees and replaces them with single leaves reducing the number of tests along a certain path. In sub tree raising, a node is moved upwards towards the root of the tree, replacing other nodes along the way.

To make actual decisions about which parts of the tree to replace or raise, the classifier uses error rates or confidence threshold (which is set to 25% by default). The confidence threshold helps the classifier to determine how specific or general the model should be. Hence, if the threshold is set to lower value (than 0.25), more pruning is performed resulting in more generalized tree. Where as if the threshold is set to greater value (than 0.25), less pruning is performed resulting in larger tree.

3.2.3 Data Evaluation

There are various test options in order to evaluate the predictive model built. Even though it is possible to make use of percentage split like by holding out some percent, say 30% for test and rest 70% for training, more standard way and more appropriate one for such limited data set is to use cross validation method. Accordingly a stratified 10-fold cross validation method is used for testing the model. In stratified 10-fold cross validation method, the data set is randomly divided in to 10 parts in which the class is represented in approximately the same proportion as in the full data set. Each part is held out in turn and learning scheme is trained on

[3] The algorithm partitions the training dataset until every leaf contains cases of a single class or until further partitioning is impossible.

the remaining nine-tenths and the execution is repeated 10 times after which the overall error rates is taken as an average.

Apart from data evaluation done by the WEKA s 10 fold validation mechanisms it is necessary to conduct external validation type that is done by inviting the experts on the area to suggest on the model built by comparing its significance with the current system. Accordingly the rules generated are given to the accountants/clerks working on the loan approval and disbursement area to judge the predictive worthiness of the model.

3.3 Experimental Design

The modeling process consists of a series of six experiments. The first experiment involved building a classifier on the entire dataset and entire attributes prior to any application of feature or data reduction mechanism because it would serve as a base line upon which to evaluate all subsequent results. In the next experiment, again the entire dataset is used but the attributes reduced to some top ranked ones (9 attributes) after applying the WEKA attribute selection mechanism that implements information gain filter. The third experiment is made by again reducing the attributes while the attribute selection is based on the WEKA ranking mechanism using information gain. Experiment Four is conducted by including some expertly selected attributes to those top ranked ones that are used in experiment three. The fifth experiment considers another dimension, which is using the loan sizes in figure form as given in the dataset.

In the sixth, last, experiment a modification is made on the sixth experiment that showed better performance than any of the previous ones. In this last experiment, the classifier confidence factor is set to 0.15 to apply more pruning than with the default value, to see if the overall performance and acceptability of the model improves.

CHAPTER FOUR
EXPERIMENTATION

4.1 Introduction

This chapter presents the most essential part of the research. In the first part, it presents the necessary preprocessing activities performed for the experimentation. The second part presents the major experiments run, interpretations and their performance evaluations, the pruned J48 tree and the essential rules generated from the tree based on the selected model.

4.2Data Understanding and data preparation

Data understanding and data preparation are some of the primary tasks that highly determine the data mining results. The model built mainly depends on how thoroughly and carefully the necessary data is obtained, analyzed and preprocessed. Hence the next subsequent sections present the data understanding and the essential preprocessing tasks performed to this end.

4.2.1 Data Understanding

Since the data available, in most cases, is generated from day to day transactions collected for some other purposes like for administrative purpose, the existing data situation should be studied to decide on the relevant aspects of the data and to get understanding of the data nature.

The dataset was obtained from the head office of the WISDOM Microfinance Institution which is regarded as social data and was collected from the different branches over the years 2006-2008. The dataset was originally present in various excel sheets as they were collected up from different areas. Hence the initial primary task was to integrate the data in to one repository (single excel sheet) as a result of which a total of 9,715 records were collected for preprocessing tasks.

The dataset in general consisted of 16 attributes even though some of the attributes have got a number of missing values, noises and inconsistencies to be handled in the data cleaning part of data preparation.

The attributes of the dataset with their data type and descriptions are given below.

Attribute	Data Type	Description
ID	Text	Identification number of borrower
Customer Name	Text	Full name of the borrowers
Group Type	Text	Type of the group
Loan Type	Text	Type of the product/service
Sex	Text	Sex of the borrower
Age	Number	Borrowers' Age
Loan Size	Number	Amount of money borrowed
Loan Cycle/Status	Number	Number of times borrowed
Sector	Text	Borrowers sector of engagement
No.Of Emp.	Number	Number of Employee
No.Of Children	Number	Number of children of borrower
OVC	Number	Orphan Volunteer Children
Area	Text	Whether the customer is from ADP(Area Development Program) or Non ADP
Location	Text	Whether the borrower is from Urban, Semi urban or Rural
No.Of Male Emp.	Text	Number of Male Employee
No.OF Female Emp	Text	Number of Female Employee

Table 4.1 Attributes of borrowers social data

4.2.2 Data Preparation

The data preparation includes any of the processes applied on the extracted data in order to make the data more suitable for the experiment to improve the overall data mining task. To this end, only most important preprocessing tasks were accomplished at this step. These include data selection, data cleaning and data aggregation/summarization as indicated below.

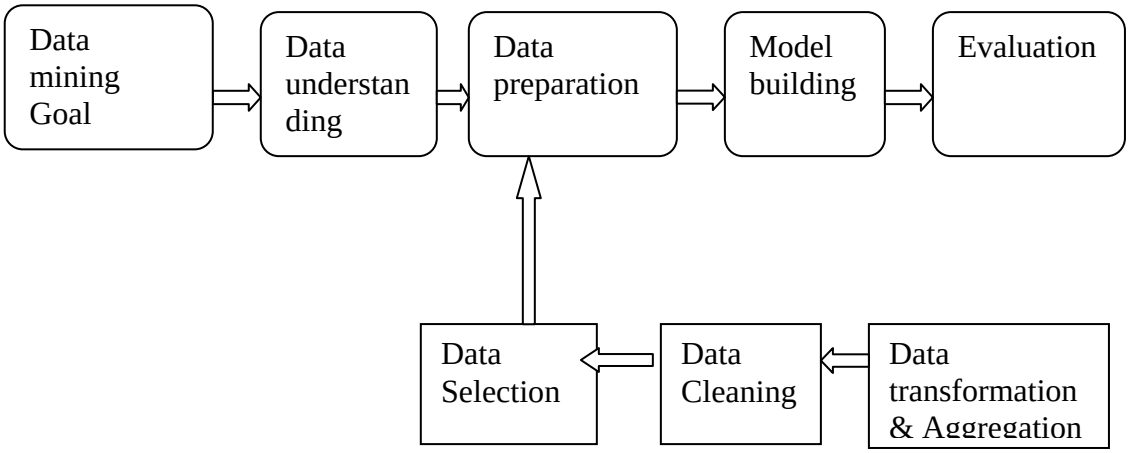


Figure 4.1 The Data Preparation Phases taken from CRISP-DM

Data Selection

At first data reduction by eliminating unnecessary or less important attributes from the original data need to be performed. This is done based on the objective of the study at hand. Hence the id number, customer's names and branch code attributes are removed in order to reduce the data to only most important ones; this would minimize the effort required for further processing.(A snapshot of the dataset used is shown as appendix I).

Among the list of attributes indicated in table 4.1 the id, name and branch code are found to be less significant or less important for the data mining task; hence they are removed and the rest, listed below, are kept for further preprocessing.

- Group Type
- Loan type
- Sector
- No. of Employee
- No.of Children
- OVC
- Area
- Location
- No.of Male Employee
- No.of Female Employee

Data Cleaning

Once the attribute selection has been done, the data cleaning is applied on the dataset with selected attributes.

Among the total dataset extracted, 165 of them have missed values for attributes such as Number of Employee, Children, OVC, Area, Location, Number of Male Employee, and Number of Female Employee. Instead of filling values on these attributes, it was found easier and more logical to remove the records that make up 1.69% of the dataset. As a result, the remaining 98.31% of the original dataset which amounted to 9,550 records were kept for further processing.

Filling in missed values, removing inconsistencies and noises were major data cleaning activities done at this stage of data preparation.

There were values missed in some of the fields. Among the various options to fill in the missed values, filling with the most probable value was used. Consequently, 43 missed values from the field “Number of children” were filled in with “0” value, 8 missed values for the field “group type” were filled with ‘Soliditary’ and 6 missed values from the field “sector” were filled in with “Agriculture”. These values were

considered to be most probable because they are values with highest modal in the original dataset.

Most of the attributes, except those with numeric values, have got their values miss spelt that create inconsistencies and noises harnessing the data mining result if used directly. Hence it was crucial to carefully inspect and correct the wrongly given values in all the dataset. For example, in the column “Group type” there were different ways of inputted values as “Individual”, ”IndividualN”,” IndividualND”, etc, just to mean the same value ”Individual”. In the column “Loan type”, there were about 7 variations of inputted values only to mean the same thing “Agriculture”. Similarly the dataset had a bulk of inconsistencies and noises in other attributes as well. Therefore removing such inconsistencies and noises by turning the variations in to uniform correctly spelt wording with the frequent consultations of domain experts were the essential tasks performed at this step.

Data Transformation and Aggregation

Transformations and aggregations of the data help to minimize the variations of the attribute values in some of the fields and also to make results more meaningful and easily interpretable.

The varying ages were transformed in to more aggregated values “young”,” adult” and “old ages” for the different age groups determined based on experiences and consultation with the domain experts. Since the dataset doesn’t contain any instance with age below 17, there was no need to include another category like “Children” even though this was also logical and permissible categorization for age. Hence young is considered to be between 18-30, while adult is considered to be between 31-50 and old age was considered to be those with ages 51 and above.

Similarly, to make it easier for discussions and interpretations of the result, other fields like loan size and loan cycle/status were generalized in to categorical values. Loan amounts below 1000 are categorized as Low, those which are between 1000 and 10,000 were categorized as medium and those loan sizes above 10,000 were categorized as high.

For the same reason stated above and also because the research is characterizing customers with respect to loan cycle/status that have direct connection with the degree of privilege the customer may have, it was necessary to turn the numeric values of Loan cycle/status in to general categorical values such as highly privileged, moderately privileged and less privileged. This was done based on consultations with the domain expert and based on the business rule of the organization. Accordingly, values from 1-3 is considered as less privileged, from 4-8 moderately Privileged and values with 9 and above are considered to be highly privileged.

According to the institutions' rule it is defined that when a customer is borrowing for large number of cycles (like 9 and above), he/she will be assumed to be highly privileged that is manifested by allowing larger loan amounts with more loan frequencies as an incentive. But there is no system or way of predicting if a new customer may stay borrowing loans for large number of cycles in order to assume him/her as highly privileged and allow him/her more loans with more frequencies. Hence the data mining research builds a classification model that helps to predict if a new customer will be highly privileged, moderately privileged or less privileged based on other essential factors.

4.3 Running the Experimentations

Several subsequent tasks are performed in each experiment utilizing the WEKA software for the model building process. Figure 4.2 depicts some of the major tasks performed during the experimentation.

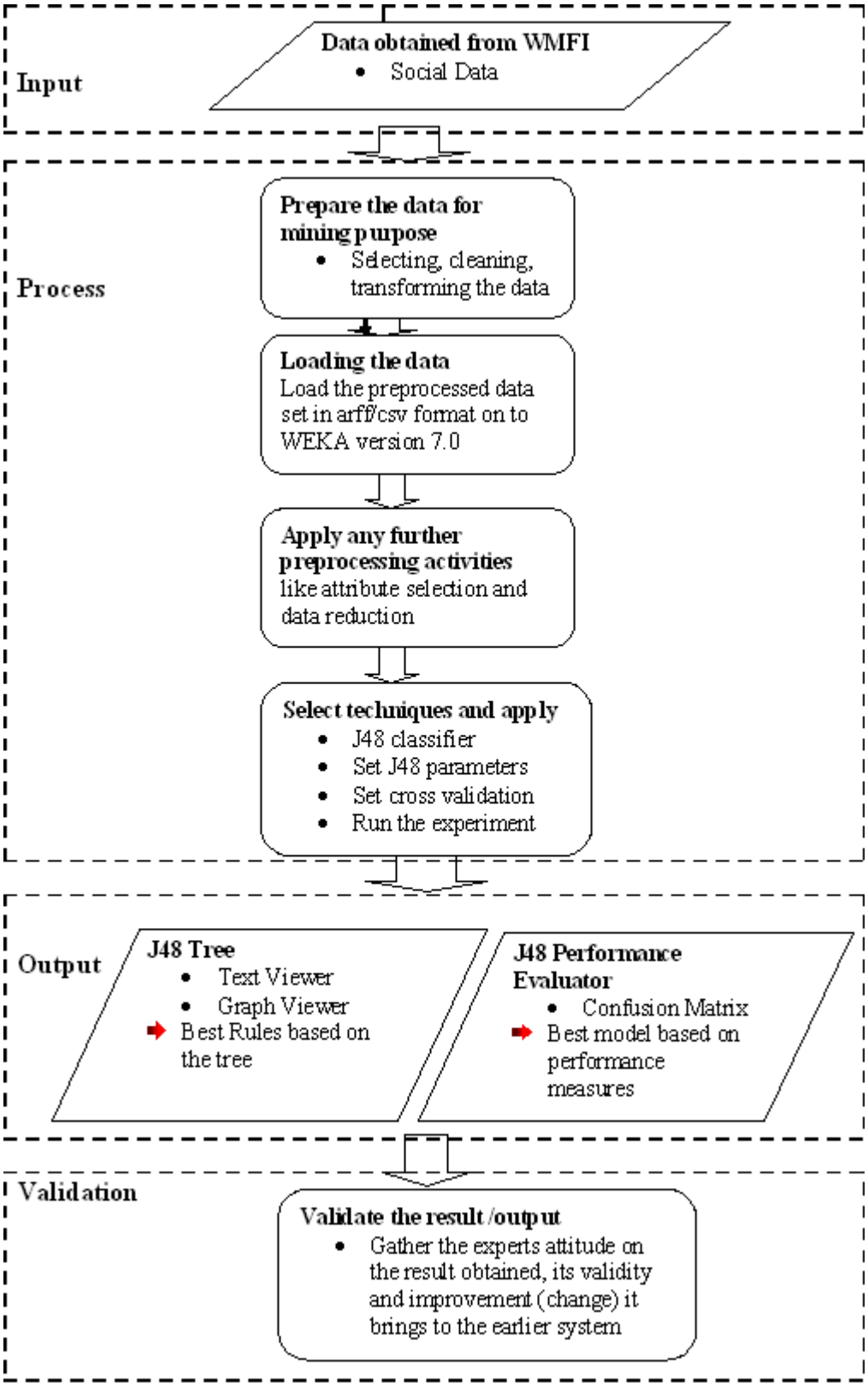


Figure 4.2 Components of the Experimentation made.

In the earlier section, all the preprocessing activities performed on the dataset were presented. This section focuses on presenting summary of the major experiments made in the process of arriving at the optimal model to achieve the objective set in chapter One.

4.3.1 Input data

Once the necessary data is passed through the preprocessing activities as described in the earlier sections, it is then loaded to the selected software for the model building required. The preprocessed data is converted using a spreadsheet program that is suitable for WEKA software. The result of the conversion is Attribute Relation File Format (ARFF format) which is then loaded directly by the WEKA's Explorer.

4.3.2 Experiments Run

Series of Experiments are conducted based on which classification tree models with varying accuracies, sizes and number of leaves are obtained.
This section presents several activities done related to running and evaluating model building experiments, selecting the best and appropriate model, and providing explanations on the selected model.
For building the classification model, J48 classifier is used.

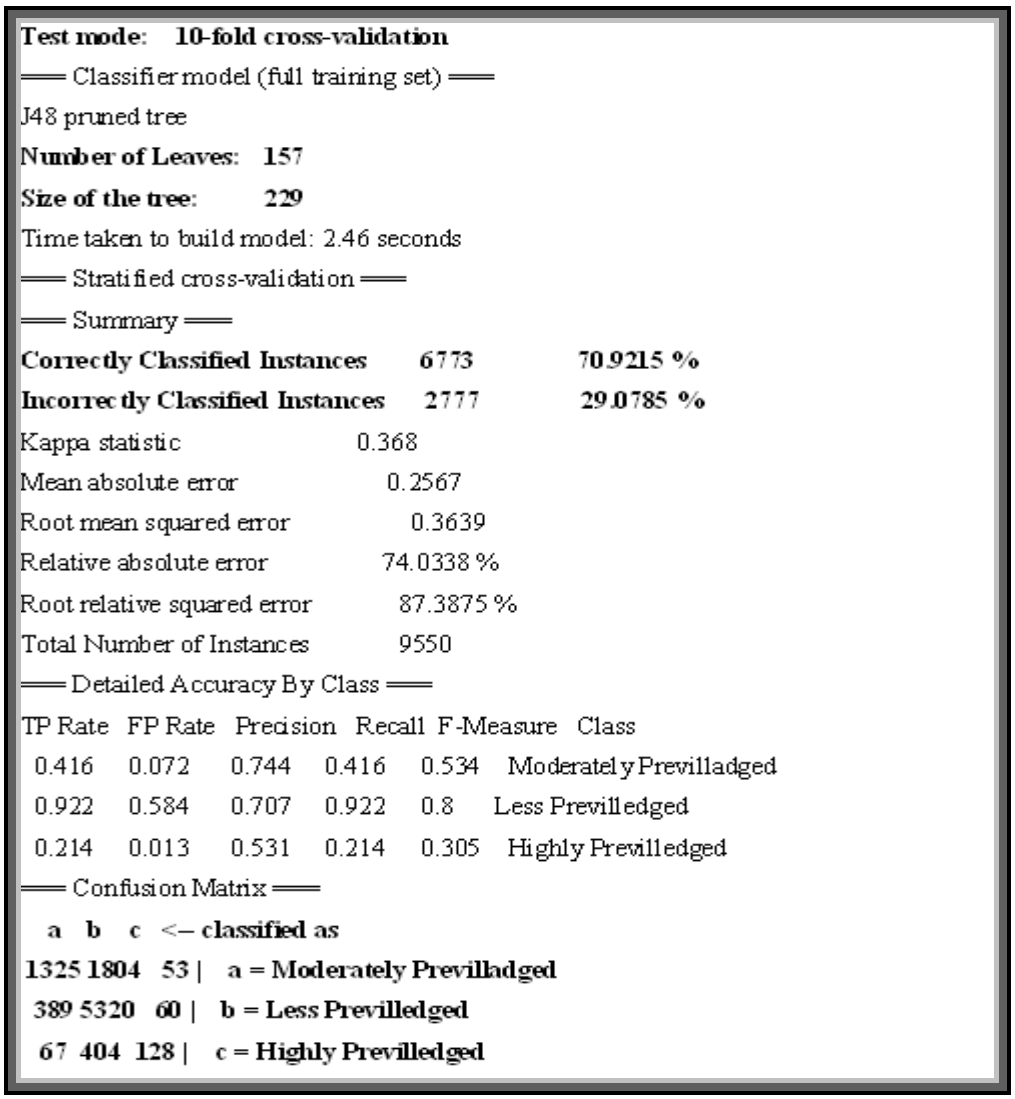
4.3.2.1 Experiment One

In this experiment a classification tree model building was done using J48 Classifier. All the dataset, 9550 instances, with all of the predictor attributes, 13 attributes were used in this experiment. The decision tree model finds the thirteen variables to be influential on the response variable borrower status, with "Loan size" being the root node of the tree. This means that loan size that the borrower wants/or borrows is the most important factor to determine the borrower's status whether he/she is highly privileged, moderately privileged or less privileged.

The results of this experiment serve as a base for classifier performance up on which to evaluate all results of subsequent experiments conducted with modifications of input variables such as modification of dataset size, incorporation of feature selection

(automatic feature selection) or by manually selecting combinations of expertly selected attributes, and modification of classifier's parameters.

The experiment yields an accuracy of 70.9215%, 157 leaves and 229 total nodes (size of the tree). Figure 4.3 is the run information.



Test mode: 10-fold cross-validation

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

J48 pruned tree

Number of Leaves: 157

Size of the tree: 229

Time taken to build model: 2.46 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	**6773**	**70.9215 %**
Incorrectly Classified Instances	**2777**	**29.0785 %**
Kappa statistic	0.368	
Mean absolute error	0.2567	
Root mean squared error	0.3639	
Relative absolute error	74.0338 %	
Root relative squared error	87.3875 %	
Total Number of Instances	9550	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.416	0.072	0.744	0.416	0.534	Moderately Previlladged
0.922	0.584	0.707	0.922	0.8	Less Previlledged
0.214	0.013	0.531	0.214	0.305	Highly Previlledged

=== Confusion Matrix ===

a	**b**	**c**	**<-- classified as**
1325	**1804**	**53**	**a = Moderately Previlladged**
389	**5320**	**60**	**b = Less Previlledged**
67	**404**	**128**	**c = Highly Previlledged**

Figure 4.3: Result of experiment One

As depicted in the confusion matrix (fig 4.3), the classifier assigns A,B and C designations to the target classes according to their alphabets: A for Moderately Privileged, B for Less Privileged and C for Highly Privileged classes.

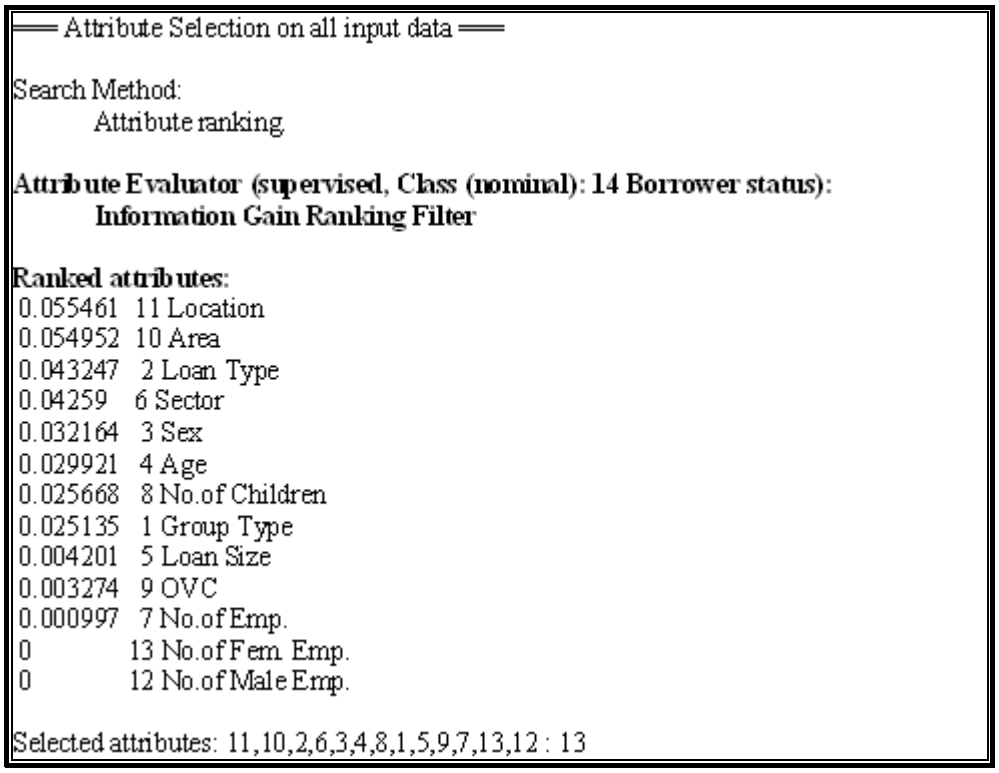
From the confusion matrix 389 instances are misclassified as class A which were actually class B, 60 instances were misclassified to class C which are actually class B, giving total misclassification of 419(389+60) of class B instances.
67 instances and 404 instances of class C were misclassified in to class A and class B respectively. However larger number of misclassifications were observed for class A instances;1804 instances and 53 instances of the class A are misclassified as class B and class C respectively, giving total of 1857 misclassifications of class A instances.

Even though the result of this experiment seems to be promising as far as accuracy is concerned it is worth running different experiments with different input in order to see if there could be a better model with more acceptable performance such as decreased size(total nodes) in addition to improvements in accuracy. Accordingly, Experiment Two was performed with some input modifications such as attribute reduction where the attributes are selected based on the information gain calculated by the WEKA software.

4.3.2.2 Experiment Two

This experiment is done on the whole dataset with reduced number of attributes to see if the accuracy level of the classifier and the complexity of the model would be improved.

To select some attributes from the total attributes, the whole dataset was loaded to the WEKA software and after choosing SELECT ATTRIBUTE and INFORMAION GAIN evaluation method, the ranking of the attributes is obtained. Hence the ranking of the attributes is based on their information gain value as calculated by the software for each attribute in determining the class label. The attributes with their relative ranking order is given in Figure 4.4.



```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 14 Borrower status):
        Information Gain Ranking Filter

Ranked attributes:
0.055461  11 Location
0.054952  10 Area
0.043247   2 Loan Type
0.04259    6 Sector
0.032164   3 Sex
0.029921   4 Age
0.025668   8 No.of Children
0.025135   1 Group Type
0.004201   5 Loan Size
0.003274   9 OVC
0.000997   7 No.of Emp.
0          13 No.of Fem. Emp.
0          12 No.of Male Emp.

Selected attributes: 11,10,2,6,3,4,8,1,5,9,7,13,12 : 13
```

Figure 4.4 Attributes ranked according to information gain calculated by the WEKA software.

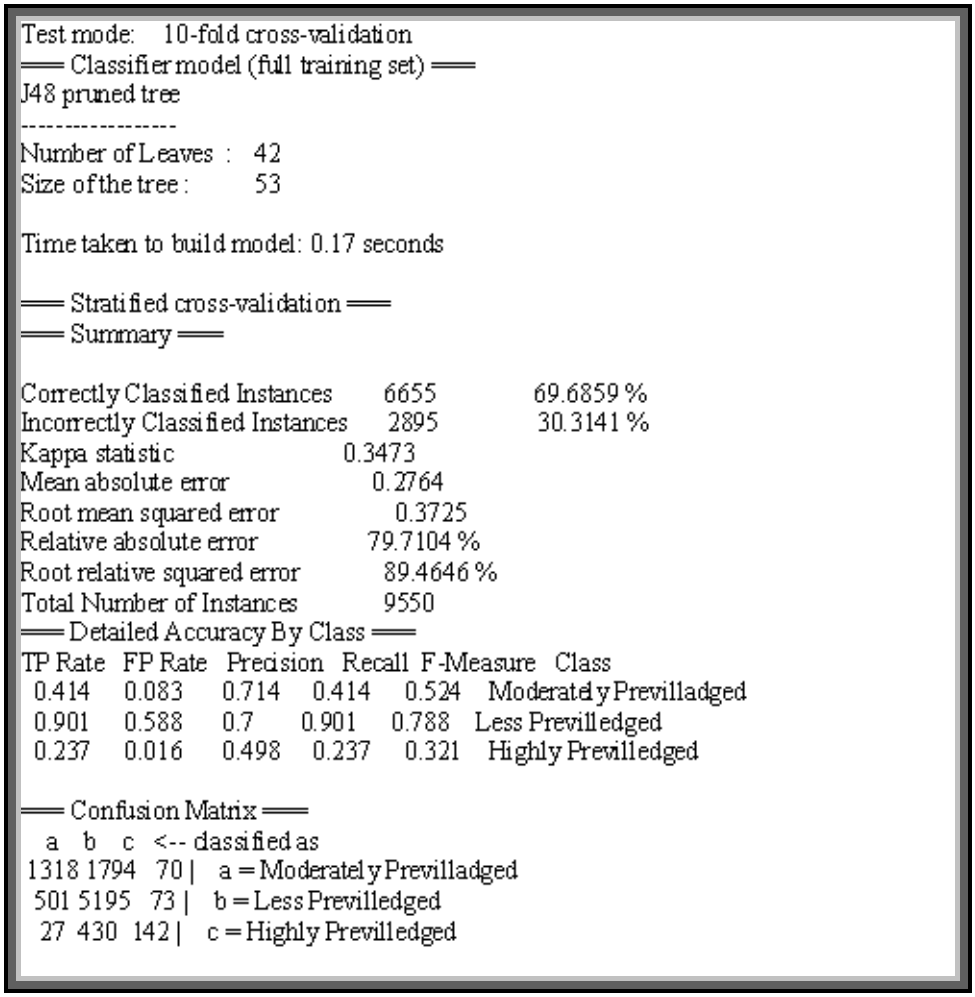
The attributes were listed down in the order of their decreasing information gain as calculated by the WEKA software. The last two attributes; No. of Male Emp. and No. of Fem Emp are removed first because both have information gain value 0. Including or excluding the third from bottom attributes, that is No.of Emp, does not have any influence on the classifier's performance, which shows that this attribute, as well, is not significantly relevant predictor attribute with respect to the target class. As a result, this experiment considers the first top ranked 10 attributes while the last three attributes namely No. of Emp., No. of Fem. Emp and No. of Male Emp are removed.

The modification on the number of attributes used brought no change on the tree built; it resulted in similar accuracy level (70.92%), number of leaves (157) and tree size (229) with Experiment One.

4.3.2.3 Experiment Three

To see if there would be any improvement, the experiment is modified in such away that the number of independent attributes is limited to six. In this case the first six attributes were selected from the ranked attributes in figure 4.4. The attributes used include Location, Area, Loan type, sector, sex and age for building classification tree based on the borrowers' status.

Figure 4.5 gives important parts of the run information as displayed by WEKA tool.



```
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
Number of Leaves  :   42
Size of the tree :    53

Time taken to build model: 0.17 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6655               69.6859 %
Incorrectly Classified Instances      2895               30.3141 %
Kappa statistic                          0.3473
Mean absolute error                      0.2764
Root mean squared error                  0.3725
Relative absolute error                 79.7104 %
Root relative squared error             89.4646 %
Total Number of Instances             9550
=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Precision   Recall  F-Measure   Class
 0.414    0.083     0.714     0.414     0.524     Moderately Previlladged
 0.901    0.588     0.7       0.901     0.788     Less Previlledged
 0.237    0.016     0.498     0.237     0.321     Highly Previlledged

=== Confusion Matrix ===
   a    b    c   <-- classified as
1318 1794   70 |   a = Moderately Previlladged
 501 5195   73 |   b = Less Previlledged
  27  430  142 |   c = Highly Previlledged
```

Figure 4.5: Result of experiment Three

The experiment conducted with less number of attributes resulted in reduced number of leaves and reduced tree size, 42 and 53 respectively. As it can be observed from Figure 4.5, even though the lesser number of nodes and tree size is a good indicator for the performance of the tree classifier, the accuracy level was decreased a bit from

the previous experiment. Hence it is necessary to go for other experiments until it would be arrived at best result that will compromise the reduced tree size and number of leaves, with improved accuracy.

4.3.2.4 Experiment Four

This experiment is again performed on the whole dataset with eight predictor attributes. The attributes are selected based on both the WEKA's attribute filter mechanism and the manual expert based attribute selection. Once the first six top ranked attributes(from figure 4.4) are taken, the other two attributes namely group type and loan size are included after discussing with the domain experts on the relevance of these attributes to be significantly influential in determining class values. The run information is given in the figure 4.6.

```
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------

Number of Leaves  :   113
Size of the tree :    152

Time taken to build model: 0.33 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances        6761               70.7958 %
Incorrectly Classified Instances      2789               29.2042 %
Kappa statistic                          0.3601
Mean absolute error                      0.2647
Root mean squared error                  0.3659
Relative absolute error                 76.3167 %
Root relative squared error             87.8684 %
Total Number of Instances             9550
=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Precision   Recall  F-Measure   Class
 0.393    0.055     0.782     0.393     0.523    Moderately Previlladged
 0.931    0.611     0.699     0.931     0.799    Less Previlledged
 0.229    0.014     0.517     0.229     0.317    Highly Previlledged
=== Confusion Matrix ===
   a    b    c   <-- classified as
1252 1866   64 |   a = Moderately Previlladged
 333 5372   64 |   b = Less Previlledged
  17  445  137 |   c = Highly Previlledged
```

Figure 4.6: Result of Experiment Four

Inclusion of some expertly selected attribute on the six top ranked ones (according to WEKA attribute filter based on information gain) resulted in increased tree and leaves while the accuracy didn't show significant improvement. Hence it is again necessary to go for another experiment.

4.3.2.5Experiment Five

This experiment is another dimension on the research which incorporates loan size with numeric values as given in the original dataset. This experiment is done on the whole dataset and selected eight attributes. The attributes and amount of dataset used were the same as with those used in experiment four but the original data in numeric value was used rather than the generalized values with an assumption that this would have effect in bringing better classifier accuracy.
The run information is given in figure 4.7.

```
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
Number of Leaves  :   176
Size of the tree :        274
Time taken to build model: 0.94 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances        7501               78.5445 %
Incorrectly Classified Instances      2049               21.4555 %
Kappa statistic                          0.5662
Mean absolute error                      0.2057
Root mean squared error                  0.3268
Relative absolute error                 59.329  %
Root relative squared error             78.4923 %
Total Number of Instances             9550
=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Precision   Recall  F-Measure   Class
 0.641    0.11      0.744     0.641    0.689    Moderately Previlladged
 0.907    0.319     0.813     0.907    0.857    Less Previlledged
 0.379    0.016     0.619     0.379    0.47     Highly Previlledged
=== Confusion Matrix ===
   a    b    c   <-- classified as
2041 1032  109 |   a = Moderately Previlladged
 505 5233   31 |   b = Less Previlledged
 198  174  227 |   c = Highly Previlledged
```

Figure 4.7 Result of experiment Five.

As the run information indicates (in figure 4.7), the summary and confusion matrix shows that this experiment resulted in better accuracy which is 78.5445%. However the number of leave and the tree size got relatively larger in this case than in the case of previous experiments. Hence a modification on this experiment by changing some classifier parameter improved the result of the tree size and number of leaves in Experiment six.

4.3.2.6 Experiment Six

Setting the J48 confidence factor to 0.15 (different from the default value 0.25) helped to essentially decrease the number of tree to 107 and the size of the tree to 164; hence this one gives the best result among all the experiments taking the compromise between the tree complexity and the accuracy level.

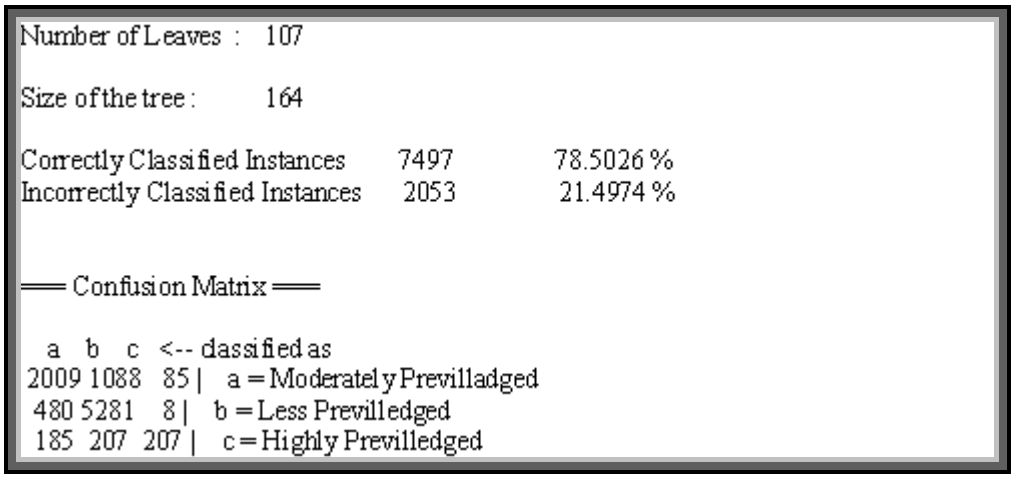


Figure 4.8: Result of experiment Six.

From this experiment with classifier parameter modified, 7497 instances were correctly classified whereas 2053 instances moved to the wrong class. Since it is difficult to depict all 107 leaves of the tree, only the top three branching of the tree and a segment/wing from the decision tree model is presented below for the purpose of illustration.

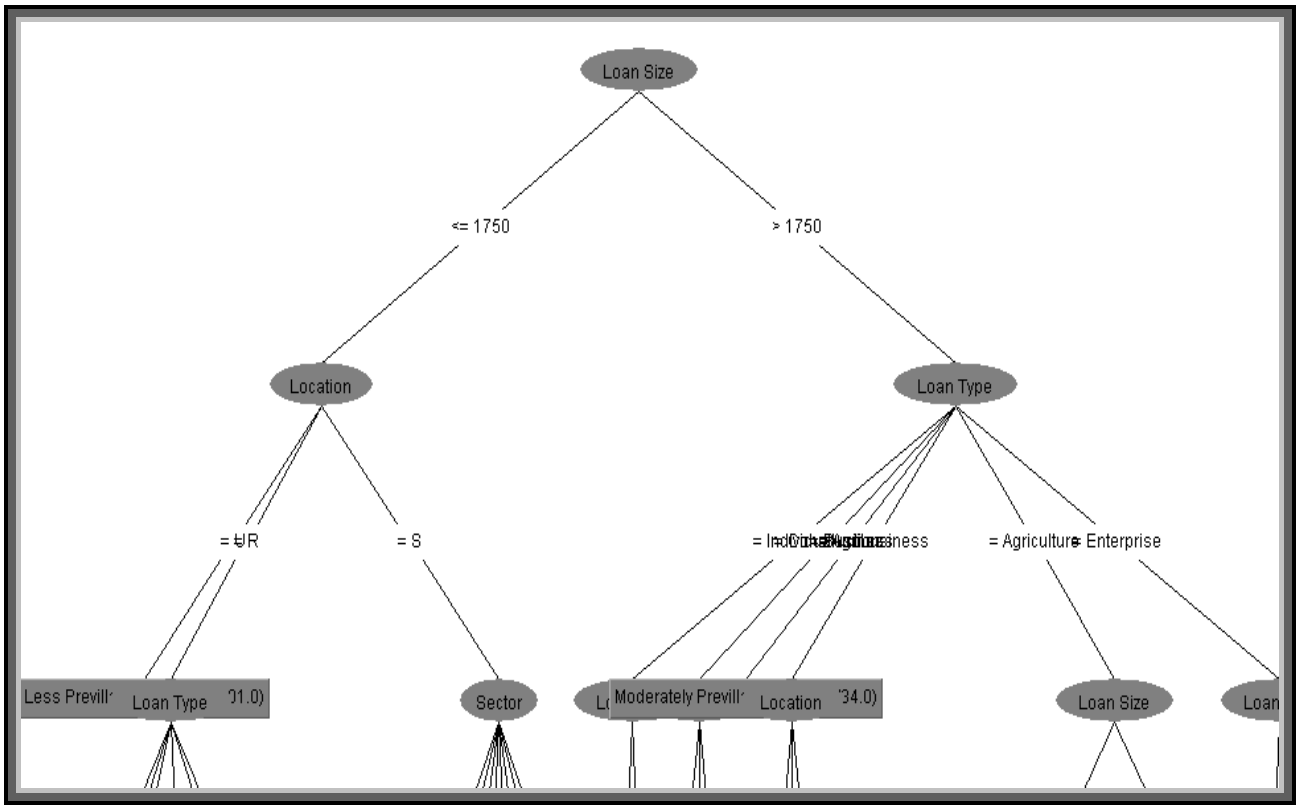


Figure 4.9: Segment from the decision tree (Graphical Representation)

Part of the j48 representation of the decision tree is shown in figure 4.10.

```
Loan Size <= 1750
|   Location = U: Less Previlledged (1559.0/301.0)
|   Location = R
|   |   Loan Type = Individual: Less Previlledged (0.0)
|   |   Loan Type = Construction: Less Previlledged (4.0)
|   |   Loan Type = Business
|   |   |   Group Type = Individual: Less Previlledged (0.0)
|   |   |   Group Type = Soliditary: Less Previlledged (28.0)
|   |   |   Group Type = Community
|   |   |   |   Loan Size <= 1100: Moderately Previlladged (3.0)
|   |   |   |   Loan Size > 1100: Less Previlledged (4.0/1.0)
|   |   Loan Type = Agribusiness: Less Previlledged (374.0/2.0)
```

Figure 4.10: Part of J48 decision tree text Representation.

Part of the decision tree shown in figure 4.10 gives examples of how the instances were classified based on the predictor attributes. For example, from the figure, while the loan size is less than or equal to 1750 br. and Location is Urban (U), there will be

1559 instances classified correctly and 301 instances classified incorrectly, as less privileged.

The numbers appearing at the leaves of the text based tree representation indicate the support in terms of observations at the leaves. For instance, the first leaf in the tree, “Loan size<=175, Location =U”, has a support of a total of 1760 records of which 1559 support the Less privileged classification and 301 do not.

4.4 Summary of the experiments

Summary of the major experiments are given below. The experiments were compared against each other based on parameters such as the number of leaves, size of the tree and the accuracy of the classifier. The following table gives the summary of the comparisons made.

No.	Experiment Label	Attributes Used	Accuracy	Leaves Generated	Size of the tree Generated
1	Experiment One	14	70.9215	157	229
2	Experiment Two	9	70.9215	157	229
3	Experiment Tree	6	69.6859	42	53
4	Experiment Four	8	70.7958	113	152
5	Experiment Five	8	78.5445	176	274
6	Experiment Six	8	78.5026	107	164

Table 4.2: Summary of the experiments

From Table 4.2, the sixth experiment, which is a modification of experiment five as stated earlier, shows best result among all the different experiments run. It is found to

be best because it shows best accuracy level with a compromised tree size and number of leaves in better way than any other experiment shown.

4.5 J48 pruned tree of the predictive model

As stated earlier (in chapter One) the main objective of building the classification model is to come up with a pattern for each borrower status (highly privileged, moderately privileged and less privileged) that would help in predicting the likely status of a new borrower in terms of these features. Different classification trees have been tried out using the J48 classifier algorithm and a classification model with relatively best accuracy and tree size is chosen in previous section.

The J48 pruned tree is attached as appendix IV out of which the best rules shown in the preceding section are selected. The algorithm result shows attributes which are evaluated to be much relevant first, on the external indention, where as the inner indention shows feature occurred as internal classification under the given attribute type in the text representation of the tree.

In the graphic representation of the tree, those attributes which are evaluated to be much relevant are indicated on top while those which are evaluated less relevant are indicated down in the tree branching and each route from the root node to a leaf node corresponds to a rule for the classification. For the purpose of illustration most of the top branching of the tree and a segment/wing from the decision tree model are presented below.

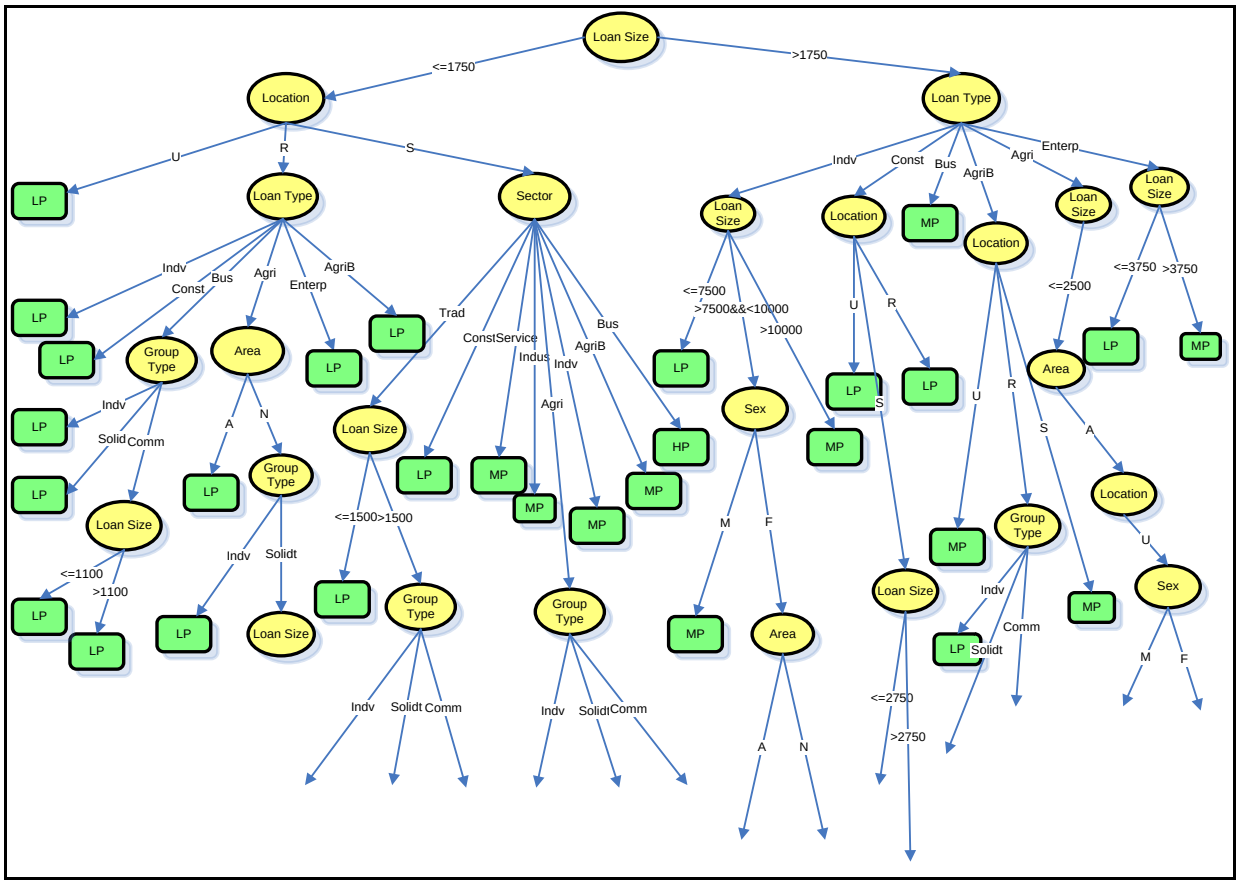


Figure 4.11 The decision tree.

The decision tree illustrated in figure 4.11 conveys information about the relevant attributes with given values in the order they were checked.

As it can be observed from the tree, there are numerous patterns generated in the way the classifier trained to classify the instances. Only some of the patterns are considered to produce the best rules class prediction of instances for the purpose of simplicity.

4.6 Best Rules Generated

Basically all the patterns observed in the classifier training can help to generate possible rules to be used for class prediction. But for the sake of simplicity and manageability, only those patterns with large number of instances correctly classified were assumed to generate the best rules and these rules are given below.

Rule #1

IF Loan Size <= 1750, Location = R, Loan Type = Agribusiness:
THEN Less Privileged = (374.0/2.0)

Rule #2

IF Loan Size <= 1750, Location = R, Area = A:
THEN Less Privileged = (2766.0/532.0)

Rule #3

IF Loan Size > 1,400 AND Loan Size <= 1,750, Location = R, Loan Type = Agriculture, Area = N, Group Type = Soliditary:
THEN Less Privileged = (150.0/32.0)

Rule #4

IF Loan Size <= 1700, Location = R, Loan Type = Agriculture, Area = N, Group Type = Community, Sex = Female:
THEN Moderately Privileged = (157.0/26.0)

Rule #5

IF Loan Size <= 1500, Sector = Trade: THEN Less Privileged= (176.0/4.0)

Rule #6

IF Loan Size <= 1750, Sector = Agriculture, Group Type = Community:
THEN Moderately Privileged= (336.0/20.0)

Rule #7

IF Loan Size > 1750 AND Loan Size<=7500, Loan Type = Individual,
THEN Less Privileged =(173.0/21.0)

Rule #8

IF Loan Size > 1750, Loan Type = Construction, Location = U:
THEN Less Privileged= (475.0/116.0)

Rule #9

IF Loan Size > 1750, Loan Type = Business:
THEN Moderately Privileged (185.0/34.0)

Rule #10

IF Loan Size > 1750 AND Loan Size <= 2500, Loan Type = Agriculture, Area = A, Location = U, Sex = Male, Group Type = Community:
THEN Highly Privileged = (58.0/16.0)

Rule #11

IF Loan Size > 1900 AND Loan Size <= 2500:, Loan Type = Agriculture, Area = A, Location = R, Group Type = Community,
THEN Less Privileged = (433.0/181.0)

Rule #12

IF Loan Size > 2500, Loan Type = Agriculture, Area = A, Location = R, Group Type = Community, Age = Adult:
THEN Moderately Privileged= (100.0/50.0)

Rule #13

IF Loan Size > 2500, Loan Type = Individual, Area = N Location = R
THEN Moderately Privileged= (440.0/86.0)

Rule #14

IF Loan Size > 2500, Loan Type = Individual, Area = N Location = R
THEN Moderately Privileged =(210.0/7.0)

Rule #15

IF Loan Size > 2500, Loan Type= Agriculture, Area = A, Sex = Male, Sector = Agriculture, Location = R:
THEN Highly Privileged= (207.0/72.0)

Rule #16

IF Loan Size > 2500, Loan Type= Agriculture, Area = A Sex = Female, Location = R: THEN Moderately Privileged= (108.0/28.0)

As depicted in the above rules generated from the tree built, the experimentation helped to get feature which characterize customers which are highly privileged, moderately privileged and less privileged. These rules will help to predict anew customer to which class label he/she may be grouped.

For example if a new customer is requesting for Loan size less or equal to 1750, and if he/she is from Rural location and Loan Type is Agribusiness, then he/she is highly regarded as a customer of type less privileged (as defined in Rule 1) meaning he/she will not probably happen to use loan service repeatedly for prolonged period of time then he/she is not among the customer groups to be given high privilege for large loan or for high frequency loan service.

There are also rules that help to extract feature that characterizes the group of customers to be regarded as highly privileged. For example if a new customer is requesting for Loan Size > 2500, from Loan Type of Agriculture, comes from Area Development, Sex is Male, engaged in Sector of Agriculture, and Location is Rural then he will be considered as Highly Privileged group (as defined in Rule 5) as a result of which he/she could be given privilege for increased loan amount and more frequencies as he/she may request.

As stated in section 2.2.3.2, except for the loan cycle 1-3 where the borrower is allowed to maximum of fixed amounts determined by the organization policy according to their group type, the maximum amount allowed will increase as the customer comes to the institution for increased loan cycle. Fifty percent increment could be granted when the customer uses the loan service from 4-6 cycles and 100% increment on his/her previous loan amount could be allowed when the customer uses for loan cycles 7 and above. In the current system, these increased loan amounts could be allowed only when it is observed that the customer has used the loan service up to the loan cycles required. Therefore rules generated could be used to predict, in advance, the borrowers' status which is equivalent to the loan cycles range as described by the rule of the organization. In accordance with the predicted borrower status the organization could grant improved loan sizes even though the borrower is using the loan service for the first time in order to improve the customer satisfaction and loyalty. For example, a borrower predicted as "less privileged" would be granted loan size to the maximum per the organization's rule for initial customers without as such any percentage increment. If a borrower is predicted as "moderately privileged", he/she may be granted maximum loan size of 50% increment on the amount determined for normal initial customers while 100% increment could be allowed for those customers predicted as "highly privileged".

This would significantly improve the customer relationship management in the organization. The attempt to predict the borrowers' status using this model would improve the customer satisfaction and loyalty strengthening the company's competitive advantage as well.

4.7 Discussion /Interpretation of the model

The confusion tree matrix and summary information displayed provides important information which is an indicator for the classifier performance. These types of information were presented in the results of each experiment as shown previously. In all the experimentations conducted, a 10 fold cross validation was set for the classifier's performance testing in the model building stage, as also stated in section 3.2.3.

A confusion matrix is probably good indicator to reflect how good or bad the model is. It not only shows how well the model predicts, but also presents the details needed to see exactly where things may have gone wrong. It shows instances that were correctly classified, and wrongly classified under another, for each class. The confusion matrix of the tree model used is presented in Table 4.3.

Actual	**Predicted**			**Total**	**Percentage of Accurate prediction**
	Moderately Privileged	Less Privileged	Highly Privileged		
Moderately Privileged	**2009**	1088	85	3182	63.13
Less Privileged	480	**5281**	8	5769	91.54
Highly Privileged	185	207	**207**	599	34.56
Total	2489	6576	300	9550	

Table 4.3 Confusion matrix of experiment Six

Table 4.3 summarizes the correct classification and incorrect classifications in each class labels for the tree model built in experiment six. According to the confusion matrix, among the 3182 instances which are actually defined to be moderately privileged, 2009 of them are correctly classified while the rest instances were misclassified; 1088 in to less privileged class and the rest 85 in to a class highly privileged. Percentage of the correct classifications for this class instances is 63.13.

Among the 5769 instances which are actually regarded as less privilege, 5281 of them were correctly classified while the rest are misclassified; 480 in to class label moderately privileged and only 8 in to the class label highly privileged. Highest percentage of the correct classification was observed for the instance of these class values where the percentage is 94.54.

Among the 599 instances which are actually considered to be highly privileged class, 207 of them were correctly classified while the rest 185 and 207 numbers of instances were incorrectly classified as moderately privileged and less privileged respectively. Percentage of the correct classifications of instances of this class is 34.56.

The overall accuracy is obtained from the percentage of the total correct classifications (2009+5281+207) to the total dataset provided to the software (9550).This gives 78.5026% which is found to be best value among all the experiments conducted. The patterns of the decision tree produced in this experiment are supported by experts on the area too. This accuracy level could be improved if more dataset than the current one could be obtained or if other classification techniques such as neural network or Bayesian network could be employed. Even though the J48 decision tree algorithm might be chosen for generating easily understandable rules and for its efficiency in time, it processes the features only one by one that would harness the accuracy levels according to Aires, et.al (2009).

The use of a predefined target classes for reasons of time and other constraints and as a fact of the objective preset, could have an impact on the accuracy level registered; Moore, et.al (N.D) states that predefined class obtained with a given dataset may

consist of records with some what class imbalance that would negatively affect the accuracy level of the model.

The researcher also believes that a combination of different techniques employed could improve the accuracy of the model developed in the current research. Davies, et.al (2007) supports this idea stating that, by selecting different classifiers at different class nodes, it tends to work better than using the same classifier at each node.

CHAPTER FIVE
CONCLUSION AND RECOMMENDATION

5.1 Conclusion

As computers and information communication technologies proliferate, there is an increasing flood of data in organizations that engaged in various sectors. Business, health, banking and finances, insurance, are some of the various sectors that continually collect, store and process large amount of data about customers, transactions, or operations. Hence it is obvious that large quantities of different data exist in organizations especially where such tools as computers and other digital technologies are properly put in place.

The availability of such large amounts of data by itself does not imply efficient and effective exploitation of the data to get the maximum benefit the organization aspires. Advanced tools and techniques are recently in common practice to enhance the utilization of existing large quantities of data, past historic data, in turning it to more meaningful information used for decision making. Data mining technology is such example which filters out certain implicit feature/pattern from vast amount of data.

The aim of the research was to investigate the potential application of data mining tools and techniques to support customer relationship management at WISDOM microfinance. The necessary data was obtained from the WISDOM microfinance head office on a scattered excel sheets which totally amounted to a dataset of 9715.

The necessary preprocessing activities were applied on the dataset after which 9550 data was prepared for the experimentation. Classification model building was experimented with J48 algorithm of the WEKA version 3.7.0 tool. Accordingly, six experiments were made with different parameters changed such as attributes and classifier parameter. Finally the last experiment conducted with selected attribute on

the full dataset and classifier confidence factor set to 0.15 was chosen since it resulted in best accuracy (78.502%) and better number of leaves and tree size. According to this experimentations the attributes loan type, loan size, location ,group type, area, sector , sex, and age are found to be relevant predictors for the target class borrower status.

The characterization of borrower status was done based on the decision tree obtained from the selected experiment. Hence most important rules were generated from the tree that characterizes the borrowers that fall under a status “highly privileged”, “moderately privileged” and “less privileged”. These characterizations or patterns will help the organization to predict the likely status of the borrower before loan disbursement. For example if the information about the borrower matches with the features extracted in Rule 11 or Rule 16, the borrower is likely “highly privileged “ category hence he/she may be given large loan amounts with frequencies as he/she request.

From the results of the experiments it can be concluded that the data mining tools and techniques especially classification techniques can be effectively applied on the microfinance social data in order to generate predictive models with an acceptable level of accuracy.

However, since the quality and size of dataset used, in addition to the mining tools and techniques are vital factors for the modeling performance, an increased size in the dataset (with increased features) could result in an improved modeling. If there were integrated ways of data handling system in the organization that could have made all data regarding the customer available in a manageable and easily accessible way, it could have enabled the research to make use of larger data with more attributes than those used in this study. Because of time and other constraints, the model building experimentation is based on just a decision tree data mining technique but it is the researcher’s belief that it would have resulted in improved accuracy if other techniques were also utilized. Hence following section presents some recommendations made based on the result of the research.

5.2 Recommendations

Even though the investigation undertaken is mainly for academic purpose, it will have paramount contribution for the organization and for other researchers interested in similar area. It has revealed the potential applicability of the data mining technology to classify borrowers in to different class labels that will help to improve the customer relationship management in the organization. In the view of the research undertaken the following recommendations are forwarded to be considered by the organization and other interested researchers as well.

- The microfinance institution should attempt to develop an integrated data warehouse that mainly consists of the customer oriented data sources. This enhances the handling of integrated, coherent and voluminous data sources encouraging further researches in the area.
- With the aid of such customer classification attempts the microfinance institution should consider what important customer relationship management strategies need to be developed for sustainable customer loyalty and for the company's competitive advantage as well.
- The model building has just focused on the dataset of social data. Therefore further data mining researches can be conducted by combining the various data available in the organization such as collection, disbursement and financial data in search of a better modeling technique for a larger dataset and larger features.
- A separate body, different from the clerk workers and accountants, need to be assigned in order to deal with loan approval and different customer complaints.
- The company should consider what important customer relationship management strategies could be applied based on the result of the research in addition to increased loan sizes and increased frequencies. Based on the result of the modeling, predicting who would churn after first few loan cycle (after first, second or third) or who would use the loan service for prolonged time, the company should identify some potential programs and offers that would entice the borrower to stay.

- Different classification algorithms such as Neural network and Bayesian networks(or combinations of any of the techniques) can be employed for customer classification by different researchers to see if it could result in a different, improved, model for the microfinance institution to help improve its customer relationship management.
- It is also the belief of the researcher that other data mining techniques such as clustering techniques would be potential further research to investigate if the clustering techniques could result in a better way of customer segmentation to be used for improving the CRM of Wisdom microfinance.

REFERENCES

1. Ahmed (2007). Classical and Incremental Classification in Data Mining Process IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.12, December 2007
2. Ali S. Koyuncugil and Nermin Ozgulbas (2008).A Data Mining Model for Detecting Financial and Operational Risk Indicators of SMEs
3. Anand Bahety (2009), Extension and Evaluation of ID3 – Decision Tree Algorithm, URL: http://www.cs.umd.edu/Grad/scholarlypapers/papers/Bahety.pdf (Accessed on 07/06/2009)
4. Anoop Singhal (2007).Data warehouse and data mining techniques for Cyber Security, Computer Security Division, USA
5. Askale Worku (2001).Possible application of data mining technology in support of Loan disbursement Activities at Dashen Bank S.C. Unpublished Thesis, Addis Ababa University.
6. B. Bakır, İ. Batmaz, F. A. Güntürkün, İ. A. İpekçi, G. Köksal, and N. E. Özdemirel (2006).Defect Cause Modeling with Decision Tree and Regression Analysis; World Academy of Science, Engineering and Technology (2006)
7. Bădulescu L. Aurelian and Nicula Adrian.(2007.) Data Mining Decision trees in economy
8. Barbara Mento and Brendan Rappel (2003).Data mining and data warehousing Association of Research Libraries, Washington D.C
9. Connally, Thomas, Carolyn, Begg and Anne (1999). Data base Systems: Practical Approach to Design implementation and Management", New Jersey, Addison-Wesley
10. Cristóbal Romerio, Sebastian Ventura, Pedro G. Espejo and Cesar Hervás (2003) Data Mining Algorithms to Classify Students Computer Science Department, Cordoba University, Spain
11. Daniel T. Larose (2005).Discovering Knowledge in data: An Introduction to Data Mining: John Wiley & Sons, Inc., Hoboken, New Jersey.
12. David Hand, Heikki Mannila and Padhraic Smyth (2001).Principles of data Mining. The MIT press Cambridge, Massachusetts Institute of Technology, London

13. Denekew Abera (2003).Application of data mining to support customer relationship Management at Ethiopian Airlines. Unpublished Thesis, Addis Ababa University
14. Elizabeth Littlefield and Richard Rosenberg (2004).Microfinance and the Poor, Finance and Development, Indonesia
15. Fayyad U. and Smyth P.(1996).From Data Mining to Knowledge Discovery In Databases
URL: http://CITESEER.Nj.Nec.Com/Fayyad96from.html , (accessed on 15/06/2009)
16. Gebrehiwot Ageba (2002).Microfinance institutions in Ethiopia: Issues of portifolio risk, Institutional arrangements and governance
URL: www.eeaecon.org/eje/v7n2/**gebrehiwot**/GEBRH-A.htm , (accessed on 23/05/2009)
17. Guangshi W. and Ning M. (2005). Research on the Application of Customer Relationship Management in Chinese Banking, *China- USA Business Review*, Volume 4, No.3 (Serial No.21), Beijing Jiaotong University
18. Han & Kamber (2001).Data mining: Concepts and techniques, San Francisco: Morgan Kaufmann Publishers
19. Herbert Edelstein (2002), Data Mining: The key to profitable Customer Relationship Management.
20. Ian H.Witten, Eibe Frank (2005).Data mining: Practical machine learning Tools and Techniques, San Francisco, Elsevier Inc.
21. Intellinova (2008).Customer Relationship management-effective techniques to dramatically improve profitability ,http: //intellinova.com.(accessed on 15/02/2009)
22. Jan Men (2006). Microfinance Services for Very Poor People: Promising Approaches from the Fiel
23. Jeffrey W.Seifert (2004).Data mining: An Overview,
URL: http://www.fas.org/irp/crs/RL31798.pdf ,(accessed on 17/06/2009)
24. Jenifer Sebstad (2003).Short study on Microfinance: Back ground documents country strategy 2003-2007,Published by SIDA-Swidish International Development cooperation Agency, Stockholm Sweden
25. Jonghoon Kim (2000).Electronic commerce and Data Mining
26. K. Boris (N.D.).Application of Data Mining in Financial Institutions, USA, Central Washington University

27. Kiva (2005).Noteson Microfinance,

URL: http://www.bsp.gov.ph/downloads/Regulations/attachments/2001,(accessed on 06/06/2009)

28. Kurt Thearling (1999). Data mining and CRM: Zeroing in on your Best customer

URL: http://www.informationmanagement.com/infodirect/19999/220/1744-1.html (accessed on 25/04/2009)

29. Madhan (2006).Data mining-competitive tool in banking,

URL: http://icai.org/resource_file/9935588-594.pdf, (accessed on 24/04/2009)

30. McKinsey & Company (2005). Banking the Unbanked: Technology's Role in Delivering Accessible Financial Services to the Poor

31. Mercy Corps (2006).The history of Microfinance,

URL: http://www.golobalenvision.org/library/4/1051 (accessed on 04/05/2009)

32. Michael J.A. Berry and Gordon S. Linoff (2004).Data mining Techniques for Marketing sales, and customer relationship management,. 2nd edition, Wiley Publishing, Inc.

33. Matthew N. Davies, Andrew Secker, Alex A. Freitas1,Jon Timmis, Miguel Mendao, Darren R. Flower (2007), An Experimental Comparison of Classification Algorithms for the Hierarchical Prediction of Protein Function. University of York, Heslington, UK

34. Nikhil R.Pal and Lakhmi Jain (2005).Advanced techniques in knowledge discovery and data mining .Springer verlarge, London

35. Ning Yang Tianrui Li Jing Song(N.D.), Construction of Decision Trees based Entropy and Rough Sets under Tolerance Relation, Research Center for Secure Application in Networks and Communications, Southwest Jiaotong University, China

36. Palous, J. (N.D.). Machine Learning And Data Mining. Prague: Gerstner Laboratory For Intelligent Decision Making And Control Czech Technical University. URL: Http://Citeseer.Nj.Nec.Com/506615.Html (accessed on 26/02/2009)

37. Paulo Guidici (2003).Applied Data Mining: Statistical Method for Business and Industry, England, John Wiley & Sons Ltd, The Atrium, Southern Gate,

Chichester, West Sussex PO19 8SQ,

38. Rachel Aires, Aline Manfrin, Sandra Aluísio, Diana Santos(2009), Which classification algorithm works best with stylistic features oPortuguese in order to classify web texts according to users' needs?,
URL: http://www.linguateca.pt/Diana/download/Airesetaltr0409relat... (accessed on 24/06/2009)
39. Rajanish Dass (2008). Data Mining in Banking and Finance: A note for bankers
URL: http://www.iimahd.ernet.in/Note-on-data-mining.
40. Rene T. Doming (N.D.).Applying data mining to Banking
URL: http://www.rtdoline.com/BMA/BSM/4.html (accessed on 06/05/2009)
41. Richard Dows (2008). Microfinance News and Information: The definition of Microfinance,URL:..http://www.microfinanceinfo.com/the-definition-of-microfinance(accessed on 04/05/2009)
42. Ruey-Shun Chen, Ruey-CHYI Wu and J.Y. Chen (2005).Data Mining Application in Customer Relationship Management Of Credit Card Business , Taiwan, Computer Society.
43. Samuel A. Moore, Daniel M. D' Addario, James Kurinskas, and Gary M. Weiss(N.D.), Are Decision Trees Always Greener on the Open(Source) Side of the Fence?, URL: http://storm.cis.fordham.edu/~gweiss/papers/dmin09-dt-evalua (accessed on 15/07/2009)
44. Shantanu Gidole and V.G. Vinod Vydiswaran (2003), DATA MINING – An Overview
45. Stefan Zemke (2003).Data mining for Prediction: Financial Series ,Case Doctorial Thesis. The Royal Institute of Technology Department of Computer and systems Sciences
46. Supatcharee Sirikulvadhana (2002) Data Mining As A Financial Auditing Tool Swedish School of Economics and Business Administration
47. Two Crows Corporation (2005), Introduction to Data Mining and Knowledge discovery ,Third Edition, Two Crows Corporation, USA
48. Valerie A., Zeithan M., Bitner J. (1976). Integrating customer focus across , 2nd Edition, USA
49. Vasilis Aggelis (2005). Predictive Model in Electronic Banking Data University of Patras Department of Computer Engineering and Informatics, Greece.
50. Vijay Prakashi and Praveen Kumar (2000).Data Mining as a tool for building and Managing customer relationship management

APPENDIX I

APPENDIX II

Interview Guide for the senior officials of the WMFI

1. How is legibility of an applicant for loan service approved at the organization?
2. Is there any mechanism for the organization to measure customer's loyalty (repeated use)?
3. Is there any special incentive mechanism in support of the customer's loyalty concept (repeated usage of a service)? If so, how are a customer's repeated usage ensured? And how much repeated use of the loan service will determine to entitle a customer as loyal and hence highly privileged or less privilege otherwise.
4. What automation efforts are there, on process, or planned for, in the organization which you think will facilitate the loan service?
5. Is there any effort (or plan) to develop an integrated, coherent and reliable customer data sources?
6. What is the view of the organization towards a premise that says "customer's loyalty (repeated use of a service) highly depends on the quality they were offered in their first approach to the business enterprise"? If you agree how you do entertain a newly approaching customer who may or may not be genuine (loyal) long term customer for the organization?

APPENDIX III

Interview Guide for the Clerk /Accountants of the WMFI

1. What types of information about a borrower do you register when the person is granted a loan for the first time?
2. What relevant information do you keep track of, about a borrower when he/she is granted the loan for a repeated time?
3. Would you please explain the meanings and roles of the different features (attributes) in the social data (records)?
4. How are the amounts of loan and frequencies for the different customers determined? Can a customer get any amount of loan in any frequency as they request or are there any requirements the borrower need to full fill in order to get the loan service as they need?
5. Is there any kind of incentive provided for the customers who used the loan service for repeated, large number of loan cycles? If so, what are the incentives mechanisms?
6. How does the organization go about it, if certain borrower (group of borrowers) after being offered the largest loan size doesn't come again for repeated loan or even for repay of the loan they were granted?
7. On the other hand, how does the organization go about it, if certain borrower (Group of borrowers) after being prohibited the large loan size they request, goes to other competitors (micro finances) and proves his /her loyalty by repeated use at the competitors market?

   (In short is there any model/pattern that helps in predicting whether a customer borrowing the loan for a single instance or may continue repeatedly borrowing)?

8. What are the roles/ purposes of the data the organization collects regarding the borrowers (especially those stored as social data)? Is it used, so far, for any purpose related to customer relationship management? If so how?,
9. Is there any automated mechanism to utilize the data for such purpose as improving customer's relationship?
10. What factors essentially makes loan service of micro finances different from that of banking and other financial institutions?(Does the difference lie on permitted loan sizes? Type of borrower granted for the loan service? Capital of the institution? or other if any?

APPENDIX IV

J48 pruned tree

Loan Size <= 1750

| Location = U: Less Privileged (1559.0/301.0)

| Location = R

| | Loan Type = Individual: Less Privileged (0.0)

| | Loan Type = Construction: Less Privileged (4.0)

| | Loan Type = Business

| | | Group Type = Individual: Less Privileged (0.0)

| | | Group Type = Soliditary: Less Privileged (28.0)

| | | Group Type = Community

| | | | Loan Size <= 1100: Moderately Privileged (3.0)

| | | | Loan Size > 1100: Less Privileged (4.0/1.0)

| | Loan Type = Agribusiness: Less Privileged (374.0/2.0)

| | Loan Type = Agriculture

| | | Area = A: Less Privileged (2766.0/532.0)

| | | Area = N

| | | | Group Type = Individual: Less Privileged (0.0)

| | | | Group Type = Soliditary

| | | | | Loan Size <= 1700

| | | | | | Sector = Trade: Less Privileged (0.0)

| | | | | | Sector = Construction

| | | | | | | Loan Size <= 1400: Moderately Privileged (2.0)

| | | | | | | Loan Size > 1400: Less Privileged (12.0/2.0)

| | | | | | Sector = Service: Less Privileged (0.0)

| | | | | | Sector = Industry: Less Privileged (51.0)

| | | | | | Sector = Agriculture

| | | | | | | Loan Size <= 1400

| | | | | | | | Loan Size <= 1250

| | | | | | | | | Sex = Male

| | | | | | | | | | Loan Size <= 1100: Moderately Privileged (11.0/4.0)

| | | | | | | | | | Loan Size > 1100: Less Privileged (7.0/1.0)

| | | | | | | | | Sex = Female

| | | | | | | | | | Loan Size <= 1050: Less Privileged (44.0/18.0)

| | | | | | | | | | Loan Size > 1050: Moderately Privileged (15.0/4.0)

| | | | | | | | Loan Size > 1250: Moderately Privileged (18.0/4.0)

| | | | | | | Loan Size > 1400

| | | | | | | | Loan Size <= 1500: Less Privileged (150.0/32.0)

| | | | | | | | Loan Size > 1500

| | | | | | | | | Sex = Male: Less Privileged (4.0)

| | | | | | | | | Sex = Female: Moderately Privileged (10.0/1.0)

| | | | | | Sector = Individual: Less Privileged (0.0)

| | | | | | Sector = Agribusiness: Less Privileged (0.0)

| | | | | | Sector = Business: Less Privileged (0.0)

| | | | | Loan Size > 1700: Less Privileged (54.0)

| | | | Group Type = Community

| | | | | Loan Size <= 1700

| | | | | | Sex = Male

| | | | | | | Sector = Trade: Less Privileged (0.0)

| | | | | | | Sector = Construction: Moderately Privileged (3.0)

| | | | | | | Sector = Service: Less Privileged (0.0)

| | | | | | | Sector = Industry: Less Privileged (0.0)

| | | | | | | Sector = Agriculture

| | | | | | | | Loan Size <= 1200

| | | | | | | | | Age = Old Age: Moderately Privileged (0.0)

| | | | | | | | | Age = Young: Moderately Privileged (5.0)

| | | | | | | | | Age = Adult: Less Privileged (4.0/1.0)

| | | | | | | | Loan Size > 1200: Less Privileged (46.0/16.0)

| | | | | | | Sector = Individual: Less Privileged (0.0)

| | | | | | | Sector = Agribusiness: Less Privileged (0.0)

| | | | | | | Sector = Business: Less Privileged (0.0)

| | | | | | Sex = Female: Moderately Privileged (157.0/26.0)

| | | | | Loan Size > 1700: Less Privileged (38.0/8.0)

| | Loan Type = Enterprise: Less Privileged (0.0)

| Location = S

| | Sector = Trade

| | | Loan Size <= 1500: Less Privileged (176.0/4.0)

| | | Loan Size > 1500

| | | | Group Type = Individual: Moderately Privileged (0.0)

| | | | Group Type = Soliditary: Less Privileged (4.0/1.0)

| | | | Group Type = Community: Moderately Privileged (15.0)

| | Sector = Construction: Less Privileged (14.0/3.0)

| | Sector = Service: Moderately Privileged (0.0)

| | Sector = Industry: Moderately Privileged (0.0)

| | Sector = Agriculture

| | | Group Type = Individual: Less Privileged (1.0)

| | | Group Type = Soliditary: Less Privileged (13.0/3.0)

| | | Group Type = Community: Moderately Privileged (336.0/20.0)

| | Sector = Individual: Moderately Privileged (0.0)

| | Sector = Agribusiness: Moderately Privileged (13.0)

| | Sector = Business: Highly Privileged (2.0)

Loan Size > 1750

| Loan Type = Individual

| | Loan Size <= 7500: Less Privileged (173.0/21.0)

| | Loan Size > 7500

| | | Loan Size <= 10000

| | | | Sex = Male: Moderately Privileged (27.0/8.0)

| | | | Sex = Female

| | | | | Area = A: Moderately Privileged (8.0/3.0)

| | | | | Area = N: Less Privileged (20.0/5.0)

| | | Loan Size > 10000: Moderately Privileged (25.0)

| Loan Type = Construction

| | Location = U: Less Privileged (475.0/116.0)

| | Location = R: Less Privileged (23.0/1.0)

| | Location = S

| | | Loan Size <= 2750: Less Privileged (16.0/3.0)

| | | Loan Size > 2750: Moderately Privileged (24.0/6.0)

| Loan Type = Business: Moderately Privileged (185.0/34.0)

| Loan Type = Agribusiness

| | Location = U: Moderately Privileged (12.0)

| | Location = R

| | | Group Type = Individual: Less Privileged (0.0)

| | | Group Type = Soliditary: Less Privileged (9.0)

| | | Group Type = Community

| | | | Area = A: Moderately Privileged (9.0)

| | | | Area = N

| | | | | Loan Size <= 2100: Less Privileged (7.0)

| | | | | Loan Size > 2100: Moderately Privileged (5.0)

| | Location = S: Moderately Privileged (34.0)

| Loan Type = Agriculture

| | Loan Size <= 2500

| | | Area = A

| | | | Location = U

| | | | | Sex = Male

| | | | | | Group Type = Individual: Highly Privileged (0.0)

| | | | | | Group Type = Soliditary

| | | | | | | Age = Old Age: Highly Privileged (1.0)

| | | | | | | Age = Young: Less Privileged (6.0)

| | | | | | | Age = Adult

| | | | | | | | Loan Size <= 1900: Highly Privileged (3.0)

| | | | | | | | Loan Size > 1900: Moderately Privileged (33.0/14.0)

| | | | | | Group Type = Community: Highly Privileged (58.0/16.0)

| | | | | Sex = Female: Moderately Privileged (52.0/17.0)

| | | | Location = R

| | | | | Group Type = Individual: Moderately Privileged (0.0)

| | | | | Group Type = Soliditary: Moderately Privileged (460.0/207.0)

| | | | | Group Type = Community

| | | | | | Loan Size <= 1900: Moderately Privileged (85.0/19.0)

| | | | | | Loan Size > 1900

| | | | | | | Loan Size <= 2000: Less Privileged (433.0/181.0)

| | | | | | | Loan Size > 2000

| | | | | | | | Loan Size <= 2450: Moderately Privileged (58.0/17.0)

| | | | | | | | Loan Size > 2450

| | | | | | | | | Age = Old Age: Moderately Privileged (26.0/10.0)

| | | | | | | | | Age = Young: Less Privileged (68.0/28.0)

| | | | | | | | | Age = Adult: Moderately Privileged (100.0/50.0)

| | | | Location = S: Moderately Privileged (8.0)

| | | Area = N

| | | | Location = U

| | | | | Loan Size <= 1900: Moderately Privileged (3.0)

| | | | | Loan Size > 1900: Less Privileged (55.0/17.0)

| | | | Location = R: Moderately Privileged (440.0/86.0)

| | | | Location = S: Moderately Privileged (210.0/7.0)

| | Loan Size > 2500

| | | Area = A

| | | | Sex = Male

| | | | | Sector = Trade: Moderately Privileged (10.0)

| | | | | Sector = Construction: Highly Privileged (0.0)

| | | | | Sector = Service: Highly Privileged (0.0)

| | | | | Sector = Industry: Highly Privileged (0.0)

| | | | | Sector = Agriculture

| | | | | | Location = U: Highly Privileged (38.0/3.0)

| | | | | | Location = R: Highly Privileged (207.0/72.0)

| | | | | | Location = S: Moderately Privileged (5.0)

| | | | | Sector = Individual: Highly Privileged (0.0)

| | | | | Sector = Agribusiness: Highly Privileged (0.0)

| | | | | Sector = Business: Highly Privileged (0.0)

| | | | Sex = Female

| | | | | Location = U: Highly Privileged (3.0)

| | | | | Location = R: Moderately Privileged (108.0/28.0)

| | | | | Location = S: Moderately Privileged (18.0)

| | | Area = N: Moderately Privileged (36.0)

| Loan Type = Enterprise

| | Loan Size <= 3750: Less Privileged (26.0/6.0)

| | Loan Size > 3750: Moderately Privileged (5.0)