ADDIS ABABA UNIVERSITY

SCHOOL OF INFORMATION SCIENCE

THE APPLICATION OF DATA MINING IN CREDIT RISK ASSESSMENT: THE CASE OF UNITED BANK SC

MENGISTU TESFAYE

June, 2013

Addis Ababa

Ethiopia

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

THE APPLICATION OF DATA MINING IN CREDIT RISK ASSESSMENT: THE CASE OF UNITED BANK SC

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

MENGISTU TESFAYE

June, 2013

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

THE APPLICATION OF DATA MINING IN CREDIT RISK ASSESSMENT: THE CASE OF UNITED BANK SC

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

MENGISTU TESFAYE

June, 2013

Name and signature of Member of the Examining Board

Name	Title	Signature	Date
_____	Chairperson	_____	
_____	Advisor	_____	_____
_____	Examiner	_____	

Declaration

I declare that this thesis is my original work and has not been presented for a degree in any other university.

_____

Date

This thesis has been submitted for examination with my approval as university advisor.

_____

Dereje Teferi (PhD)

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God for his peace and blessings upon my life and indeed for granting me the chance and the ability to successfully complete this study.

I wish to express my deepest gratitude to my advisor, Dr. Dereje Teferi for his constructive comments and overall guidance.

I would also like to thank the credit department staffs and loan officers of United Bank for their unreserved and detailed explanation of credit activities.

I am very much grateful to my wife, W/ro Martha Jimma and my sweet daughter, Blen Mengistu for their care and understanding during my study times.

Last but not least, I would like to thank my instructors and friends for the constant assistance and encouragement they rendered to me since the time of my admission to the postgraduate program.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Credit facilities and investments are the cornerstones of the growing economy of Ethiopia. United Bank being one of the former private banks has played its own role in the economy by rendering loan facilities to the individuals and companies which are running business in various sectors. The bank uses internal and NBE credit policies, procedures and strictly followed manuals in various levels of credit committees before disbursing loan to customers. However, there are total defaulters and inconsistent loan repaying customers which declines the profitability of the bank in particular and threatens the growing economy of the country in general. While fueling the sprinting economy in the country, minimizing the possible defaulters is the prime concern of the bank.

Identifying customers and contracts which are more likely to be inconsistent loan payers or defaulters is an important issue. This data mining research has been carried out to identify trends of good and bad or NPL (non-performing loan) patterns from the historic data and build predictive Model to assist the management of the bank.

This research has used the last 7 years credit data of United Bank and applied various preprocessing activities to clean the data. An experiment has been conducted using the CRISP-DM (2000) Model using WEKA tool. Different parameters of WEKAJ48 Decision tree and Naïve Bayes classification algorithm were applied. The model developed using the J48 decision tree algorithm has showed highest classification accuracy of 96.6%.

Generally, the result of this study has showed that the application of data mining in Credit data can bring valuable input to assist the decision of credit committees and management.

ACRONYMS

AAU: Addis Ababa University

AI: Artificial Intelligence

ARFF: Attribute Relation File Format

CRA: Credit Risk Assessment

CRC: Collaborative Research Center

CRISP-DM: CRoss-Industry Standard Process for Data Mining

CSV: Comma Separated Value

DM: Data Mining

DT: Decision Tree

DTS: Domestic trade and services

KD: Knowledge Discovery

KDD: Knowledge Discovery in Databases

LC: Letter of Credit

NBE: National Bank of Ethiopia

NPL: Non-performing Loan

OD: Overdraft

OLAP: On-Line Analytical Processing

ROC: Receiver Operating Characteristics

SEMMA: Sample, Explore, Modify, Model, and Assess

SMOTE: Synthetic Minority Oversampling Technique

SQL: Structured Query Language

UB: United Bank

WEKA: Waikato Environment for Knowledge Analysis

CHAPTER 1

INTRODUCTION

1.1. Background

Ethiopia is one of the poorly developed countries in terms of infrastructure and services. The banking sector is one of the cornerstones which play its own role in the development of the country. The number of private banks in the country is increasing from year to year. The presence of many banks in the country has created fierce competition in the traditional brick and mortar based market. All private and governmental banks have partially or fully automated their banking operations as per the national bank directive to implement standard core banking solution.

Many of these private and governmental banks have a huge amount of data which is used for statement, auditors' verification of transactions and functional level reporting purposes. In order to discover the set of critical success factors that will help banks reach their strategic goals and remain in the competition, they need to move beyond standard business reporting and sales forecasting. They should learn from their abundant historical data, by applying data mining and predictive analytics to extract actionable intelligent insights and quantifiable predictions. These insights can support the decision of management, auditors and other clerical staffs in the pillar activities of the bank like credit risk assessment (CRA) to grant loan to a customer.

The credit risk assessment of customers involves structured and non-structured management decision elements. The structured decisions are those where the processes necessary for the granting of loan are known beforehand and several computational tools to support the decisions are available. For non-structured decisions, only the managers' intuition and experience are used. Specialists may support these managers, but the final

decisions involve a substantial amount of subjective elements. Data mining comes here into picture to assist the unstructured decision of management in predicting and execution of the necessary follow up procedures.

United bank S.C being one of the technologically rich banks in the country, has introduced many channels like SMS, internet, telephone and now ATM and POS banking services. It has been using legacy system application since its inception and is using core banking solutions for the last 7 years. These years of day-to-day transactions created terabytes of data which are collected, generated, printed, and stored only for the sake of customer statements, reports for lower level functional managers and auditors.

United Bank's Core banking system includes the loan module, where credit contract detailed information is booked, and a huge number of credit related transactions like repayment, daily interest accruals, status changes, full liquidations and contract amendment information is captured over several years.

The analysis of these data may lead to a better understanding of the customer's profile and attached contract, thus supporting the offer of new products or services and identification of risky disbursements. These data usually hold valuable information like trends and patterns, which can be employed to improve credit assessment. The bulky nature of the data makes its manual analysis an impossible task. In many cases, several related features need to be simultaneously considered in order to accurately model credit contract behavior.

Credit facility is the corner stone of United Bank's existence and profitability. As a result, the automatic extraction of useful credit knowledge from its historic data will be an input to the credit risk grading process. Hence, in order to assist the management of the bank to make effective

decisions in reducing risk of defaulters and focus on risk free sectors and customers, implementing data mining will play a great role.

1.2. Statement Of The Problem

Since the enactment of the proclamation for the licensing and supervision of banks, the numbers of competent and powerful banks with the state of the art technology are increasing every year. Despite the number of banks and potential customers, the banking services are so limited and nearly homogenous which may be attributed to several factors. The presence of many private banks and homogenous service strategy has created a good opportunity for credit customers to identify the structured requirement of banks easily. These customers will apply credit for different banks, some with misleading formalities at different times. This environment has created a brutal competition between banks and kept them struggling in cash collection and in rendering risk free credit facility.

Management and measurement of risk is at the core of every financial institution. Today's major challenge in the banking and insurance world is therefore the implementation of risk management systems in order to identify, measure, and control business exposure. Here credit and market risk present the central challenge. One can observe a major change in the area of how to measure and deal with them, based on the advent of advanced database and data mining technology (Rajanish,(n.d)).

Loan (credit facility) is one of the major services that contribute to the lion share of profitability of any bank in Ethiopia. In order to lend money to customers, banks need to collect more cash from customers and other various means. Currently each and every bank is extending its effort in deposit mobilization by various methods like branch expansion, increased interest rate for fixed time depositors etc. This hard fetched deposit is later given as a credit to various customers. But there will be inconsistent loan

repayments, defaulters and corruptions related to various sector credit facility customers. Therefore, Banks are in need of loan repayment prediction and customer credit analysis from their historic credit facility data. In order to alleviate these stated problems and find new, applicable and interesting classifications and predictions, the banks need to look deep into their historic data.

There is a huge amount of data in Ethiopian banks which is used mainly for reporting purpose. The application of data mining to this data is assumed to yield extra knowledge to the banks.

Therefore, this research will try to address the following basic and general questions:-

- Can useful patterns be extracted through data mining for the credit risk area to assist the decision makers of the Ethiopian banks in general and to united bank in particular?
- Which data mining techniques can best be used for the credit facility (loan) risk assessment area?
- What are the interesting patterns and relationships for the risky and risk free credit contracts of United Bank?
- What is the response of domain experts on the patterns identified?

1.3. Scope and Limitations

The main target of this research is to identify the applicability of data mining in the credit risk assessment of united bank. This is an area of study that will be more fruitful if it were conducted widely by including all private and governmental banks of Ethiopia.

Taking in to consideration the time and budget constraint, the coverage of this research would be on united bank S.C. only. United bank has been using “flexcube (Oracle financials) core banking solution” with underlying

oracle database for the last 7 years. It has also been using a legacy system before the introduction of flexcube. Due to resource and time limitations to merge the two system information into a data warehouse, this research will be conducted on the last 7 years flexcube core banking data.

1.4. Objective(s)

1.4.1. General Objectives

This research assesses the applicability of data mining to the credit risk assessment of United Bank in order to get a competitive edge and better customer satisfaction.

1.4.2. Specific Objectives

The Specific objectives of this research paper concentrates on the following key points:

- To identify the possible patterns of good and bad loans based on the transaction history of existing and previous customers;
- To illustrate the accuracy of preferred data mining techniques and model for the pre-processed credit data;
- To select the best prediction model and find out interesting patterns from the output of the selected Model;
- To assess applicability of the pattern(knowledge) generated by the model by discussing with the domain experts;

Based on the above analysis and assessment, the researcher will report the knowledge assimilated for all concerned.

1.5. Significance of the Research

The study contributes a lot for united bank as a starting point if the bank needs to implement data mining on its database to get extra competitive edge. Hence this study will be conducted to provide sufficient information

on credit risk assessment using data mining to uncover hidden knowledge based on the bank's historic data.

As the services and style of product development is nearly similar for Ethiopian banks, the research can also be used by other banks in the country to facilitate decision making and earn better profit margins by reducing credit risks.

It also serves as a starting point for those individuals who would like to undertake broader research on the topic.

1.6. Methodology

In order to follow the industry standard knowledge discovery process for this research, the researcher has followed CRISP-DM (Cross Industry Standard Process for Data Mining) phases. According to kdnuggets (2012), CRISP-DM is a data mining process model that describes commonly used approaches that expert data miners use to tackle problems. Polls conducted in 2002, 2004, and 2007 show that it is the leading methodology used by data miners. The only other data mining standard named in these polls was SEMMA. However, 3-4 times as many people reported using CRISP-DM. Data collected from united bank core banking solution will be analyzed through qualitative and quantitative research techniques. CRISP-DM is used for this research. CRISP-DM breaks the process of data mining into six major phases:

- Business Understanding
    - Determine business objectives
    - Assess situation
    - Determine data mining goals
    - Produce project plan
- Data Understanding

        - Collect initial data
        - Describe data
        - Explore data
        - Verify data quality
- Data Preparation
    - Select data
    - Clean data
    - Construct data
    - Integrate data
    - Format data
- Modeling
    - Select modeling technique
    - Generate test design
    - Build model
    - Assess model
- Evaluation
    - Evaluate results
    - Review process
    - Determine next steps
- Deployment (This step will not be done in this research)

1.7. Organization of the Thesis

This thesis is organized into five chapters.

- Chapter 1: Introduction
- Chapter 2: Literature Review
- Chapter 3: Survey of Credit Risk Assessment at United bank SC
- Chapter 4: Experimentation
- Chapter 5: Conclusion and Recommendation

CHAPTER 2

DATA MINING AND KNOWLEDGE DISCOVERY IN DATABASE

2.1. Overview of Data Mining and KDD

Living in the age of digital information, the problem of data overload is an eminent phenomenon. Our ability to analyze and understand massive datasets lags far behind our ability to gather and store the data. A new generation of computational techniques and tools is required to support the extraction of useful knowledge from the rapidly growing volumes of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD) and data mining (Fayyad et al, 1996).

Progress in digital data acquisition and storage technology has resulted in the growth of huge databases. This has occurred in all areas of human endeavor, from the mundane(such as supermarket transaction data, credit card usage records, telephone call details, and government statistics) to the more exotic (such as images of astronomical bodies, molecular databases, and medical records). Little wonder, then, that interest has grown in the possibility of tapping these data, of extracting from them information that might be of value to the owner of the database (Hand et al, 2001).

Conventionally the idea of searching pertinent patterns in data has been referred using different names such as data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. Among these terms KDD and data mining are used widely (Fayyad et al, 1996).

According to the Gartner, Inc. (2012), Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting

through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.

The above definition is strengthened by Hand et al(2001) as they explain Data mining as the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. The relationships and summaries derived through a data mining exercise are often referred to as models or patterns. Examples include linear equations, rules, clusters, graphs, tree structures, and recurrent patterns in time series.

The definition above refers to "observational data," as opposed to "experimental data". 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 means that the objectives of the data mining exercise play no role in the data collection strategy. 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. For this reason, data mining is often referred to as "secondary" data analysis.

The definition also mentions that the data sets examined in data mining are often large. If only small data sets were involved, the discussion would be merely on classical exploratory data analysis as practiced by statisticians. When large bodies of data are faced, new problems arise. Some of these relate to housekeeping issues of how to store or access the data, but others relate to more fundamental issues, such as how to determine the representativeness of the data, how to analyze the data in a reasonable period of time, and how to decide whether an apparent relationship is merely a chance occurrence not reflecting any underlying reality.

As per the explanation of Zaki and Wong (2003), Data mining is generally an iterative and interactive discovery process. The goal of this process is to mine patterns, associations, changes, anomalies, and statistically significant structures from large amount of data. Further-more, the mined results should be valid, novel, useful, and understandable. These qualities that are placed on the process and outcome of data mining are important for the following reasons:

(1) ***Valid***: It is crucial that the patterns, rules, and models that are discovered are valid not only in the data samples already examined, but are generalizable and remain valid in new data samples as well. Only then can the rules and models obtained be considered meaningful.
(2) ***Novel***: It is desirable that the patterns, rules, and models that are discovered are not already known to experts. Otherwise, they would yield very little new understanding of the data samples and the problem at hand.
(3) ***Useful***: It is desirable that the patterns, rules, and models that are discovered allow us to take some useful action. For example, they allow us to make reliable predictions on future events.
(4) ***Understandable***: It is desirable that the patterns, rules, and models that are discovered lead to new insight on the data samples and the problem being analyzed.

As per Zaki and Wong (2003), the goals of data mining are often that of achieving reliable prediction and/or that of achieving understandable description. The former answers the question “what”, while the latter the question “why”. With respect to the goal of reliable prediction, the key criterion is that of accuracy of the model in making predictions on the problem being analyzed. How the prediction decision is arrived at may not be important. With respect to the goal of understandable description, they key criteria is that of clarity and simplicity of the model describing the problem being analyzed.

There is sometimes a dichotomy between these two aspects of data mining in the sense that the most accurate prediction model for a problem may not be easily understandable, and the most easily understandable model may not be highly accurate in its predictions. According to Cristianini and Scholkopf (2002), on many analysis and prediction problems, support vector machines are reported to hold world records in accuracy. However, the maximum error margin models constructed by these machines and the quadratic programming solution process of these machines are not readily understood to the non-specialists. In contrast, the decision trees constructed by tree induction classifiers such as C4.5 are readily grasped by non-specialists, even though these decision trees do not always give the most accurate predictions.

Hand et al (2001), states that Data mining is often set in the broader context of knowledge discovery in databases, or KDD. This term (KDD) is originated in the artificial intelligence (AI) research field. The KDD process involves several stages: selecting the target data, preprocessing the data, transforming them if necessary, performing data mining to extract patterns and relationships, and then interpreting and assessing the discovered structures.

The process of seeking relationships within a data set— of seeking accurate, convenient, and useful summary representations of some aspect of the data—involves a number of steps:

- determining the nature and structure of the representation to be used;
- deciding how to quantify and compare how well different representations fit the data (that is, choosing a "score" function);
- choosing an algorithmic process to optimize the score function; and
- Deciding what principles of data management are required to implement the algorithms efficiently.

2.2. Foundations Of Data Mining

According to Thearling (2012), Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

(a) Massive data collection

(b) Powerful multiprocessor computers

(c) Data mining algorithms

In the evolution from business data to business information, each new step has built upon the previous one. For example, dynamic data access is critical for drill-through in data navigation applications, and the ability to store large databases is critical to data mining. From the user's point of view, the four steps listed in Table 2.1 were revolutionary because they allowed new business questions to be answered accurately and quickly.

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record

		(SQL), ODBC		level
Data Warehousing &Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

Table 2.1: Steps involved in Evolution of Data Mining (Thearling, 2012)

The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments.

2.3. Supervised and Unsupervised Learning

As per CRC 649 (2013), Data and Knowledge Mining is basically learning from data. In this context, data are allowed to speak for themselves and no prior assumptions are made. This learning from data comes in two flavors: supervised learning and unsupervised learning. In supervised learning (often also called directed data mining) the variables under investigation can be split into two groups: explanatory variables and one (or more) dependent variables. The target of the analysis is to specify a relationship between the explanatory variables and the dependent variable as it is done in regression analysis. To apply directed data mining

techniques the values of the dependent variable must be known for a sufficiently large part of the data set.

Unsupervised learning is closer to the exploratory spirit of Data Mining. In unsupervised methods, no target variable is identified as such. Instead, the data mining algorithm searches for patterns and structure among all the variables. There is no distinction between explanatory and dependent variables. However, in contrast to the name undirected data mining there is still some target to achieve. This target might be as general as data reduction or more specific like clustering (Ding, 2007).

The large amount of data that is usually present in Data Mining tasks allows splitting the data file in three groups: training cases, validation cases and test cases. Training cases are used to build a model and estimate the necessary parameters. The validation data helps to see whether the model obtained with one chosen sample may be generalized to other data. In particular, it helps avoiding the phenomenon of over fitting. Iterative methods incline to result in models that try to do too well. The data at hand is perfectly described, but generalization to other data yields unsatisfactory outcomes. Not only different estimates might yield different models, usually different statistical methods or techniques are available for a certain statistical task and the choice of a method is open to the user. Test data can be used to assess the various methods and to pick the one that does the best job on the long run.

Although we are dealing with large data sets and typically have abundant cases, partially missing values and other data peculiarities can make data a scarce resource and it might not be easily achievable to split the data into as many subsets as there are necessary. Re-sampling and cross-validation techniques are often used in combination with data and computer intensive methods in Data Mining.

2.4. Data Mining Tasks

The cycle of data and knowledge mining comprises various analysis steps, each step focusing on a different aspect. CRC 649 (Collaborative Research Center) (2003) proposes the following categorization of data mining tasks.

2.4.1. Description and Summarization

At the beginning of each data analysis, there is the wish and need to get an overview of the data, to see general trends as well as extreme values rather quickly. It is important to familiarize with the data, to get an idea what the data might be able to tell you, where limitations will be, and which further analyses steps might be suitable. Typically, getting the overview will at the same time point the analyst towards particular features, data quality problems, and additional required background information. Summary tables, simple univariate descriptive statistics, and simple graphics are extremely valuable tools to achieve this task.

Checking data quality is by no means a negative part of the process. It leads to deeper understanding of the data and to more discussions with the data set owners. Discussions lead to more information about the data and the goals of the study.

Speed of the data processing is an important issue at this step. For simple tasks and data summary and description are typically considered to be simple tasks, although it is generally not true users are not willing to spend much time. A frequency table or a scatter plot must be visible in the fraction of a second, even when it comprises a million observations. Only some computer programs are able to achieve this. Another point is a fast scan through all the variables: if a program requires an explicit and lengthy specification of the graph or table to be created, a user typically will end this tedious endeavor after a few instances. Generic functions with context-sensitive and variable-type-dependent responses provide a viable solution to

this task. On the level of standard statistical data sets this is provided by software like XploRe, S-Plus and R with their generic functions summary and plot. Generic functions of this kind can be enhanced by a flexible and interactive user environment which allows navigating through the mass of data, to extract the variables that show interesting information on the first glance and that call for further investigation. Currently, no system comes close to meet these demands, future systems hopefully will do.

2.4.2. Descriptive Modeling

General descriptions and summaries are an important starting point but more exploration of the data is usually desired. While the tasks in the previous section have been guided by the goal of summary and data reduction, descriptive modeling tries to find models for the data. In contrast to the subsequent section, the aim of these models is to describe, not to predict models. As a consequence, descriptive models are used in the setting of unsupervised learning. Typical methods of descriptive modeling are density estimation, smoothing, data segmentation, and clustering. Clustering is a well-studied and well-known technique in statistics. Many different approaches and algorithms, distance measures and clustering schemes have been proposed. With large data sets all hierarchical methods have extreme difficulties with performance. The most widely used method of choice is k-means clustering. Although k-means is not particularly tailored for a large number of observations, it is currently the only clustering scheme that has gained positive reputation in both the computer science and the statistics community. The reasoning behind cluster analysis is the assumption that the data set contains natural clusters which, when discovered, can be characterized and labeled. While for some cases it might be difficult to decide to which group they belong, we assume that the resulting groups are clear-cut and carry an intrinsic meaning. In segmentation analysis, in contrast, the user typically sets the number of groups in advance and tries to partition all cases in homogeneous subgroups.

2.4.3. Predictive Modeling

Predictive modeling falls into the category of supervised learning; hence, one variable is clearly labeled as target variable Y and will be explained as a function of the other variables X. The nature of the target variable determines the type of model: classification model, if Y is a discrete variable, or regression model, if it is a continuous one. Many models are typically built to predict the behavior of new cases and to extend the knowledge to objects that are new or not yet as widely understood; Predicting the value of the stock market, the outcome of the next governmental election, or the health status of a person. Banks use classification schemes to group their costumers into different categories of risk.

2.4.4. Discovering Patterns and Rules

The realm of the previous tasks has been much within the statistical tradition in describing functional relationships between explanatory variables and target variables. There are situations where such a functional relationship is either not appropriate or too hard to achieve in a meaningful way. Nevertheless, there might be a pattern in the sense that certain items, values or measurements occur frequently together. Association Rules are a method originating from market basket analysis to elicit patterns of common behavior.

2.4.5. Retrieving Similar Objects

The World Wide Web contains an enormous amount of information in electronic journal articles, electronic catalogs, and private and commercial homepages. Having found an interesting article or picture, it is a common desire to find similar objects quickly. Based on key words and indexed meta-information search engines are providing us with this desired information. They do not only work on text documents, but to a certain extent also on images. Semi-automated picture retrieval combines the

ability of the human vision system with the search capacities of the computer to find similar images in a data base.

2.5. Data Mining Techniques

Moin and Ahmed (2012) have reported the presence of several data mining techniques and algorithms that have been developed and used in data mining like association, classification, clustering, prediction and sequential patterns, Regression, Neural Networks etc. We will briefly examine some of these techniques as follows:

2.5.1. Classification

Moin& Ahmed (2012) have considered classification as the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to new data sets. For a fraud detection application, this would include complete records of both fake and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper judgment. The algorithm then encodes these parameters into a model called a classifier (Moin& Ahmed, 2012).

Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

2.5.2. Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. Association and correlation is usually used to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis (Moin& Ahmed, 2012). For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit. The various types of associations include (Moin& Ahmed, 2012):

- Multilevel association rule.
- Multidimensional association rule
- Quantitative association rule
- Direct association rule.
- Indirect association rule.

2.5.3. Clustering

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes. As per Kargupta, Joshi, Kumar, and Yesha (2005),Classification approach can also be used for effective means of distinguishing groups or classes of object

but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For Example: The customer of a given geographic location and of a particular job profile demand a particular set of services, like in banking sector the customers from the service class always demand for the policy which ensures more security as they are not intending to take risks. Similarly the same set of service class people in rural areas have a preference for some particular brands which may vary from their counterparts in urban areas. This information will help the organization in cross-selling their products, The bank's customer service representatives can be equipped with customer profiles enriched by data mining that help them to identify which products and services are most relevant to callers. This technique will help the management in finding the solution of 80/20 principle of marketing, which says: Twenty per cent of your customers will provide you with 80 per cent of your profits, then problem is to identify those 20 % and the techniques of clustering will help in achieving the same (Bhambri, 2011). Types of clustering methods

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

2.5.4. Prediction

According to Moin& Ahmed (2012), the prediction as it name implies is a data mining techniques that discovers relationship between dependent and independent variables. Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may

depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. Types of regression methods include:

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

2.5.5. Sequential Patterns

Sequential patterns analysis in one of data mining technique that seeks to discover similar patterns in data transaction over a business period. The uncovered patterns are used for further business analysis to recognize relationships among data.

2.6. Steps in Data Mining Process

The data mining process is often characterized as a multi-stage iterative process involving data selection, data cleaning, and application of data mining algorithms, evaluation, and so forth.

Some of the most serious errors in data analysis result from a poor understanding of the problem—an understanding that must be developed before we get into the details of algorithms to be used. Here is a list of steps to be taken in a typical data mining effort (Shmueli et al, 2010):

1. Develop an understanding of the purpose of the data mining project(if it is a one-shot effort to answer a question or questions) or application (if it is an ongoing procedure).
2. Obtain the dataset to be used in the analysis. This often involves random sampling from a large database to capture records to be used in an analysis. It may also involve pulling together data from different databases. The databases could be internal (e.g., past purchases

made by customers) or external (credit ratings). While data mining deals with very large databases, usually the analysis to be done requires only thousands or tens of thousands of records.

3. Explore, clean, and preprocess the data. This involves verifying that the data are in reasonable condition. How missing data should be handled? Are the values in a reasonable range, given what you would expect for each variable? Are there obvious outliers? The data are reviewed graphically: for example, a matrix of scatterplots showing the relationship of each variable with every other variable. We also need to ensure consistency in the definitions of fields, units of measurement, time periods, and so on.
4. Reduce the data, if necessary and (where supervised training is involved) separate them into training, validation, and test datasets. This can involve operations such as eliminating unneeded variables, transforming variables (e.g., turning "money spent" into "spent > $100" vs. "spent ≤ $100"), and creating new variables (e.g., a variable that records whether at least one of several products was purchased). Make sure that you know what each variable means and whether it is sensible to include it in the model.
5. Determine the data mining task (classification, prediction, clustering, etc.). This involves translating the general question or problem of step 1 into a more specific statistical question.
6. Choose the data mining techniques to be used (regression, neural nets, hierarchical clustering, etc.).
7. Use algorithms to perform the task. This is typically an iterative process—trying multiple variants, and often using multiple variants of the same algorithm (choosing different variables or settings within the algorithm). Where appropriate, feedback from the algorithm's performance on validation data is used to refine the settings.
8. Interpret the results of the algorithms. This involves making a choice as to the best algorithm to deploy, and where possible, testing the final choice on the test data to get an idea as to how well it will

perform. (Recall that each algorithm may also be tested on the validation data for tuning purposes; in this way the validation data become a part of the fitting process and are likely to underestimate the error in the deployment of the model that is finally chosen.)

9. Deploy the model. This involves integrating the model into operational systems and running it on real records to produce decisions or actions. For example, the model might be applied to a purchased list of possible customers, and the action might be "include in the mailing if the predicted amount of purchase is > $10."

2.7. The Data Mining Models

There are different DM process model standards that are used in different research and business data mining projects.

- ❖ KDD process (Knowledge Discovery in Databases),
- ❖ SEMMA (Sample Explore Modify Model Assess)
- ❖ CRISP-DM (CRoss Industry Standard Process for Data Mining), and

2.7.1. The KDD process

The KDD process, as presented in (Fayyad et al, 1996) is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database. The general KDD process is depicted in Figure 2.1. It comprises the following steps (Han and Kamber, 2000).

1. **Selection –** This stage consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.
2. **Preprocessing** – This stage consists on the target data cleaning and preprocessing in order to obtain consistent data.
3. **Transformation** – This stage consists on the transformation of the data using dimensionality reduction or transformation methods.

4. **Data Mining** – This stage consists on the searching for patterns of interest in a particular representational form, depending on the data mining objective (usually, prediction)
5. **Interpretation/Evaluation** – This stage consists on the interpretation and evaluation of the mined patterns.

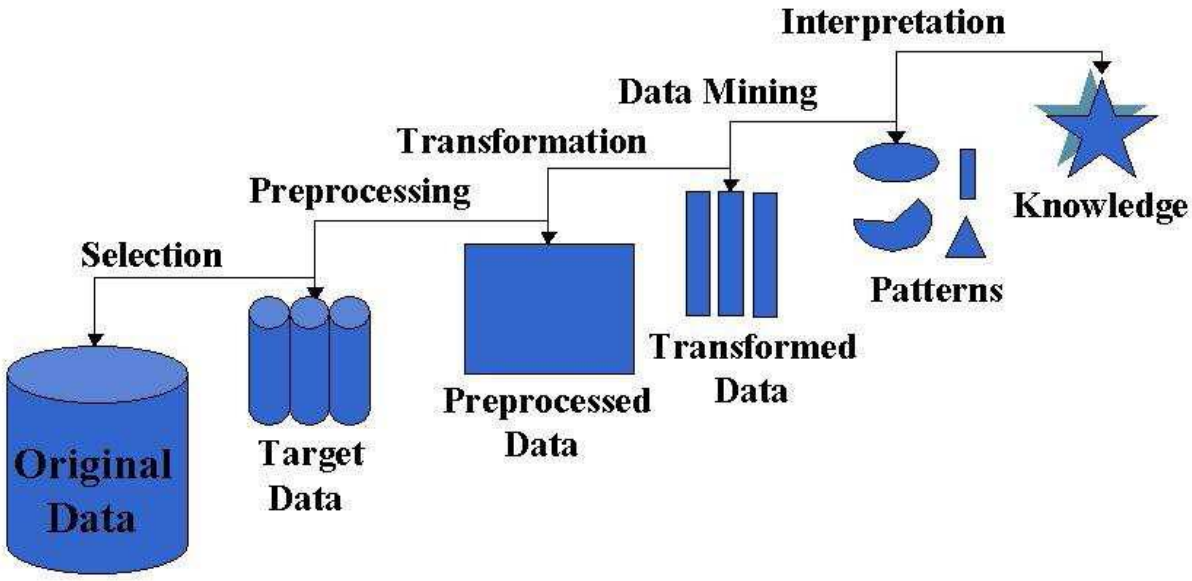

Figure 2.1: The five stages of KDD (Zaki& Wong, 2003).

The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user (Brachman and Anand, 1996).
Additionally, the KDD process must be preceded by the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user. It also must be continued by the knowledge consolidation by incorporating this knowledge into the system (Fayyad et al, 1996).

2.7.2. The SEMMA process

The SEMMA process was developed by the SAS Institute. The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to

the process of conducting a data mining project. The SAS Institute considers a cycle with 5 stages for the process (SAS Institute Inc., 1998):

1. **Sample** – This stage consists on sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly. This stage is pointed out as being optional.
2. **Explore** – This stage consists on the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.
3. **Modify** – This stage consists on the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.
4. **Model** – This stage consists on modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.
5. **Assess** – This stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs.

Although the SEMMA process is independent from de DM chosen tool, it is linked to the SAS Enterprise Miner software and pretends to guide the user on the implementations of DM applications.

SEMMA offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for his conception, creation and evolution, helping to present solutions to business problems as well as to find DM business goals. (SAS Institute Inc., 1998).

2.7.3. The CRISP-DM process

Analyzing the problems of DM & KD projects, a group of prominent enterprises (Teradata, SPSS – ISL, Daimler-Chrysler and OHRA) developing DM projects, proposed a reference guide to develop DM & KD projects. This

guide is called CRISP-DM (CRoss Industry Standard Process for Data Mining) (Chapman et al., 2000). CRISP-DM is vendor-independent so it can be used with any DM tool and it can be applied to solve any DM problem.

The CRISP-DM methodology is described in terms of a hierarchical process model, comprising four levels of abstraction (from general to specific): phases, generic tasks, specialized tasks, and process instances.

CRISP-DM defines the phases to be carried out in a DM project. CRISP-DM also defines for each phase the tasks and the deliverables for each task. CRISP-DM is divided into six phases (see Figure 2).

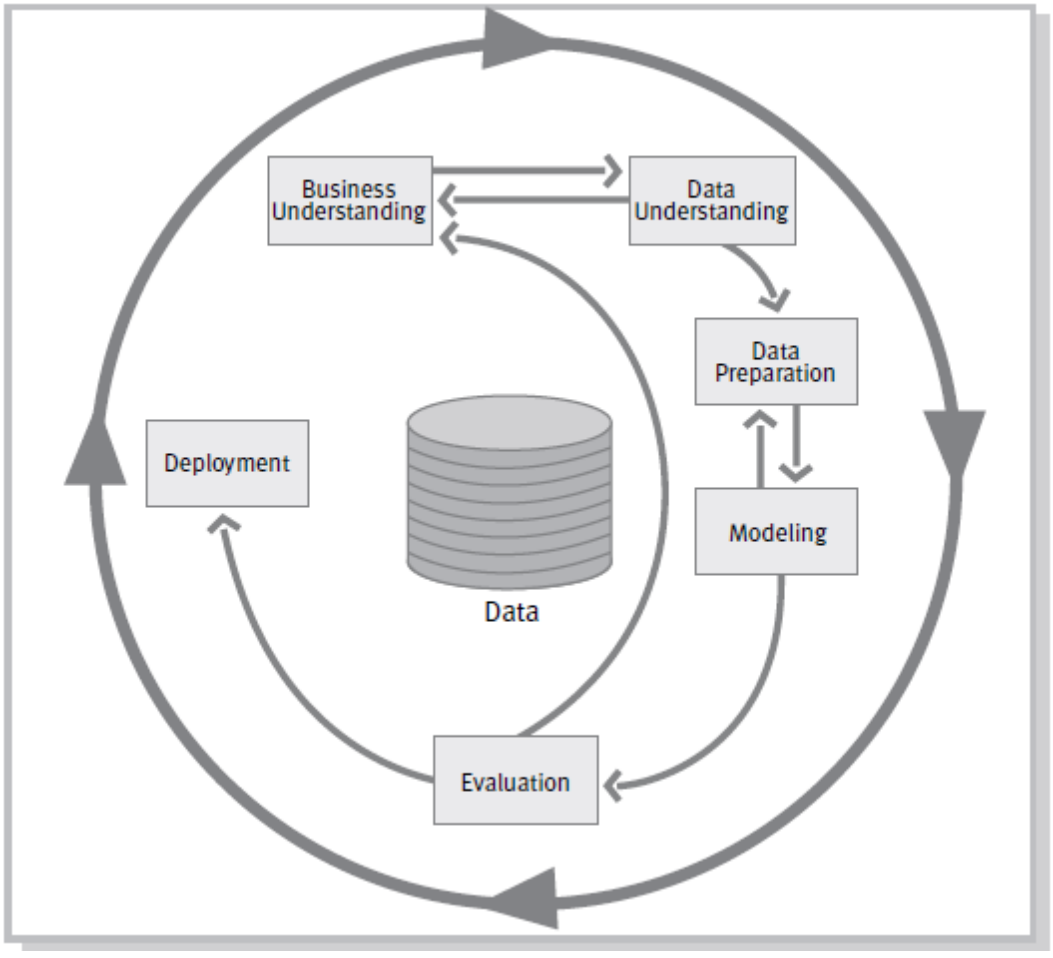

Figure 2.2: Phases of the CRISP-DM reference model(Chapman et al., 2000)

The life cycle of a data mining project is broken down in six phases which are shown in Figure 2.2.The sequence of the phases is not strict. The arrows indicate only the most important and frequent dependencies between

phases, but in a particular project, it depends on the outcome of each phase, or which particular task of a phase, has to be performed next.

The outer circle in Figure 2.2 symbolizes the cyclic nature of data mining itself. Data mining is not finished once a solution is deployed. The lessons learned during the process and from the deployed solution can trigger new, often more focused business questions. Subsequent data mining processes will benefit from the experiences of previous ones(Berry and Linoff, 1997).

Wirth and Hipp (n.d) have outlined each phase of CRISP-DM briefly as follows:

- ***Business Understanding***

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary project plan designed to achieve the objectives.

- ***Data Understanding***

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. There is a close link between Business Understanding and Data Understanding. The formulation of the data mining problem and the project plan require at least some understanding of the available data.

- ***Data Preparation***

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection, data cleaning, construction of new attributes, and transformation of data for modeling tools.

- ***Modeling***

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are

several techniques for the same data mining problem type. Some techniques require specific data formats. There is a close link between Data Preparation and Modeling. Often, one realizes data problems while modeling or one gets ideas for constructing new data…

- ***Evaluation***

At this stage in the project you have built one or more models that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

- ***Deployment***

Creation of the model is generally not the end of the project. Usually, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the user, not the data analyst, who will carry out the deployment steps. In any case, it is important to understand up front what actions will need to be carried out in order to actually make use of the created models.

Considering the presented analysis Azevedo and Lourenço (2008)conclude that SEMMA and CRISP-DM can be viewed as an implementation of the KDD process described by (Fayyad et al, 1996).

The correspondence of the three data mining models is described in Table 2.2 below:

KDD	SEMMA	CRISP-DM
Pre KDD	……..	Business Understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data Preparation
Data Mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	…….	Deployment

Table 2.2 Summary of the correspondences between KDD, SEMMA and CRISP-DM (Azevedo and Lourenço, 2008)

2.8. Application Of Data Mining

The field of data mining has been growing rapidly due to its achievements, scientific progress and broad applicability in various domains like fraud detection, retail, health care, finance, telecommunication, risk analysis etc. (Kumar D. and Bhardwaj D., 2011). The most notable applications areas are listed below:

2.8.1. In Medical Science

Ruben D. (2009) has surveyed the growing number of data mining applications in medicine and public health which includes the analysis of health care centers for better health policy-making, detection of disease outbreaks and preventable hospital deaths, and detection of fraudulent insurance claims.

Antonie, Zaiane, and Coman (2001) mentioned the large scope application of Medical data mining through Diagnosis of diesis, health care, patient profiling and history generation.

As per Kusiak, Kernstine, &Kern (2000), the data mining algorithms significantly reduce patient's risks and diagnosis costs. Using the prediction

algorithms the observed prediction accuracy was 100% for 91.3% cases. The neural networks with back-propagation and association rule mining used for tumor classification in mammograms. Such data mining techniques are effectively used in the diagnosis of lung abnormality that may be cancerous or benign.

2.8.2. Market Basket Analysis

Retailers have been collecting large amount of data like customer information and related transactions, product information etc. This significantly improves the applications like product demand forecasting, assortment optimization, product recommendation and assortment comparison across retailers and manufacturers (Ghani, Probst, Liu, Krema, and Fano, n.d)

Data mining technique is used in MBA (Market Basket Analysis). As information management (2010) explains, Market basket analysis, or MBA for short, is the process of analyzing transaction-level data to drive business value. At this level of detail, the information is very useful as it provides the business users with direct visibility into the market basket of each of the customers who shopped at their store. The data becomes a window into the events as they happened, understanding not only the quantity of the items that were purchased in that particular basket, but how these items were bought in conjunction with each other. In turn, this capability enables advanced analytics such as:

- Item affinity: Defines the likelihood of two (or more) items being purchased together.
- Identification of driver items: Enables the identification of the items that drive people to the store that always need to be in stock.
- Trip classification: Analyzes the content of the basket and classifies the shopping trip into a category: weekly grocery trip, special occasion, etc.

- Store-to-store comparison: Understanding the number of baskets allows any metric to be divided by the total number of baskets, effectively creating a convenient and easy way to compare stores with different characteristics (units sold per customer, revenue per transaction, number of items per basket, etc.).

2.8.3. Data Mining In the Education System

Padhyet al (2012) believe that, with huge number of higher education aspirants, the data mining technology can help bridging knowledge gap in higher educational systems. The hidden patterns, associations, and anomalies that are discovered by data mining techniques from educational data can improve decision making processes in higher educational systems. This improvement can bring advantages such as maximizing educational system efficiency, decreasing student's drop-out rate, and increasing student's promotion rate, increasing student's retention rate in, increasing student's transition rate, increasing educational improvement ratio, increasing student's success, increasing student's learning outcome, and reducing the cost of system processes.

Data mining in distance learning automatically generate useful information to enhance the learning process based on the vast amount of data generated by the tutors and student's interactions with web based distance-learning environment (Luis, Redol, Simoes, &Horta, 2003). As per Deshpande and Thakare (2010), the Data Mining Applications transfers the data into information and feedback to the e-learning environment. This solution transforms large amounts of useless data into an intelligent monitoring and recommendation system applied to the learning process.

As Romero, Ventura & De-Bra (2004) emphasizes, In Web-based Education the data mining methods are used to improve courseware. The relationships are discovered among the usage data picked up during students' sessions. This knowledge is very useful for the teacher or the author of the course,

who could decide what modifications will be the most appropriate to improve the effectiveness of the course.

2.8.4. Data mining in Sports

Solieman (2006) elaborates that, the sports world is known for the vast amounts of statistics that are collected for each player, team, game, and season. There are also many types of statistics that are gathered for each – a basketball player will have data for points, rebounds, assists, steals, blocks, turnovers, etc. for each game. This can result in information overload for those trying to derive meaning from the statistics. Hence, sports are ideal for data mining tools and techniques.

Data mining can be used by sports organizations in the form of statistical analysis, pattern discovery, as well as outcome prediction. Patterns in the data are often helpful in the forecast of future events. Data mining can be used for scouting, prediction of performance, selection of players, coaching and training and for the strategy planning (Solieman, 2006). The data mining techniques are used to determine the best or the most optimal squad to represent a team in a team sport in a season, tour or game (Chodavarapu, n.d).

2.8.5. Data Mining in CRM

According to Ngai et al. (2009), the application of data mining tools in CRM is an emerging trend in the global economy. Analyzing and understanding customer behaviors and characteristics is the foundation of the development of a competitive CRM strategy, so as to acquire and retain potential customers and maximize customer value. Appropriate data mining tools, which are good at extracting and identifying useful information and knowledge from enormous customer databases, are one of the best supporting tools for making different CRM decisions (Berson et al., 2000). As such, the application of data mining techniques in CRM is worth pursuing in a customer-centric economy.

2.8.6. Credit Scoring

Credit scoring has become very important issue due to the recent growth of the credit industry, the credit department of banks face the huge numbers of consumers' credit data to process, but it is difficult analyzing this huge amount of data both in economic and manpower terms (Padhy et al, 2012).
As per the study of Padhyet al (2012), many of the proposed models can only classify customers into two classes “good” or “bad” ones. The most used applied methods for doing credit scoring task are derived from classification technique. Generally classification is used when we predict something which is possible by using the previous available information. It is one type of methods which can be defined as classification where the members of a given set of instances into some groups where the different types of characteristics are to be made. Classification task is very suited to data mining methods and techniques.

2.8.7. Intrusion Detection

Joshi and Pimprale (2013) suggest that network security and Intrusion Detection system (IDS) are having challenging issues with the tremendous growth of information technology. IDS are an essential component of the network to be secured. The traditional IDS are unable to manage various newly arising attacks. The intrusion detection in the Network is very difficult and needs a very close watch on the data traffic. To deal with these new problems of networks, data mining based IDS are opening new research avenues. Data mining is used to identify new patterns which were not known previously from large volume of network dataset. New Intrusion Detection Systems are based on sophisticated algorithms in spite of signature based detection.

The classification method of data mining is used to classify the network traffic normal traffic or abnormal traffic (Cai, & Li, 2004). If any TCP header does not belong to any of the existing TCP header clusters, then it can be considered as anomaly.

With data mining, it is easy to identify valid, useful and understandable pattern in large volume of data. Thus the efficiency and accuracy of Intrusion Detection system are increased and security of network so is also enhanced.

2.9. Related Works

Local research endeavors have been checked from AAU (Addis Ababa university) space site and some efforts including the application of data mining in CRM (Customer relationship management) on Ethiopian airlines by Henock Wubshet, Fraud detection on Africa insurance by Tariku Adane, customer segmentation and prediction by Bleached Regained are found. These papers are industry specific, with different area of concern and with different dataset to be compared with this research.

CHAPTER 3

A SURVEY OF CREDIT RISK ASSESSMENT AT UNITED BANK S.C

3.1. Overview of Credit Risk

According to Basel (1999),Credit risk refers to the risk that a bank borrower or counterparty will default (fail to meet its obligations)on any type of debt by failing to make payments which it is obligated to doing accordance with agreed terms. The risk is primarily that of the lender and includes lost principal and interest, disruption to cash flows, and increased collection costs. The loss may be complete or partial and can arise in a number of circumstances.

To reduce the lender's credit risk, the lender may perform a credit check on the prospective borrower, may require the borrower to take out appropriate insurance, such as mortgage insurance or seek security or guarantees of third parties, besides other possible strategies. In general, the higher the risk, the higher the interest rate that the debtor will be asked to pay on the debt.

3.2. Overview of Credit Risk Assessment

Credit risk assessment/analysis is a largely standardized process that attempts to evaluate the desirability of a particular account based on its estimated reliability and profitability as part and parcel of banks money lending activity. Banks and other lenders conduct credit investigations in order to minimize the probability that they will experience losses from late and delinquent payments (Wills, 2012).

The world of credit risk assessment is a wide one. This is why so many credit assessment institutions can be found all over the world and why each player focuses on a specific population. This specification can be caused by location (National Central Banks for example), portfolio limits (commercial

banks), whether a company is publicly listed or not, has recourse to bond and other traded securities markets or not (international agencies) etc. Due to these different scopes, each player uses the information he/she has access to in order to design the most appropriate rating scale. As a result a variety of definitions of default are in use nowadays (European Committee of Central Balance Sheet Data Offices (ECCBSO), 2007).

The goal of credit risk Assessment as discussed by Basel (1999) is to maximize a bank's risk-adjusted rate of return by maintaining credit risk exposure within acceptable parameters. Banks need to manage the credit risk inherent in the entire portfolio as well as the risk in individual credits or transactions. Banks should also consider the relationships between credit risk and other risks. The effective management of credit risk is a critical component of a comprehensive approach to risk management and essential to the long-term success of any banking organization.

3.2.1. Credit Risk Analysis Metrics

Underlining the importance of understanding the metrics and process of credit risk assessment Wills (2012) has pinpointed the following Credit Risk Analysis Metrics:

- Reliability — Measures of reliability include credit payment history, references from current and past suppliers, and the qualitative character of the management or owners.
- Ability to Pay — financial models and business plans need to demonstrate that the applicant can generate enough revenue and consistent cash flows to make payments within terms. This includes evidence that the business has been (and continues to be) operating successfully and paying its bills on time.
- Economic Conditions — Economic and industry trends contribute to banks' risk assessments as an overall predictor of a business's viability. If an industry is rapidly expanding, that bodes well for a

successful credit arrangement; conversely, if it's shrinking, the bank may err on the side of caution when considering a credit application.

- Collateral — Willingness to back the desired credit terms or loan with asset(s) is a critical consideration in credit risk analysis. If the bank can be assured recourse to recover losses via liquidation of the applicant's property, it has good reason to feel secure in such an arrangement. Secured credit and loans are much more common in difficult economic conditions.

The importance of each metric can vary considerably from applicant to applicant. Not only do they help a lender decide whether or not to issue credit, they also influence payment terms, credit limit, and if there are additional assurances that need to be made.

Riskglossary (2003) has defined Credit risk as risk due to uncertainty in a counterparty's (also called an obligor's or credit's) ability to meet its obligations.

Because there are many types of counterparties—from individuals to sovereign governments—and many different types of obligations—from auto loans to derivatives transactions—credit risk takes many forms. Institutions manage it in different ways.

In assessing credit risk from a single counterparty, an institution must consider three issues (Riskglossary, 2003):

- **Default probability**: What is the likelihood that the counterparty will default on its obligation either over the life of the obligation or over some specified horizon, such as a year? Calculated for a one-year horizon, this may be called the expected default frequency.
- **Credit exposure**: In the event of a default, how large will the outstanding obligation be when the default occurs?

- **Recovery rate**: In the event of a default, what fraction of the exposure may be recovered through bankruptcy proceedings or some other form of settlement?

3.2.2. Credit Risk Functions

Credit Risk specialists are responsible for portfolio/transactional management, including (UniCredit Group, 2012):

- Assigning ratings and coordinating the rating override process;
- Monitoring of credit risks and defining related strategies;
- Defining and controlling credit risk limits;
- Assessing/evaluating large credit transactions;
- Enforcing limits within the approved risk appetite;
- Developing credit risk measurement methodologies;
- Formulating group rules on credit risks;
- Managing credit process harmonization among legal entities;
- Performing credit stress tests;
- Participating in the credit risk regulatory process.

A separate, independent group of specialists manage the Special Credit process, being responsible for:

- Directing the restructuring and workout activities for the Group;
- Setting Special Credit policies and guidelines;
- Managing Group-wide default propagation processes.

3.3. Types of Credit Risk

According to UniCredit Group (2012), Credit risk can be classified in the following way:

1. **Credit default risk** - The risk of loss arising from a debtor being unlikely to pay its loan obligations in full or the debtor is more than 90 days past due on any material credit obligation; default risk may impact all credit-sensitive transactions, including loans, securities and derivatives.

2. **Concentration risk** - The risk associated with any single exposure or group of exposures with the potential to produce large enough losses to threaten a bank's core operations. It may arise in the form of single name concentration or industry concentration.
3. **Country risk** - The risk of loss arising from a sovereign state freezing foreign currency payment (transfer/conversion risk) or when it defaults on its obligations (sovereign risk).

3.4. Introduction to United Bank (UB)

United Bank is incorporated as a Share Company on 10 September 1998 in accordance with the 1960 Commercial Code of Ethiopia and the Licensing and Supervision of Banking Business Proclamation No. 84/1994.

United Bank built itself into a progressive and modern banking institution, endowed with a strong financial structure and strong management, as well as a large and ever-increasing customers and correspondent base. At the end of June 2012, United Bank reported a net profit with a return on equity of 52.83%. Today, United Bank is a full service Bank that offers its customers a full range of commercial banking services with a network that includes 71 branches.

United Bank's priority in the coming years is to strengthen its capital base, maximizing return on equity and benefit from the latest technology in order to keep abreast with the latest developments in the local and international financial services industry. Table 3.1 below shows Figures from five years activity of the bank.

	2007/08	2008/09	2009/10	2010/11	2011/12
Deposits	2,443,352	3,615,752	4,724,855	6,065,827	6,757,616
Loans & Advances	1,859,662	2,152,976	2,613,610	3,276,959	4,085,376
Total Assets	3,250,281	4,625,443	5,896,233	7,725,441	8,786,860
Paid Up Capital	330,277	355,203	373,187	523,298	580,943
Total Capital	467,872	519,975	637,554	901,365	1,101,851
Gross Profit	125,832	133,543	247,667	322,540	406,496
Earnings Per Share	29.51	27.06	47.73	52.81	52.83
No. of Branches	34	40	42	48	63
No. of Employees	1,202	1,358	1,462	1,708	1,975

Table 3.1: Figures for quick reference in '000 Birr

(UB Annual Report,2012)

3.4.1. Services Rendered at UB

United bank avails the following services to its customers:

- Deposits
  - Saving Account
  - Current Account
  - Fixed Time Account
  - Foreign Currency Account
- Loans
  - Short Term Loan
  - Overdraft
  - Letter of Credit
  - Merchandise Loan etc…
- other Services
  - Internet Banking
  - SMS Banking
  - Telephone Banking
  - BLMT (Broad Band Money Transfer)

3.4.2. Basic steps of Loan processing in UB

As per the united bank website (2013), the following are among the basic points to take note in the lending process by customers

1. ***Customers should satisfy the following criteria:***

- ❖ Open accounts in the branch where the credit would be requested.
- ❖ Fill an application form mentioning the objective and the amount of the credit.
- ❖ Supplement a renewed trade license and other relevant licenses.
- ❖ Memorandum and articles of association for businesses with legal personalities.
- ❖ Profile of management members
- ❖ Business plan.
- ❖ Businesses with legal personality should submit audited financial statements preferably done by external auditors
- ❖ Financial statements showing the business and financial position.
- ❖ Balance Sheet.
- ❖ Income Statement.
- ❖ Cash flow statement
- ❖ Other relevant documents as deemed necessary.

2. ***Obligation by loanees***

- ❖ Providing the bank with accurate information on their business.
- ❖ Requesting and enjoying the amount of loans only to the extent of the requirement of their business.
- ❖ Paying regularly the periodic repayment installment and finishing the total proceeds and interest over the agreed terms and duration.
- ❖ Not diverting the loan proceeds to other purpose than the agreed purposes.
- ❖ In case of failures to repay the loan due to problems facing the business, seeking for solution jointly with the bank.

3. ***What is to be done after the loan has been approved.***

- ❖ The loanee or the security holder should sign the loan and collateralization contract.

- ❖ The collateralized assets are to be covered with insurance in the name of the loanee and the loaning branch.
- ❖ The collateralized asset should be registered by state registration and documentation authorities or their agents.

3.5. UB Credit Policy

It is evident that a substantial proportion of the total revenue of all banks in Ethiopia comes from interest on loans and advances. Loans and advances comprise a very large portion of a bank's total assets, and they also form one of the most essential operations of a bank. Wise and prudent policies and procedures in regard to credit management are considered important factors inspiring confidence in depositors and prospective customers of a bank.

The ability of UB to maintain a profitable operation and adequate control base is largely dependent on its capacity to engage in a profitable yet safe lending based on such policies. The process by which the bank seeks, evaluates, extends and monitors its loans is of a paramount importance in the achievement of these objectives.

The use of UB's credit policy (2009) as well as lending procedure Document (credit manual) will enhance efficient customer service, maintain consistency in credit extension, promote transparency and control in credit decision-making.

3.5.1. Credit Functions structure and Approval Authority

Although the organizational structure of the lending functions of a bank varies with its size and types of business, the credit structure of UB shall at all times ensure maximum efficiency in credit processing, clearly delineate

responsibility and accountability, allow effective supervision and monitoring, and ensure efficient credit reporting and control.

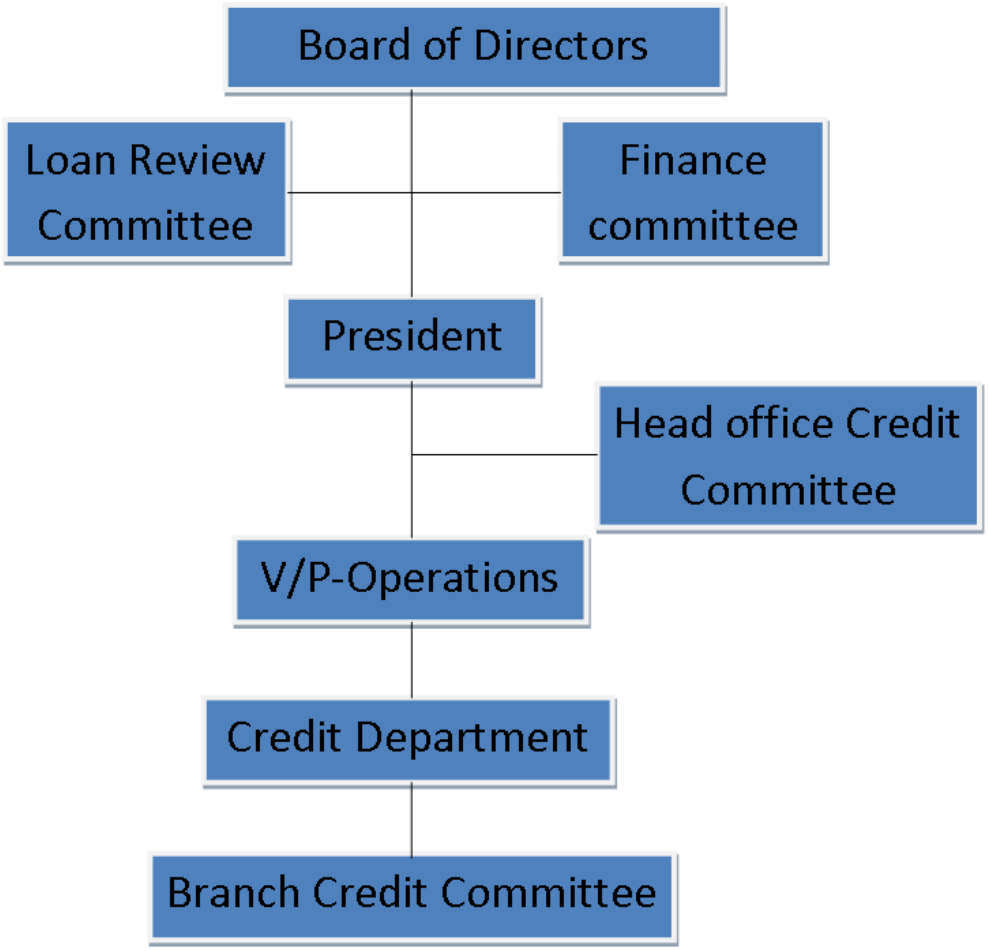

Figure 3.1: United Bank's Credit Functional structure (UB credit policy, 2009)

3.5.2. Forms of Loans and Advances

Loans and advances at UB are made at the following forms:

- Term Loans

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

- Short term loans

It is availed to meet working capital requirements of a business. The loan life is limited to a maximum of two years unless the nature of the business and cash flow pattern of the customer justifies to give a longer repayment period. Short term loan repayments are made by equal monthly, bi-monthly, quarterly or at lump-sum repayments depending upon the cash flow pattern of the business.

- Medium term loans

A medium term loan is granted for purchase of capital goods and to cover the initial working capital requirement of the business; for financing industrial, agriculture, commercial, real estate, construction and development projects. It is also availed for acquisition of trucks, trailers, and public transport buses. Based on the applicants repaying capacity, the repayment of the term loan can be extended up to five years.

- Long term loans

Long term loans are usually financed for big investments requiring long period of grace period before starting to generate cash flow. Such loans have durations exceeding five years and it can be extended up to ten years. Agriculture, mining, industry and building and construction sectors are among the beneficiaries of long term loans.

❖ Overdraft facility

An overdraft facility is a method of loan financing where by a customer is allowed to withdraw money in excess of the balance kept in his current accounts up to certain limits. As a general policy, UB grants overdraft (OD) facilities to firms and/or individuals engaged in profit generating activities.

❖ Advance against import bills

UB avails this facility to businessmen engaged in import. Importers usually request letter of credit (LC) facilities when they face financial constraints to cover payments for LCs to be opened or when they don't want to tie up their funds until the arrival of the import documents. Each advance shall have tenure of 90 days.

- ❖ Pre-shipment advance

Pre-shipment advances are classified into three:

- Pre-shipment advance availed on clean basis
- Pre-shipment finance upon presentation of set of documents
- Pre or post shipment advance availed against NBE guarantee

- ❖ Letter of Guarantee

Letter of guarantee is an unconditional commitment given to a third party on behalf of the bank's customers'. The guarantee issued could be for a local or foreign entity. UB issues the following types of letter of guarantee on behalf of its credit-worthy customers.

- Bid Bond Guarantees
- Advanced payment Guarantees
- Performance Bond Guarantees
- Customers Bond

- ❖ Merchandise Loans

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

3.5.3. Credit Objectives and Economic Sectors to be Served

The primary purpose of extending credit facilities shall be to generate a reasonable return on the funds that the bank extends to customers. Extension of credit shall be undertaken with prudence and paramount concern for maintaining public confidence in the bank at all times. Moreover, UB shall not extend loans to activities which run counter to accepted social or moral values.

UB categorizes loans to different economic sectors and the details of categories of loans currently entertained by the bank are as follows:

1. Domestic trade and services loans (DTS)
2. Agricultural loans
3. Manufacturing production loans
4. Export loans
5. Import loans
6. Building and construction loans
    a. Mobilization fund to contractors
    b. Acquisition of construction machinery and dump trucks
    c. Construction loan for completion of commercial buildings
    d. Staff construction loan
7. Transport loans
8. Hotels & Tourism
9. Personal loans
    a. Personal loan to businessmen
    b. Personal loan for residential house construction
    c. Personal loans to staffs of the bank
10. Non-accrual loans
11. Loans under foreclosure category

3.5.4. Collateral

It is the general credit policy of UB (2009) that all loans to be extended shall be backed by acceptable collateral. Under exceptional cases, however, UB may extend loans and advances with no specific collateral or guarantee, but with the understanding that the business establishment to which the loan or advance is being granted is an exceptionally credit-worthy customer.

The following shall constitute acceptable collateral for UB loans as per its credit policy (2009):

1. Houses and buildings (fixed Assets) (private and office houses, various buildings and ware houses, buildings and houses under construction, buildings transferred to private ownership, and buildings on leased spaces.
2. Motor vehicles (Trucks and trailers, Tankers, Buses, Dump Trucks, Automobiles)
3. Business Mortgage (Good will, Trade name, patent or copy right, movable properties of a business etc.)
4. Cash Deposits
    a. Saving Accounts
    b. Current Accounts
    c. Fixed Time Deposits
5. Merchandize
6. Negotiable Instruments (Treasury bills, government bonds, cash surrender values of life insurance policy, Warehouse receipts)
7. Share Certificates
8. Guarantees (Bank Guarantee & Personal Guarantee)
9. Second degree mortgage

3.5.5. Credit Follow-up and Review

Credit risk is one of the most worrisome of all risks which a bank faces because of the importance of the potential loss it may entail. UB put in place a sound credit follow-up and review system, in order to maintain the

quality of its loan portfolio and to ensure compliance with credit policies, directives and regulations.

Credit follow-up shall encompass analysis of performance in relation to such variables as trends in growth and portfolio mix, trends in loan quality as indicated by NBE's directives on loan classification, delinquencies, non-performing loans, write-offs, etc.

Loan review shall be carried out regularly and on an ongoing basis on all loans and advances.

- To monitor the quality of the Bank's loan portfolio.
- To prevent deterioration of loan portfolio.
- To identify early warning signals.
- To check the compliance with UB's policies and NBE regulations.

3.5.6. Classification and loan loss provisioning

All loans of the bank for the purpose of provisioning shall be classified according to the standards set by NBE. The minimum provisioning for loans and advances shall be according to Table 3.2 below:

Category	***Minimum provision***	***Consecutive Past due days (x)***
Pass	1%	x<30
Special mention	3%	30<=X<90
Substandard	20%	90<=X<180
Doubtful	50%	180<=X<360
Loss	100%	x>=360

Table 3.2: Loan categoryas per the elapsed repayment days

Provision to be held is calculated on the principal outstanding balance and before applying the minimum provision percentage laid out above, UB deducts from the outstanding loans and advances cash collateral and substitutes held as collateral, and in the case of loans and advances secured by physical collateral.

3.5.7. Regulatory Body Regulations

The Bank offers credit facilities only to legitimate commercial and personal activities. When doing so, the Bank complies with all applicable laws and regulations of the government and all directives and guidelines issued by the supervisory authority (NBE).

3.6. UB Property Estimation Guideline

One of the major businesses of commercial banks is availing loans and advances to viable ventures based on financial analysis conducted to determine the performance of the business. It is well known that analyzing the performance of the business is very crucial to determine possible financial risks. It is evident that screening deserving loan applicants will help to minimize loan losses, but it is not totally avoid risks. Therefore, to fill this gap properties are usually held as collateral.

A well prepared or designed property estimation guideline is very important to estimate property values in a transparent and precise way to mitigate credit risks that may ensure closing the said gap.

Valuation is the art of assessing fair value of property in terms of money at a given time. It requires a close analysis of various factors like trends in inflation, cost of capital, and cost of labor, location, and suitability for the intended purpose etc. It is a complex process which highly depends on expertise in the field with adequate knowledge of the current property market.

There are three methods of estimation:

1. Investment worthiness: Two basic real-estate capital inputs i.e. land and construction costs are to be major considerations.
2. Rent Based Capitalization: Assumes the income produced or the potential to produce by renting the property.
3. Replacement cost method: Calculated by taking the current unit construction cost per square meter or cubic meter of a building plus average current value of land.

3.7. Credit Risk Grading

United Bank has a credit risk grading mechanism to mitigate the risk of default and inconsistent repayment. There are about five factors of risk grading as indicated in Table 3.4 which are used to discriminate risky loan applicants by the credit officers of branches. Unfortunately these factors are not followed strictly and the related data is not properly captured by the core banking and CIS (credit information system) of the bank.

No.	Risk Grading Factors	Details	Allotted point
1	Account Performance	Term loans, OD (overdraft), LC (Letter of credit), Pre-shipment loans etc…	30
2	Financial Standing	• Type of financial stat. ○ Audited/Provisional/CCR • Liquidity Risk (Current Ratio, Quick Ratio) • Gearing Ratio (Debt to Asset) • Operational Risk	30
3	Management Quality	• Experience • Qualifications	15
4	Banking R/Ship	Relationship Length	10

5	Collateral	• Collateral Coverage • Realize ability	15
	Total		100%

Table 3.4: Risk grading factors used by branches

3.8. UB Core Banking Solution

United Bank has implemented FLEXCUBE Core Banking Solution since 2006.

Oracle FLEXCUBE Core Banking offers comprehensive technical features to support the full spectrum of banking operations—from small community banks to large, multinational financial institutions (Oracle Financial Services, 2009).

Supported Products and Services:

- ❖ Current and savings accounts (CASA)
- ❖ Deposits
- ❖ Loans
- ❖ Collections
- ❖ Standing instructions
- ❖ Payments (Society for Worldwide Interbank Financial Telecommunication, or SWIFT, accredited) etc…

3.8.1. Flexcube Implementation Of Loans And Advances

The FLEXCUBE implementation of loans and advances captures the day to day credit activities of the branches and automated contract interest calculations and status tracking.

FLEXCUBE having the LD (Loans and Deposits) module enables the bank to capture the customer information, define loan products, disburse the approved amount for customers, calculate interest on daily basis, schedule the principal and interest repayment schedule, and amend the loan when the need arises, liquidate the loan through periodic payments or once and tracks the status of the loan based on the repayment schedule and actual repayment period differences.

3.9. UB Credit Information System

Credit Information is also captured in an external system as per the requirement of National Bank of Ethiopia (NBE). This database has information about customers which is not linked in the core banking database.

Hence, the blended information from these two systems will give a nearly complete view of the loans and advances in UB.

CHAPTER 4

EXPERIMENTATION

4.1. Overview

In this chapter, the researcher depicts the actual application of data mining process in a stepwise fashion on the credit data of United Bank.
In line with CRISP-DM the following steps of data mining are followed.

Figure 4.1: CRISP-DM steps

The Business understanding step focuses on producing a project plan by determining business objectives, Assess situation and determining data mining goals.

4.1.1. Business Objectives

Credit facilities and investments are the cornerstones of the growing economy of Ethiopia. United Bank being one of the former private banks has played its own role in the economy by rendering loan facilities to the individuals and companies which are running business in various sectors. The bank uses internal and NBE credit policies, procedures and strictly followed manuals in various levels of credit committees before disbursing loan to customers. However, there are total defaulters and inconsistent loan repaying customers which declines the profitability of the bank in particular and threatens the growing economy of the country in general. While fueling the sprinting economy in the country, minimizing the possible defaulters is the prime concern of the bank.

4.1.2. Data Mining Goals

Data mining is an innovative way of gaining new and valuable business insights by analyzing the information held in a company database (IBM, 2013).

Thus the overall goal of this data mining process is to extract information from United Bank's Credit data set and transform it into new, valuable and an understandable structure by identifying risky credit scenarios and facilitate the making of well-informed business decisions.

This data mining will be applied on the bank's credit related information extracted from Flexcube Core banking system and is expected to identify rules that will assist to classify contracts as good and bad loans.
By uncovering patterns of credit failure and success through data mining the bank will benefit in many ways:

- Will enable the loan officers to scrutinize control on the existing loans which do have a bad loan pattern.
- Will be able to avoid/screen expected bad loan applicants and target a marketing campaign for possibly good loan applicants.
- Can lead to an improvement in the quality and dependability of strategic business decision making with regard to bank's credit activity.

4.1.3. Data Mining Tool Selection

There are various potential commercial and open source tools which are available for data mining. Weka is selected as a matter of convenience and earlier experience.

4.2. Data Understanding

Having an insight about the need of data mining the next basic thing for the process is getting the credit data and creating an understanding for it. Data

understanding step of CRISP-DM has different components of learning before the actual application of data mining techniques.

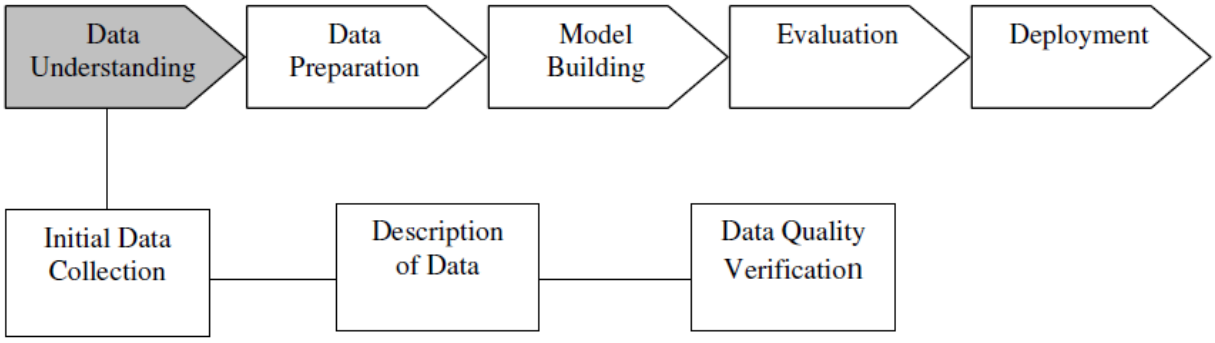

Figure 4.2: The data understanding phase of CRISP-DM

4.2.1. Initial Data Collection

The data for this research has been compiled from United Bank Core banking system. The core banking system contains numerous tables regarding credit activities and many other modules of the bank. Among these tables the following are considered to be relevant for this thesis:

1. Customer Base table
2. Loan contract record master table,
3. Loan contract schedules,
4. Historic accounting transactions table containing any transaction related with the loan contract like disbursement and regular repayment amounts and non-performing loan movements and
5. Loan contract event table for the daily events.

4.2.2. Description of Data

Only selected fields and records as shown below in Table 4.1, Table 4.2, Table 4.3, Table 4.4 and Table 4.5 are to be used for the task of this thesis. Having an explanation of each table and relevant fields will give a glimpse of the dataset to be constructed from these tables.

1. Customer Base table: This table contains basic customer's information without attaching to the account type they are opening. This table contains a unique ID called CIF (Customer Identification File) which is used as a basis to open any type of account for the customer. This table contains around 81 fields. Out of which the following are considered important for this thesis:

Field Name	Data Type	Description
CUSTOMER_NO	VARCHAR2	Unique Identifier for any customer (CIF No)
CUSTOMER_TYPE	CHAR	Identifies the customer as Individual, Corporate or Financial organization
RECORD_STAT	CHAR	Is it open or closed

Table 4.1: Customer Base table selected fields

2. Loan contract record master table: This table contains most of the loan contract information based on the paper signed as a binding agreement between the bank and the customer while granting the credit facility. There are about 95 fields in this table, out of which the following will be used for this thesis.

Field Name	Data Type	Description
CONTRACT_REF_NO	VARCHAR2	Unique Identifier for the contract by a customer
PRODUCT	VARCHAR2	Classification based on product type and tenor of loan
PAYMENT_METHOD	CHAR	
COUNTERPARTY	VARCHAR2	CIF No for the

		customer
LCY_AMOUNT	NUMBER	Amount granted in local currency
VALUE_DATE	DATE	Disbursement date of the loan
MATURITY_DATE	DATE	The Final Liquidation date
TENOR	NUMBER	Expected Lifetime of the loan in Number of days
USER_DEFINED_STATUS	VARCHAR2	The status of the loan based on repayment activity
CONTRACT_STATUS	CHAR	Active, closed, reversed etc…
MAIN_COMP_RATE	NUMBER	The interest rate for the contract
RISK_FREE_EXP_AMOUNT	NUMBER	Collateral Amount in Cash value
LOCATION (Derived)	VARCHAR2	Northern Ethiopia, A.A, others (South, East, West)

Table 4.2: Loan Contract Master Table selected attributes

3. Loan contract schedules: This table is included primarily for the sake of frequency of payment like Monthly, quarterly etc.

Field Name	**Data Type**	**Description**
CONTRACT_REF_NO	VARCHAR2	Unique Identifier for the

		contract by a customer
FREQUENCY	CHAR	Frequency of payment

Table 4.3: Loan Contract schedules Table selected attributes

4. Accounting transactions table: This table holds every detail of any transaction which are generated by the system or passed by users. There are 46 fields in this table out of which following are found important for this thesis.

Field Name	**Data Type**	**Description**
TRN_REF_NO	VARCHAR2	Unique Identifier for the transaction (contract_ref_no for contracts)
AC_NO	VARCHAR2	New status of the loan (GL)

Table 4.4: Accounting transactions Table selected attributes

5. Loan contract event log table: This table contains any event record on the contract like accrual, amendment, Liquidation etc. This table has 14 fields, out of which the following ones are relevant here.

Field Name	**Data Type**	**Description**
CONTRACT_REF_NO	VARCHAR2	Unique Identifier for the contract by a customer
EVENT_CODE	VARCHAR2	Event codes for anything on the contract

Table 4.5: Loan contract event log table selected attributes

4.2.3. Data Quality Verification

From the above description it can be seen that extracting data needs joining different tables and dropping redundant fields in order to come-up with a reliable set of columns and content for the data-mining task. Moreover certain fields have null values and others need categorization to meaningful elements.

4.3. Data Preparation

As per Han and Kamber (2006), preprocessing assists to fill some missing values; to detect some outliers that may jeopardize the result of data mining; and to detect and remove/correct some noisy data. In relation to this, data normalization, discretization and related activities need to be performed. Moreover, to conduct the experimentation, the dataset must be prepared in the appropriate format.

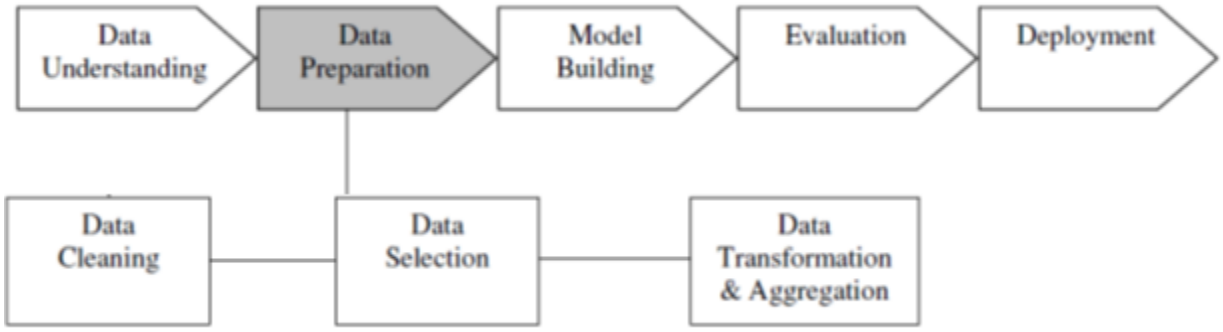

Figure 4.3: CRISP-DM Data Preparation phase

As indicated in the Figure 4.3, this phase included data cleaning, selection, transformation and aggregation, integration, and formatting.

4.3.1. Data Selection

On this phase, the relevant data for the data mining process is selected. The cornerstone of selection of data was based on relevance, availability and quality of variables.

The most challenging part of this step was getting the relevant information as needed. Most of the core banking reports is designed for daily routine and periodic administrative decisions per branch and specific products. Getting credit information bank wide for the whole period and fiscal year needs a tailored approach and exposure to the core banking system.

The core banking system, stacked with various tables and views is not an easy prey to get the desired information without losing focus. There were more than 3,500 tables and 600 views in the Core banking system. Identifying the important ones and further reducing the number of tables to minimize the number of joins has taken extended time. Finally the tables and views mentioned in the data description part were selected with the relevant fields.

While selecting the data the following criteria were used:

- All the 5 tables were used in one query to generate the data
- The final modified versions of each loan contract were used
- Only contracts of Loan Module were used.
- NPL (Non-Performing Loan) GL entries were counted per contract
- The following initial field names were used to extract data from Oracle database to MS_EXCEL sheet using “SQL Navigator” tool:

CUSTOMER_NO,CUSTOMER_TYPE,RECORD_STAT,CONTRACT_REF_NO,PRODUCT,PRODUCT_DESCRIPTION,PAYMENT_METHOD,LCY_AMOUNT,VALUE_DATE,MATURITY_DATE,TENOR,USER_DEFINED_STATUS,CONTRACT_STATUS,MAIN_COMP_RATE,RISK_FREE_EXP_AMOUNT,LOCATION,FREQUENCY,NPL_COUNT

These fields are the initial attributes that are used for further processing of the 27,439 extracted data.

4.3.2. Data Cleaning

The data cleaning task was partially managed in the data extraction step. While making an inner join and setting certain filtering criteria, the invalid or incomplete data was suppressed before extraction. The researcher has

used “SQL Navigator 3.0” to extract data from oracle database of United Bank core banking system. The following data mining activities were performed:

- The missing values of CUSTOMER_TYPE field were substituted with the default value (I=Individual).
- 1,418 records of reversed loans were avoided as they create duplication. These Loans are recaptured to the system with modifications after reversal. Since the reversed contract doesn’t stay to the full life time of the contract (as it is replaced with the new one), removing it has a positive impact on the final dataset analysis.
- The VALUE_DATE which is the actual loan granting (disbursement) date and the MATURITY_DATE, which is the final Liquidation date of the contract, have little or no importance to the final analysis as there is TENOR field which is the lifetime of the contract in number of days. TENOR is nearly identical to the difference of MATURITY_DATE and VALUE_DATE. Hence, the whole columns of MATURITY_DATE and VALUE_DATE are removed.
- PAYMENT_METHOD attribute which was supposed to depict last principal payment method has the same values for the entire contract. Such records have no value for the data-mining algorithms. So the corresponding column was removed. Actually the more important field for the domain experts regarding payment is the repayment frequency which entails monthly, quarterly, Semi-Annually or annually which is represented by FREQUENCY attribute.
- The CUSTOMER_NO and CONTRACT_REF_NO are used to uniquely identify the customer and loan contract respectively. They are not as such important to the data mining process and are primarily removed for the confidentiality of the data to be used for the analysis.
- Three (3) records of USER_DEFINED_STATUS attribute were having missing value. Further analysis has been done on the transaction of

the contracts and the missed values were determined to have values of “NORM”.

- The CONTRACT_STATUS field which shows the current status of the loan has an option Active (A), Liquidated (L) and Hold (H) statuses. But the Hold (H) status is to mean that the contract is under modification and the modification is saved but not authorized. So the researcher has categorized the Hold status contracts (18 in number to Active contracts, as the modification happens only to Active contracts.
- Eleven (11) records of “TENOR” attribute were missing. 4 of the values were generated by subtracting maturity date from value date attribute values. The remaining 7 records which have also missing information on maturity date or value date and 3 records with zero value were removed from the dataset as there number is few as compared to the whole dataset and as there is no easy way to populate the content.
- The MAIN_COMP_RATE which is the interest rate attribute has 1,444 missing values. 330 of the missing values were having a product of EMERGENCY STAFF LOAN product category. And all of the EMERGENCY STAFF LOANs are non-interest bearing loans, so 0.00 values were generated. The remaining records were given values by loading the excel CSV file to oracle table and creating a match for the contract product group, amount, tenor and loan granting value date. After this the null values were updated from the corresponding interest rule table.
- The RISK_FREE_EXP_AMOUNT attribute which is the cash equivalent collateral amount which is held by the bank has 2,085 missing values. Fortunately the missing values were from EMERGENCY STAFF LOAN, STAFF SALARY ADVANCE, and ADVANCE ON IMPORT BILLS which are zero by default. So Zero amount is updated. 1 record has an outlier figure of more than one billion (1,309,707,021.00) in this attribute. This seems a wrong figure and the record has been deleted.

- The LCY_AMOUNT which is the amount granted for the debtor has 100 records which were less than 500 birr. The amount which is less than 500 has no business logic. Hence, these amounts are expected to be recaptured loans per outstanding interest or principal. The researcher has removed these values.

4.3.3. Data Transformation and Aggregation

Data transformation and aggregation assists in reducing the variations of field values and changes to a meaningful and understandable form.

- The Field "LOCATION" was generated as a derived field from the CONTRACT_REF_NO by taking the first number of the reference, as this indicates the geographic location of the debtor and creditor branch. The bank has classified its debtors into 3 regions namely, Addis Ababa, north and others (East-West-South) of Ethiopia.
- Fiscal year has an influence on interest rate and business activity of banks due to the global impact and macro-economic shifts. Hence a derived field of BOOKING_YEAR, which is the disbursement year of the contract, is generated from the CONTRACT_REF_NO, by taking the 8th and 9th numeric values of the contract.
- The product field has many related but different values. So the researcher has mixed the related products into one as follows:

PRODUCT_DESCRIPTION	PRODUCT
ADVANCE AGAINST EXPORT BILL	ADVANCE AGAINST EXPORT BILL
ADVANCE ON IMPORT BILLS	ADVANCE ON IMPORT BILLS
BUILDING & CONS - PROJECT -NORMAL	BUILDING and CONSTRUCTION
BUILDING and CONSTRUCTION - MEDIUM	
BUILDING and CONSTRUCTION - SHORT T	
DOMESTIC TRADE & SER - PROJECT NORM	DOMESTIC TRADE SERVICE
DOMESTIC TRADE SERVICE - MEDIUM	
DOMESTIC TRADE SERVICE - MEDIUM-NOR	
EXPORT MEDIUM TERM	EXPORT
EXPORT MEDIUM TERM-NORMAL WITH BEAR	
EXPORT SHORT TERM	

HEALTH SERVICE - MEDIUM TERM	
HEALTH SERVICE - SHORT TERM	
HEALTH SERVICE -ML-NORMA	HEALTH SERVICE
HOTEL and TOURISM SERVICE -MEDIUM	
HOTEL and TURISM SERVICES - SH	HOTEL and TOURISM
IMPORT - MEDIUM TERM	
IMPORT - SHORT TERM	IMPORT
MANUFACTURING - MEDIUM TERM	
MANUFACTURING - PROJECT -NORMAL	
MANUFACTURING - SHORT TERM	MANUFACTURING
MERCHANDIZE - DTS	
MERCHANDIZE - EXPORT	
MERCHANDIZE - IMPORT	
MERCHANDIZE - MANUFACTURING	
MERCHANDIZE LIMIT - EXPORT	
MERCHANDIZE LIMIT - IMPORT	
MERCHANDIZE LIMIT - MANUFACTURING	
MERCHANDIZE LOAN LIMIT - DTS	MERCHANDIZE
PERSONAL - MEDIUM TERM	
PERSONAL - MEDIUM TERM-NORMAL	
PERSONAL - PROJECT -NORMAL	
PERSONAL - SHORT TERM	
PERSONAL LOAN STAFF - A	
PERSONAL LOAN STAFF - B	
PERSONAL LOAN STAFF - C	
PERSONAL LOAN STAFF - C-Normal	PERSONAL
PRE-SHIPMENT ADVANCE AGAINST NBE GU	
PRE-SHIPMENT ADVANCE LIMIT	
PRE-SHIPMENT ADVANCE ON CLEAN BASIS	PRE-SHIPMENT ADVANCE
PROJECT LOAN	
PROJECT LOAN- NORMAL	
PROJECT-AGRICULTURE	
PROJECT-BUILDING & CONS	
PROJECT-DOMESTIC TRADE SERVICE	
PROJECT-EXPORT	
PROJECT-IMPORT	
PROJECT-PERSONAL	PROJECT
STAFF SALARY ADVANCE	
EMERGENCY STAFF LOAN	EMERGENCY STAFF LOAN
TERM LOAN - DTS - SHORT	
TERM LOAN - DTS - SHORT-NORAML WITH	DTS
TRADE BILLS DISCOUNT	
TRADE BILLS DISCOUNTING LIMIT	TRADE BILLS DISCOUNT

TRANSPORT - MEDIUM TERM	
TRANSPORT - PROJECT -NORMAL	
TRANSPORT - SHORT TERM	TRANSPORT

Table 4.6: Transformed products

The NPL_COUNT attribute shows the entries which were passed per contract when the debtor fails to pay on time or defaults. A debtor who didn't pay for more than 3 months for a monthly basis payment will have a non-zero count on this field. Those having zero count are classified as "GOOD" (risk free loans) and those greater than Zero as "BAD" (risky loans). Thus NPL_COUNT is used to determine the class and renamed to STATUS and the MAIN_COMP_RATE field is renamed to INTEREST_RATE for ease of understanding.

4.3.4. Final Dataset Preparation

After passing the above preprocessing stage the researcher has got 27,310 records. Then the preprocessed dataset in excel is converted to Comma Separated Values (.csv) and Attribute Relation File Format (.arff) to make it compatible with WEKA software.

The STATUS field which was chosen for the class attribute has an imbalanced count of records. The count of BAD loans is 5,417 (19.83%) while the GOOD ones are 21,893 (80.2%).

According to Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, P. (2002), a dataset is imbalanced if the classification categories are not approximately equally represented. Performance of DM algorithms is typically evaluated using predictive accuracy. However, this is not appropriate when the data is imbalanced and/or the cost difference of error is large.

Under-sampling of the majority (normal) class has been proposed as a good means of increasing the sensitivity of a classifier to the minority class. However, Chawla, N. et al. (2002), showed a method of over-sampling the minority (abnormal)class and under-sampling the majority (normal) class can achieve better classifier performance (in ROC space)than only under-sampling the majority class.

The Credit Data of UB has a higher imbalance for the class variable (STATUS). Therefore, the researcher used re-sampling of Weka (weka.filters.supervised.instance.Resample) to over sample the minority classes (BAD instances) and under sample the majority (GOOD instances) of loans. As a result, the class distribution in the dataset changes and probability of correctly classifying the instances of the class increases.

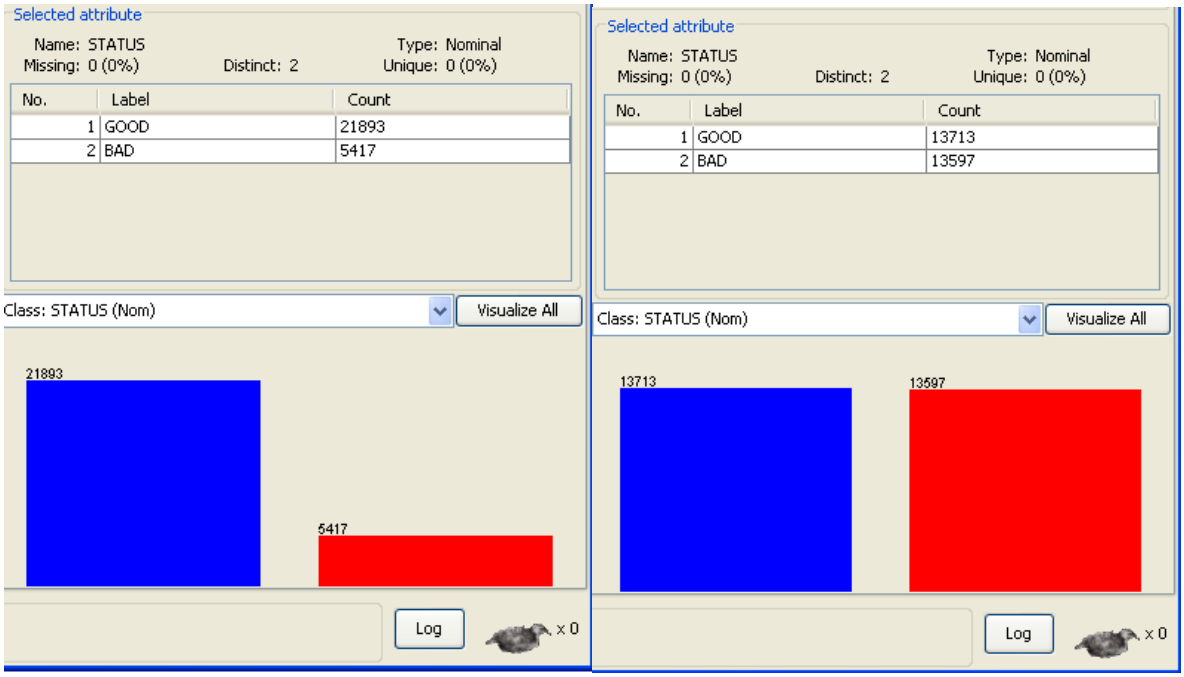

Figure 4.4 a. Imbalanced Data b. balanced data after resampling

Figure 4.4 depicts a comparative view of the class variable before (a) and after (b) resampling, which is over sampling the minority which has moved the instances from 5,417 to 13,597and under sampling the majority diminishes from 21,893 to 13,713.

Finally four of the numeric attributes namely TENOR, INTEREST_RATE, RISK_FREE_EXP_AMOUNT and LCY_AMOUNT have been discretized with bin value of 10 each.

4.4. Modeling

Modeling phase of data mining is conducted on the pre-processed data by selecting the modeling technique, generating test design, building model and assessing and selecting the best model.

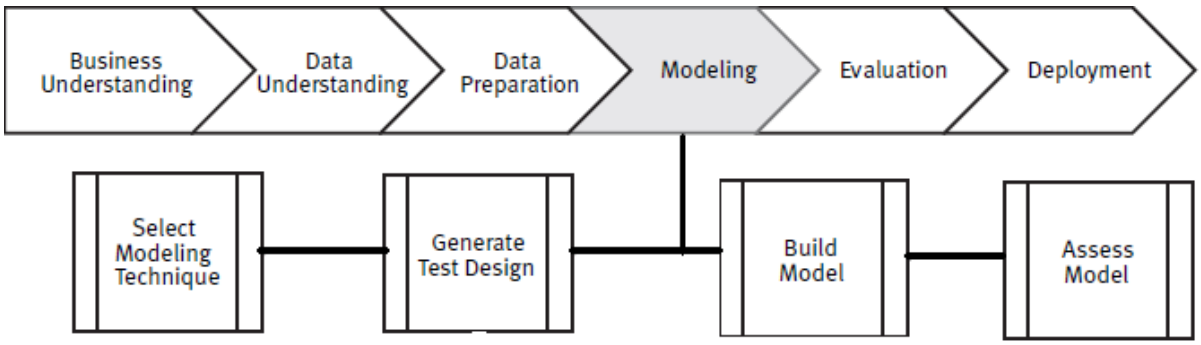

Figure 4.5: The CRISP-DM Modeling phase

4.4.1. Selection Of Modeling Techniques

The major business objective of this study is to minimize the number and amount of non-performing loans (bad loans) by proactively identifying or predicting the future status of the loan from the mix of variables during application or after disbursement for committed follow-ups. So this research applies predictive model using the J48 decision tree algorithms and Naïve Bayes classification techniques of WEKA 3.6.6.

4.4.2. Test Design

According to CRISP-DM (2000), a plan should be first set to guide the training, testing and evaluation process of the model. Mostly researchers split the dataset into training and test sets. Normally, training should be done on large proportion of the total data available, whereas testing is done

on small percentage of the data that has been excluded during training of the model.

Accordingly, all the resampled dataset has been used for training and testing of this research. While conducting the J48 decision tree and Naïve Bayes classification models, different experiments have been done by splitting the dataset into training and testing set and by adjusting the default parameter values. Finally, the classification model that shows better accuracy performance has been selected.
The interpretation of the rules generated by the selected model was made by the researcher and the domain experts of the bank. The Loan officers of selected branches and the credit department staffs were selected as domain experts for this research.

4.4.3. Building Classification Model

In order to construct classification model, the decision tree (in particular theJ48 algorithm) and the Naïve Bayes methods are selected. To classify the records based on their values for the given cluster index, the model is trained by changing the default parameter values of the algorithms. The main interest here is to find rules that predict the class labels which are novel to the domain experts.

4.4.3.1. J48 Decision Tree Model Building

J48 algorithm is used for building the decision tree model and contains some parameters that can be changed to further improve classification accuracy.
The algorithm used: weka.classifiers.trees.J48

Option	Description
binarySplits	Whether to use binary splits on nominal attributes

	when building the trees.
confidenceFactor	The confidence factor used for pruning (smaller values incur more pruning).
debug	If set to true, classifier may output additional info to the console.
minNumObj	The minimum number of instances per leaf.
numFolds	Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree.
reducedErrorPruning	Whether reduced-error pruning is used instead of C.4.5 pruning.
saveInstanceData	Whether to save the training data for visualization.
seed	The seed used for randomizing the data when reduced-error pruning is used.
subtreeRaising	Whether to consider the subtree raising operation when pruning.
unpruned	Whether pruning is performed.
useLaplace	Whether counts at leaves are smoothed based on Laplace.

Table 4.7: The J48 Decision Tree Model building parameters

The Experiment has been done by changing default values of the above parameters progressively. The table below shows some of the experiments which are found to be relevant for comparison. The background of the test mode variable is highlighted with grey scale and subsequent changes are marked with bold face.

Exp. No	Cross validation Folds	% Split for Training	Confidence Factor	Unpruned	F Measure	Tree size	Time (sec)
1	**10**	66	0.25	False	0.956	1496	0.33
2	10	66	0.25	False	0.949	1496	0.18
3	10	70	0.25	False	0.949	1496	0.22
4	10	75	0.25	False	0.951	1496	0.03
5	10	80	0.25	False	0.953	1496	0.03
6	10	85	0.25	False	0.953	1496	0.03
7	10	90	0.25	False	0.948	1496	0.01
8	10	95	0.25	False	0.949	1496	0.01
9	10	66	0.25	**True**	0.966	3280	0.15
10	10	66	0.25	**True**	0.959	3280	0.04
11	10	90	0.25	**True**	0.962	3280	0.01
12	10	66	**0.20**	**True**	0.966	3280	0.12
13	10	66	**0.15**	**False**	0.950	840	0.12

Table 4.8: Input parameters and resulting J48 Decision Tree.

The above J48 decision tree experiments were basically done by using cross validation and % split test modes.

Classification using the cross validation folds has been tested by the default value of confidence factor (0.25) and got 95.6% of accuracy. Other confidence factors above and below 0.25 yields less than 95.6% of accuracy for the 10 fold cross validation. Changing the value of unpruned parameter to TRUE and FALSE, while changing the confidence factor has not given better result than the default options.

The other test mode that has been tested with various combinations is the % split for training. In this regard the 66% split for training and 0.25 confidence factor with unpruned tree has given the highest accuracy of 96.6% which is not exceeded by any other combinations of parameters during the experiment.

J48 Decision Tree Model with the highest accuracy has the following Confusion Matrix and related information:

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	26371	96.5617 %
Incorrectly Classified Instances	939	3.4383 %
Kappa statistic	0.9312	
Mean absolute error	0.0396	
Root mean squared error	0.1622	
Relative absolute error	7.922 %	
Root relative squared error	32.4389 %	
Coverage of cases (0.95 level)	98.9381 %	
Mean rel. region size (0.95 level)	54.0205 %	
Total Number of Instances	27310	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.961	0.030	0.970	0.961	0.966	0.931	0.988	0.985	GOOD
0.970	0.039	0.961	0.970	0.966	0.931	0.988	0.981	BAD

Weighted Avg. 0.966 0.034 0.966 0.966 0.966 0.931 0.988 0.983

=== Confusion Matrix ===

a b <-- classified as
13177 536 | a = GOOD
403 13194 | b = BAD

Table 4.9: J48 DT accuracy, confusion matrix & summary

The above confusion matrix of J48 decision tree depicts that of 13,713 GOOD contracts 13,177 are classified as GOOD (96.10%) and the actually good 536 were classified as BAD loans (3.90%). On the other hand out of the resampled 13, 597 BAD loans 13,194 were classified as bad (97.04%) and 403 of them were wrongly classified as GOOD loans (2.96%). This entails that the records with the GOOD class are classified with higher error.

4.4.3.2. Naïve Bayes Model Building

Naïve Bayesian Classifier is a machine-learning algorithm that maps (classifies) a data example into one of several predefined classes.
The algorithm used: weka.classifiers.bayes.NaiveBayes
The experiment is done by changing cross validation folds and % split test modes and their underlying options.

Exp. No	Cross validation Folds	% Split for Training	Precision	Recall	F Measure	ROC area	Time (sec)
1	**10**	**66**	**0.897**	**0.885**	**0.884**	**0.980**	**0.05**
2	10	66	0.896	0.882	0.882	0.981	0.05
3	10	70	0.895	0.881	0.880	0.981	0.03
4	10	80	0.896	0.883	0.882	0.981	0.03

5	10	90	0.894	0.883	0.882	0.980	0.03

Table 4.10: Naïve Bayesian classification experiment result

The experiment depicted above shows that the Naïve Bayes classification algorithm with a 10 fold cross validation works good in terms of precision, recall and F-Measure. The classification which has been done changing the cross validation fold value or % split for training value has decreased accuracy.

Time taken to build model: 0.05 seconds

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

Correctly Classified Instances	24167	88.4914 %
Incorrectly Classified Instances	3143	11.5086 %
Kappa statistic	0.77	
Mean absolute error	0.1121	
Root mean squared error	0.3069	
Relative absolute error	22.4161 %	
Root relative squared error	61.3815 %	
Coverage of cases (0.95 level)	94.8041 %	
Mean rel. region size (0.95 level)	55.4211 %	
Total Number of Instances	27310	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.798	0.028	0.967	0.798	0.874	0.782	0.980	0.980	GOOD
	0.972	0.202	0.827	0.972	0.894	0.782	0.980	0.982	BAD
Weighted Avg.	0.885	0.114	0.897	0.885	0.884	0.782	0.980	0.981	

=== Confusion Matrix ===

a	b	<-- classified as
10949	2764	a = GOOD

379 13218 | b = BAD

Table 4.11: Confusion matrix and related information of Naïve Bayes model

The above confusion matrix of Naïve Bayes depicts that of 13,713 GOOD contracts 10,949 are classified as GOOD (79.84%) and the actually good 2,764 were classified as BAD loans (20.16%). On the other hand out of the resampled 13, 597 BAD loans 13,218 were classified as bad (97.21%) and 379 of them were wrongly classified as GOOD loans (2.79%). This entails that the records with the GOOD class are classified with higher error.

4.4.4. Selecting Classification Model

After the experimentation of this study with the J48 decision tree algorithm and Naive Bayes classifier DM model, comparing and selecting the one which performs the best is one of the deliverables of this phase.

All the experiments were carried out with the same dataset. From the output these experiments the highest accuracy is found by the J48 decision tree method. From Table 4.8 and 4.11 it can be seen that all the J48 decision tree algorithm experiments have performed better than the Naïve Bayes classifier method.

4.4.4.1. Generating Rules from J48 Decision Tree

The experiments of credit information using the decision tree technique showed better performance as compared to Naïve Bayes. But all the experiments which are carried out using decision tree didn't have equal performance. Hence it will be good to select the model which has performed best among them. To this end, the 9th experiment of decision tree from table 4.8is selected. The set of rules are extracted simply by traversing through the output of the decision tree.

The researcher has extracted rules that are believed to be unambiguous, relevant and novel to the domain experiments and shared and discussed the result with the loan officers and credit department domain experts.

The following are the selected set of rules which are in line with the survey of credit risk assessment and primarily got the attention of domain experts:

Rule1:

IF PRODUCT = MANUFACTURING and FREQUENCY = M and USER_DEFINED_STATUS = NORM and RISK_FREE_EXP_AMOUNT = '(-inf-0.05]': GOOD (11.0).

Explanation:

Loans given to the manufacturing industry which do have a monthly repayment with a current Normal status and without any risk free collateral are risk free (GOOD).

Rule 2:

IF PRODUCT = ADVANCE_AGAINST_EXPORT_BILL and TENOR = '(-inf-91.5]': GOOD (131.0)

Explanation:

Loans and Advances rendered for the coverage of export bill having a life time of less than 3 months (91.5 days) are risk free (GOOD).

Rule 3:

IF PRODUCT = TRANSPORT and FREQUENCY = M and TENOR = '(91.5-366.5]' and LOCATION = Addis_Ababa and (BOOKING_YEAR = 2008: BAD (15.0/1.0) or

BOOKING_YEAR = 2009: BAD (6.0/1.0) or BOOKING_YEAR = 2012: BAD (5.0))

Explanation:

Transport term Loans disbursed for Addis Ababa customers having a life time between 3 months and one year at 2008, 2009 and 2012 are risky (BAD).

Rule 4:

IF PRODUCT = DOMESTIC_TRADE_SERVICE and CUSTOMER_TYPE = C: BAD (165.0/1.0)

Explanation:
Nearly all of the DOMESTIC_TRADE_SERVICE loans granted for the corporate customers are risky (BAD)

Rule 5:
IF PRODUCT = DOMESTIC_TRADE_SERVICE and CUSTOMER_TYPE = I: BAD (9427.0/2.0)

Explanation:
Nearly all of the DOMESTIC_TRADE_SERVICE loans granted for the Individual customers are risky (BAD).

Rule 6:
IF PRODUCT = HOTEL_and_TOURISM: BAD (1342.0)
Explanation:
Nearly all of the HOTEL and TOURISM loans are risky for the bank (BAD).

Rule 7:
IF PRODUCT = HEALTH_SERVICE: BAD (143.0)
Explanation:
Nearly all of the HEALTH_SERVICE loans are risky for the bank (BAD).

Rule 8:
IF PRODUCT = ADVANCE_AGAINST_EXPORT_BILL and TENOR = '(91.5-366.5]' and LOCATION = Addis Ababa and LCY_AMOUNT<=2001445.97: BAD (5.0)
Explanation:

The ADVANCE_AGAINST_EXPORT_BILL loans granted in Addis Ababa branches with a life time of 3 months and one year and amount less than 2,001,445.97 are risky for the bank (BAD).

Rule 9:

IF PRODUCT = PRE-SHIPMENT_ADVANCE and TENOR = '(-inf-91.5]' and BOOKING_YEAR = 2006: GOOD (33.0) or BOOKING_YEAR = 2007: GOOD (149.0) or BOOKING_YEAR = 2008: GOOD (97.0) or BOOKING_YEAR = 2009: GOOD (80.0) or BOOKING_YEAR = 2010: GOOD (74.0) or BOOKING_YEAR = 2013: GOOD (17.0)

Explanation:

The PRE-SHIPMENT_ADVANCE loans granted during 2006, 2007, 2008, 2009, 2010 and 2013 calendar year with a life time of less than 3 months (91.5 days)are risk free for the bank (GOOD).

Rule 10:

IF PRODUCT = PRE-SHIPMENT_ADVANCEandTENOR = '(1096.5-inf)': BAD (9.0)

Explanation:

The PRE-SHIPMENT_ADVANCE loans granted with a life time of more than 3 years are risky for the bank (BAD).

Rule 11:

IF PRODUCT = IMPORT and FREQUENCY = M and CUSTOMER_TYPE = I and TENOR = '(1095.5-1096.5]'and INTEREST_RATE = '(11.725-12.03]' and LOCATION = Addis Ababa: GOOD (92.0/3.0)

Explanation:

The IMPORT term loans granted with a life time of around 3 years (1095.5 to 1096.5 days) and with an interest rate between 11.725 and 12.03 for the Addis Ababa branch customers are risk free for the bank (GOOD).

Rule 12:

IF PRODUCT = PROJECT and TENOR = '(-inf-91.5]': BAD (9.0) Or TENOR = '(600.5-729.5]': BAD (6.0)

Explanation:

The PROJECT term loans granted with a life time of less than 3 months (191.5 days) or with a range of tenor of 600.5 to 729.5 days are risky for the bank (BAD).

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1. CONCLUSION

Most of the Ethiopian banks have automated their operations by branch computerization and local and wide area network connectivity. This automation and implementation of core banking solutions have created centralized terabytes of data. There are Valuable bits of information which are embedded in these databases of each bank. The historic data is used commonly for customer statement, auditors' verification and sometimes for functional level report consumption purpose only.

The bulky nature of the banking data is inconvenient to harness interesting information by a human analyst as it was used to be in the old manual days. As a result we need a means of extracting valuable information which is hidden in the terabytes of data. Data mining techniques become important to uncover useful but hidden knowledge through an efficient use of information stored in the databases to support the decision-making process of the business owners and other interested parties.

United Bank being one of the former private banks in Ethiopia, has played its own role in the economy by rendering credit facilities to the individuals and companies which are running business in various sectors. The bank uses internal and NBE credit policies, procedures and strictly followed manuals in various levels of credit committees before disbursing loan to customers. However, there are total defaulters and inconsistent loan repaying customers which declines the profitability of the bank in particular and threatens the growing economy of the country in general. While fueling the sprinting economy in the country, minimizing the possible defaulters is the prime concern of the bank.

The presence of political, economic, social and technological correlations in the financial market forces the creditors to use substantial amount of subjective elements in the identification of risk free customers as it becomes hard to express through deterministic rules.

This research has assessed the application of DM technology on the credit information of United Bank to predict the pattern of risky and risk free contracts by developing a classification model using Weka tool.

This research has been conducted according to the CRISP_DM Model approach. After many pre-processing effort a data set with 27,310 total credit records was used to develop a classification model.J48Decision Tree and Naïve Bayes algorithms were employed to conduct various experiments on the prepared dataset. A Model built by 10-fold cross-validation test mode of unprunedJ48 Decision Tree which registered the highest accuracy (96.6%), was selected as best model for prediction purpose.

The finding of this research has generated various rules of risky and risk free contracts which do have an acceptance by the domain experts. The researcher suggests the use of this model to assist in the non-structured decisions where only the credit committee or managers' intuition and experience are used in the granting process of loans.

5.2. Recommendations

Banks do have the most liquid asset (cash) in their control. This cash comes through various marketing and deposit mobilization techniques and an interest is paid for it. As a result it should not be granted for customers who are not to pay it consistently. So banks need Information on creditworthiness of customers which can be converted to knowledge, which is the most valuable asset in this generation. The researcher believes that

findings of this study will give an insight on the application of data mining techniques to make an informed decision by the bank officials.

Based on the findings discussed above, the following recommendations are forwarded:

- Even though results from this study were encouraging, further classification techniques like neural network should be undertaken by including data before the implementation of the core banking system (2006) to have the full picture of the bank's credit history.
- Currently the bank performs credit scoring activities which includes relationship with the bank, management quality etc. to grade the risk level of the customer and business. However, these grading is filled on papers and not encoded into the core banking system. A data mining research which includes this data source will have a better chance of predicting the future status of any contract before disbursement. So capturing the workflow of loan processing will be a great input for any data mining or decision support system to be carried out for the bank.
- There is a need to develop a credit risk assessment prototype or knowledge base system for the practical implementation of this academic research endeavor.
- In order to tackle the problem of defaults nationwide, an extensive classification experiment is needed from National Bank of Ethiopia central database and or the other private and governmental banks.

Reference

1. Fayyad, U., Piatesky -Shapiro, G., and Smyth, P. (1996). From Data Mining To Knowledge Discovery in Databases, AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–26256097–6. MIT.
2. Hand, D.,Manila., &Smyth,P. (2001). Principles of Data Mining. The MIT Press. ISBN: 026208290x
3. Shmueli, G., Patel, N.R., Bruce, P.C. (2005). Data Mining In Excel: Lecture Notes and Cases, Arlington, USA, Resampling Stats, Inc.
4. Gartner, Inc (2012), Data Mining, Retrieved from : http://www.gartner.com/it-glossary/data-mining/
5. Thearling, K. (2012), An Introduction to Data Mining, Retrieved on 25-Feb-2013, from: http://www.thearling.com/text/dmwhite/dmwhite.htm
6. Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C. & Wirth, R. (2000).CRISPDM 1.0 step-by-step data mining guide. Technical report, CRISP-DM
7. Zaki,M.J., & Wong, L.(2003), Data Mining Techniques,
8. Brachman, R. J. &Anand, T., 1996. The process of knowledge discovery in databases.
9. SAS Institute Inc., (1998), A SAS Institute Best Practices Paper, Data Mining and the Case for Sampling: Solving Business Problems Using SAS® Enterprise Miner™ Software, Cary, NC: SAS Institute Inc.
10. Padhy,N., Dr. Mishra, P.,&Panigrahi, R.(2012). The Survey of Data Mining Applications and Feature Scope. International Journal of Computer Science, Engineering and Information Technology (IJCSEIT). 2(3)
11. Kumar, D. &Bhardwaj, D. (2011). Rise of Data Mining: Current and Future Application Areas. 8(5)
12. Ruben d. (2009). Data mining in healthcare: current applications and issues.CarnegieMellon University. Australia

13. Antonie, M. L., Zaiane, O. R.,Coman, A. (2001), “Application of Data Mining Techniques for Medical Image Classification”, Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD 2001) in conjunction with ACM SIGKDD conference, San Francisco.
14. Kusiak, A., Kernstine, K.H., Kern, J.A. (2000), McLaughlin, K.A., and Tseng, T.L., “Data Mining: Medical and Engineering Case Studies”. Proceedings of the Industrial Engineering Research 2000 Conference,Cleveland, Ohio, pp. 1-7.
15. Ghani, R., Probst, K., Liu, Y., Krema, M., Fano, A. (n.d). “Text Mining for Product Attribute Extraction”, SIGKDD Explorations , 8(1)
16. Deshpande,S. P. &Thakare,V. M. (2010). Data Mining System and Applications: A Review. International Journal of Distributed and Parallel systems (IJDPS). 1(1). DOI : 10.5121/ijdps.2010.1103 32
17. Information management (2013). "Demystifying Market Basket Analysis". Retrieved on 03 March 2013 from: http://www.information-management.com/specialreports/20061031/1067598-1.html
18. Luis, R., Redol, J., Simoes, D., Horta, N. (2003). “Data Warehousing and Data Mining System Applied to ELearning, Proceedings of the II International Conference on Multimedia and Information & Communication Technologies in Education, Badajoz, Spain.
19. Romero, C., Ventura, S. and De-Bra, P. (2004). “Knowledge Discovery with Genetic Programming for Providing Feedback to Courseware Authors, Kluwer Academic Publishers, Printed in the Netherlands.
20. Solieman, O. K., “Data Mining in Sports: A Research Overview, A Technical Report, MIS Masters Project, August 2006”. Retrieved on 05-Mar-2013, from: http://ai.arizona.edu/hchen/chencourse/Osama-DM_in_Sports.pdf

21. Chodavarapu Y., “Using data-mining for effective (optimal) sports squad selections”. Retrieved on 05-Mar-2013, from:http://insightory.com/view/74//using_data-mining_for_effective_(optimal)_sports_squad_selections
22. Berson, A., Smith, S., &Thearling, K. (2000). Building data mining applications for CRM. McGraw-Hill.
23. Ngai,Xiu L., &Chau,D.C.K.(2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications 36, 2592–2602
24. Cai, W. and Li L., “Anomaly Detection using TCP Header Information, STAT753 Class Project Paper, May 2004”.
25. Joshi,S.A., Pimprale,V.S. (2013).Network Intrusion Detection System (NIDS) based on Data Mining. International Journal of Engineering Science and Innovative Technology (IJESIT).2(1).
26. Shmueli, G., Patel, N.R., & Bruce, P.C.(2010) Data mining for business intelligence : concepts, techniques, and applications in Microsoft Office Excel with XLMiner (2nd ed.). John Wiley & Sons, Inc., Hoboken, New Jersey
27. Han, J.& Kamber, M. (2000). Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco, CA.
28. Berry, M. J. A.; Linoff, G. (1997). Data Mining Techniques. For Marketing, Sales and Customer Support. Wiley Computer Publishing
29. Wirth, R., Hipp, J. (n.d). CRISP-DM: Towards a Standard Process Model for Data Mining
30. Azevedo, Lourenço, A.I.R. (2008). KDD, SEMMA and CRISP-DM: a parallel overview
31. Collaborative Research Center (CRC) 649. (2013). Economic Risk.Humboldt-Universitätzu Berlin. School of Business and Economics. Retrieved on 20-Mar-2013 from:http://sfb649.wiwi.hu-berlin.de/fedc_homepage/xplore/ebooks/html/csa/node204.html

32. Ding W. (2007). Data Mining. Retrieved on 20-Mar-2013 from: http://dm-dingwang.blogspot.com/2007/05/supervised-versus-unsupervised-methods.html
33. Cristianini,N. and Scholkopf, B. (2002). Support vector machines and kernel methods: The new generation of learning machines. AI Magazine, pages31-41.
34. Moin,K.I., & Ahmed, Q.B. (2012).Use of Data Mining in Banking. International Journal of Engineering Research and Applications (IJERA). 2(2), pp.738-742. ISSN: 2248-9622.
35. Kargupta,H., Joshi, A., Kumar,K.S., &Yesha,Y. (2005). "Data Mining: Next Generation Challenges and Future Directions"
36. Bhambri, V.(2011). "Application of Data Mining in Banking Sector", International Journal of Computer Science and Technology. 2(2).
37. EUROPEAN COMMITTEE OF CENTRAL BALANCE SHEET DATA OFFICES (2007). Credit Risk Assessment Revisited Methodological Issues and Practical Implications. WORKING GROUP ON RISK ASSESSMENT
38. Basel(1999). Principles for the Management of Credit Risk. Consultative paper issued by the Basel Committee on Banking Supervision.
39. Wills, B. (2012).How Banks Conduct Credit Risk Analysis–and How It Can Affect Your Business. Retrieved on 01-Apr-2013. From: http://creditbuilding.dnb.com/corporate-credit/how-banks-conduct-credit-risk-analysis-and-how-it-can-affect-your-business/
40. Riskglossary(2003). Credit Risk. Retrieved on 02-Apr-2013. From: http://www.riskglossary.com/link/credit_risk.htm

41. UniCredit Group (2012). Credit Risk. Retrieved on 05-Apr-2013. From: https://www.unicreditgroup.eu/en/investors/risk-management/credit.html
42. United Bank S.C. (2009). Credit policy
43. United Bank S.C. (2010). Property estimation Guideline
44. United Bank S.C. (2009). Credit Information system user guide manual.
45. Oracle Financial Services (2009). Transform Your Retail Banking Strategy. Retrieved on 10-Apr-2013. From: http://www.oracle.com/us/industries/financial-services/046059.pdf
46. United Bank SC (2009). United Bank Website. Retrieved on 10-Apr-2013. From: http://www.unitedbank.com.et/
47. UB Annual Report (2012). Annual Report for the year ended June 30, 2012. Retrieved on: 10-Apr-2013. From: http://www.unitedbank.com.et/Annual_Report2012.pdf
48. IBM. (2013).DB2 Business Intelligence. Retrieved on 20-Apr-2013. From: http://publib.boulder.ibm.com/infocenter/db2luw/v9/index.jsp?topic=%2Fcom.ibm.im.easy.doc%2Fc_dm_goals.html
49. J. Han and M. Kamber (2006).Data Mining Concepts and Techniques. Second Edition, Morgan Kaufmann Publishers, San Francisco.
50. Chawla, N.,Bowyer,K., Hall, L.,Kegelmeyer, P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16 (2002), 321–357
51. Rajanish, D. (n.d), Data Mining in Banking and Finance: A Note for Bankers. Indian Institute of Management Ahmedabad
52. Kdnuggets(2013) .What main methodologies are you using for data mining? Retrieved on 01-Mar-2013. From: http://www.kdnuggets.com/polls/2002/methodology.htm

ANNEX I

Sample WEKA System Understandable ARFF Format for UB Credit dataset

```
@relation  Loan_before_Sampling2-weka.filters.supervised.instance.Resample-
B1.0-S1-Z100.0-weka.filters.unsupervised.attribute.Discretize-F-B10-M-1.0-
Rfirst-last

@attribute USER_DEFINED_STATUS {NORM,SPME,SUBS,LOSS,DOUB}
@attribute CONTRACT_STATUS {L,A}
@attribute LOCATION {Addis_Ababa,EAST-WEST-SOUTH,North}
@attribute FREQUENCY {M,B,Q,H,Y}
@attribute CUSTOMER_TYPE {I,C,B}
@attribute RECORD_STAT {O,C}
@attribute BOOKING_YEAR {2006,2007,2008,2009,2010,2011,2012,2013}
@attribute                                                          PRODUCT
{EMERGENCY_STAFF_LOAN,MANUFACTURING,ADVANCE_ON_IMPORT__BILLS,DOMESTIC_TRAD
E_SERVICE,HOTEL_and_TOURISM,PROJECT,HEALTH_SERVICE,PERSONAL,IMPORT,BUILDIN
G_and_CONSTRUCTION,EXPORT,TRANSPORT,PRE-
SHIPMENT_ADVANCE,TRADE_BILLS_DISCOUNT,ADVANCE_AGAINST_EXPORT_BILL}
@attribute      TENOR      {'\'(-inf-91.5]\'','\'(91.5-366.5]\'','\'(366.5-
600.5]\'','\'(600.5-729.5]\'','\'(729.5-730.5]\'','\'(730.5-
731.5]\'','\'(731.5-1094.5]\'','\'(1094.5-1095.5]\'','\'(1095.5-
1096.5]\'','\'(1096.5-inf)\''}
@attribute      INTEREST_RATE      {'\'(-inf-3]\'','\'(3-8.51]\'','\'(8.51-
9.035]\'','\'(9.035-9.51]\'','\'(9.51-10.325]\'','\'(10.325-
10.77]\'','\'(10.77-11.495]\'','\'(11.495-11.725]\'','\'(11.725-
12.03]\'','\'(12.03-inf)\''}
@attribute        RISK_FREE_EXP_AMOUNT        {'\'(-inf-0.05]\'','\'(0.05-
177734.8]\'','\'(177734.8-310479.635]\'','\'(310479.635-
507826.865]\'','\'(507826.865-799684.5]\'','\'(799684.5-
1150376.55]\'','\'(1150376.55-1818024.3]\'','\'(1818024.3-
3030133.76]\'','\'(3030133.76-5933537.965]\'','\'(5933537.965-inf)\''}
@attribute   LCY_AMOUNT   {'\'(-inf-5961]\'','\'(5961-11979]\'','\'(11979-
34161.085]\'','\'(34161.085-121318.765]\'','\'(121318.765-
249994.555]\'','\'(249994.555-383216.74]\'','\'(383216.74-
599302.13]\'','\'(599302.13-1000235.005]\'','\'(1000235.005-
2001445.97]\'','\'(2001445.97-inf)\''}
@attribute STATUS {GOOD,BAD}

@data
NORM,L,Addis_Ababa,B,I,O,2008,ADVANCE_ON_IMPORT__BILLS,'\'(-inf-
91.5]\'','\'(3-8.51]\'','\'(-inf-0.05]\'','\'(599302.13-
1000235.005]\'',GOOD
NORM,L,Addis_Ababa,M,I,O,2009,EMERGENCY_STAFF_LOAN,'\'(730.5-
731.5]\'','\'(-inf-3]\'','\'(-inf-0.05]\'','\'(11979-34161.085]\'',GOOD
NORM,L,Addis_Ababa,M,I,O,2009,EMERGENCY_STAFF_LOAN,'\'(600.5-
729.5]\'','\'(-inf-3]\'','\'(-inf-0.05]\'','\'(5961-11979]\'',GOOD
NORM,L,Addis_Ababa,M,C,O,2010,HOTEL_and_TOURISM,'\'(1096.5-
inf)\'','\'(12.03-inf)\'','\'(5933537.965-inf)\'','\'(2001445.97-
inf)\'',BAD
NORM,L,Addis_Ababa,M,I,O,2008,DOMESTIC_TRADE_SERVICE,'\'(366.5-
600.5]\'','\'(8.51-9.035]\'','\'(0.05-177734.8]\'','\'(34161.085-
121318.765]\'',BAD
```

```
NORM,L,Addis_Ababa,B,I,O,2008,MANUFACTURING,'\'(-inf-91.5]\'','\'(9.035-
9.51]\'','\'(1818024.3-3030133.76]\'','\'(34161.085-121318.765]\'',GOOD
NORM,L,Addis_Ababa,Q,I,O,2009,BUILDING_and_CONSTRUCTION,'\'(366.5-
600.5]\'','\'(10.77-11.495]\'','\'(1818024.3-3030133.76]\'','\'(383216.74-
599302.13]\'',BAD
NORM,L,Addis_Ababa,M,I,O,2010,DOMESTIC_TRADE_SERVICE,'\'(730.5-
731.5]\'','\'(10.325-10.77]\'','\'(507826.865-799684.5]\'','\'(121318.765-
249994.555]\'',BAD
NORM,L,Addis_Ababa,M,I,O,2008,IMPORT,'\'(729.5-730.5]\'','\'(8.51-
9.035]\'','\'(507826.865-799684.5]\'','\'(249994.555-383216.74]\'',GOOD
NORM,L,Addis_Ababa,B,I,O,2011,ADVANCE_ON_IMPORT__BILLS,'\'(-inf-
91.5]\'','\'(11.495-11.725]\'','\'(-inf-0.05]\'','\'(599302.13-
1000235.005]\'',GOOD
NORM,L,EAST-WEST-SOUTH,M,I,O,2008,HOTEL_and_TOURISM,'\'(730.5-
731.5]\'','\'(9.035-9.51]\'','\'(177734.8-310479.635]\'','\'(34161.085-
121318.765]\'',BAD
NORM,L,North,M,I,O,2010,HOTEL_and_TOURISM,'\'(730.5-731.5]\'','\'(10.77-
11.495]\'','\'(177734.8-310479.635]\'','\'(34161.085-121318.765]\'',BAD
NORM,L,Addis_Ababa,M,C,O,2008,IMPORT,'\'(729.5-730.5]\'','\'(9.035-
9.51]\'','\'(3030133.76-5933537.965]\'','\'(599302.13-1000235.005]\'',GOOD
```