ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

APPLICATION OF DATA MINING TECHNIQUES TO PREDICT CUSTOMERS'CHURN AT COMMERCIAL BANK of ETHIOPIA

KASSAHUN GEBREMESKEL

SEPTEMBER 2013

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

APPLICATION OF DATA MINING TECHNIQUES TO PREDICT CUSTOMERS'CHURN AT COMMERCIAL BANK of ETHIOPIA

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

KASSAHUN GEBREMESKEL

September 2013

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

APPLICATION OF DATA MINING TECHNIQUES TO PREDICT CUSTOMERS'CHURN ATCOMMERCIAL BANK of ETHIOPIA

By

KASSAHUN GEBREMESKEL

Name and signature of Members of the Examining Board:

Name	**Title**	**Signature**	**Date**
	Chairperson	_____	_____
Dereje Teferi (PhD)	Adviser	_____	_____
Tibebe Beshah (PhD)	Examiner	_____	_____

DEDICATION

I would like to dedicate this work to all my families especially to my beloved sister Alganesh G/meskel and all others who would like to see the fruits of my effort.

ACKNOWLEDGEMENT

First and foremost, I would like to thank the almighty God to let me finish my work.

I would like to express my deepest gratitude to my advisor, Dr. Dereje Teferi, for his time and excellent guidance.

My sincere thanks also go to Ato Yilma Atlabachew (MIS Directorate Director at CBE), Ato Ephrem Aberra (MIS Directorate technical head), Ato Fikresillasie (CATS-CBC Directorate Director), Achamyelesh Borshe and all the staffs of HRM directorate and CBE corporate communication for providing me the business domain knowledge and data resources to conduct this research.

Finally, I would like to thank my friends Tesfaye Shiferaw and Kidist Wondimu for their support in editorial tasks and their valuable suggestions.

TABLEOF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ATM:	Automatic Teller Machine
CATS:	Customers' Account Transaction Service
CBE:	Commercial Bank of Ethiopia
CRISP-DM:	CRoss-Industry Standard Process for Data Mining
CRM:	Customer Relationship Management
DM:	Data Mining
KDD:	Knowledge Discovery in Databases
KDP:	Knowledge Discovery Process
LR:	Logistic Regression
MB:	Mobile Banking
MER:	Misclassification Error
POS:	Point of Sale Terminal
SMOTE:	Synthetic Minority Oversampling Technique
SMS:	Short Messaging Service
WEKA:	Waikato Environment for Knowledge Analysis

ABSTRACT

Data mining tools and techniques are being used to solve different types of problems in various industries. Predicting customers' churn is one of the areas where data mining can be applied. Customers' churn, which is the common measure of lost customers, is one of the major problems in industries such as banks where there is a fierce competition. By minimizing the number of churning customers companies can maximize their profit and sustainability. For this reason, customer retention is critical for a good marketing and a customer relationship management strategy. This paper presents the prediction of customers, who are prone to move to a competitor, in Commercial Bank of Ethiopia. The data of 13172 customers with 9 attributes and their corresponding 628,634 transactions with 10 attributes is collected from the bank. The CRISP-DM methodology is followed to conduct the data mining process. After the business is thoroughly analyzed and the goals are clearly identified, successive steps of a data preparation processes are undertaken. A dataset of 6045 instances and 18 attributes is prepared. A WEKA (Waikato Environment for Knowledge Analysis) tool is used for modeling. The dataset is partitioned into different sets of testing and training sets. As the proportion of the churn class is very small as compared to the active (non-churn) class, SMOTE (Synthetic Minority Oversampling Technique) has been applied to minimize the class imbalance problem. Three modeling techniques are used for predicting churn. These are J48, Logistic Regression, and Bagging. The training models are built using cross validation and tested for reliability by separate test sets. The models are evaluated by their F-Measure values (which is the harmonic mean of recall and precision). The results of the study show that J48 modeling technique is the best model with a performance of 94.8% followed by bagging (93.9%) and Logistic Regression (76.6%).

CHAPTER ONE

1 Background

1.1 Introduction

Banking industry is one of the most important service industries which touch the livelihood of millions of people. Banking can be traced back to the year 1694 with the establishment of the bank of England. The bank was started by a few individuals who were actually money lenders with an aim of lending money at interest (Ombati, Magutu, Nyamwange, & Nyaoga, 2010). The banking sector consists of the public sector, private sector and foreign banks apart from smaller regional and cooperative banks. (Bhasin, 2006).The industry has brought several changes through time in the service provision mechanisms, in its marketing strategy and also in the amount of capital and availability of the services.

The banking industry in general has experienced some profound changes in recent decades, as innovations in technology and the inexorable forces driving globalization continues to create both opportunities for growth and challenges for banking managers to remain profitable in this increasingly competitive environment (Scott & Arias, 2011). To survive in such competitive environment, much has to be done in customer handling activities.

Many researchers agreed that banks should adopt customer-centric marketing strategies for maximizing their profit and increasing the loyalty of their customer. In today's competitive banking industry, customers can make a choice among various service providers by making a trade-off between relationships and economies, trust and products, or service and efficiency (Sachdev & Verma, 2004).

CRM considered by many researchers, as the key mechanism of handling customers, increase the loyalty of customers and create a better profitability by the banking and other industries. The companies who initiate a good CRM Practice will maximizing the potential of existing customers, acquiring new customers that are profitable or likely to have the potential and retaining customers who are profitable and enhance the firm performance (Boulding et al., 2005). According to (Bhatnagar, 2012), CRM has become inevitable for growth and profitability of Banks in present scenario marked by rising competition, technological advancement and empowered customers.

CRM has been defined by many authors in different ways but the concept behind is more or less similar. According to (Bhatnagar, 2012), CRM is the strategy for building, managing and

strengthening loyal and long-lasting customer relationships. CRM is a customer centric approach based on customer insight. Its ultimate objective is towards 'Personalized' handling of customers as distinct entities through the identification and understanding of their differentiated needs, preferences and behaviors. (Parvatiyar & Sheth, 2002) also noted that CRM is a comprehensive strategy and the process of acquiring, retaining and partnering with selective customers to create superior value for the company and the customer.

According to (Thanuja, Venkateswarlu, & Anjaneyulu, 2011), CRM consists of four dimensions. These are customer identification, customer attraction, customer retention, and customer development. Among the four dimensions, customer retention is the key component of CRM. (Hosseni & Tarokh, 2011) states that customer retention is a vital part of CRM because of the high costs for identifying and attracting the new customers. Customer acquisition and retention are very important concerns for any industry, especially the banking industry (Bhasin, 2006). Predicting churners (customers who quit buying products or services) and taking appropriate action beforehand is one of the major programs in customer retention.

When the number of customers recorded in the banks database gets higher and higher, banks face difficulty in clearly understanding the need and wants of their customers and to effectively apply CRM processes. (Bhambri, 2012) stated that the huge size data bases makes it impossible for the organizations to analyze these data bases and to retrieve useful information as per the need of the decision makers. The manual process of data analysis becomes tedious as size of data grows and the number of dimensions increases, so the process of data analysis needs to be computerized (Beniwal & Arora, 2012).

DM technologies will be helpful in analyzing the data and provide helpful (supporting) results for a better customer handling procedures through CRM. DM is a tool used to extract important information from existing data and enable better decision-making throughout the banking and retail industries (Moin & Ahmed, 2012). According to (Bhasin, 2006), DM allows extracting knowledge from the historical data, and predicting outcomes of future situations. It helps optimize business decisions, increase the value of each customer and communication, and improve customer satisfaction.

The Banking industry is growing relatively at a faster rate in Ethiopia than ever before. The type of services being provided is improved and also being supported by modern technologies, the number of banks has increased significantly and the competition among each other becomes fierce. According to (National Bank of Ethiopia, 2012), currently there are more than 18 government-owned and private

banks operating in the country. CBE, Construction and Business Bank and Development Bank of Ethiopia are government owned banks, while the rest are private. Almost all the banks have many branches in different regional states of the country. They implement core banking system and other advanced technologies for a better management of their customers' data and for delivering their services in a better way.

CBE, which is the leading bank in Ethiopia, was established in 1942. It is a government owned bank having about 673 branches (as of the date of this research) stretched in most of cities and towns of the country. Having more than 12,800 employees, the bank is playing a catalytic role in the economic progress & development of the country. CBE is one of the leading African banks with assets of Birr 155 billion as on June 30th 2012. The bank has more than 4 million account holders currently.

According to (Commercial Bank of Ethiopia, 2012), the bank is mainly engaged in three categories of banking activities. These are Domestic Banking, International Banking, and E-payment. The Domestic banking services of the bank are:

1. Deposits,
2. Credit facilities and
3. Local transfer.

The deposit service is being provided in three ways. These are: Saving accounts, Current (checking) account, and Diaspora account.

There are different types of Credit facilities being provided by CBE. These are: Overdraft, Merchandise loan facility, Pre-shipment Export Credit facility, Revolving Export Credit Facility, Special Truck Loan Financing, Short term loan, Medium and long term loans, Agricultural Input Loan, Agricultural Investment Loan, Coffee farming Term Loan Financing, and Micro–Finance Institution's Loan. The bank is also engaged in Local money Transfer service in different ways. Some of the most common ways of money transfer are: Mail Transfers, Telegraphic or Telephone Transfers, Local Drafts, and Cashier Payment Order (CPO).

The major International Banking services provided at CBE are Trade Services, Money Transfer (cross country), and Foreign Exchange (Forex Service). The Trade service is available in all branches of the bank in different ways. These are: Documentary credit (L/C), Documentary collection, Advance payment, Consignment basis payment, Guarantee, Franco-valuta license (permit), and Small export items license (permit). In the E-Payment service of CBE, Internet Banking, Card Banking (ATM, POS, and CBE Reliable VISA card), and MB services are being provided.

CBE passes through different banking technological advancements. Currently, the bank is in a process to migrate its customers' data from the earlier "SMART system" to an advanced "Core Banking system" and also working to integrate the overall banking activities of all the branches with network. Core banking system is used by the bank to manage all the routine financial transactions in an integrated manner. As most of the branches of CBE serve thousands of customers every day, the amount of data recorded is too much. But apart from supporting the day-to-day transaction, this huge amount of accumulated data is not being utilized effectively for supporting the various CRM activities.

The intention of this research is therefore to apply different DM tools and techniques in order to convert the accumulated data of CBE into useful knowledge. Such knowledge, which is created by building different DM models, will be helpful in predicting the customers' behavior and supports the CRM processes of the bank.

1.2 Statements of the problem

It is stated by many researchers that the banking industry is operating in an environment where there is fierce competition. CRM process should be applied to overcome most of the challenges in relation to customer handling and increasing the loyalty of customers. Today, the invention of several DM tools and techniques and the availability of ample customer related data in the banking industry created a good opportunity to support the CRM Process. But only limited banks are using DM technology to enjoy its benefit in segmenting customers based on homogeneity in behavior, predicting potential customers and credit defaults; and predicting churning customers from their behavior.

In one hand, CBE has millions of customers and the number of customers is still increasing as the bank has different promotional strategies to attract new customers. But, on the other hand, a number of customers are churning out for unknown reasons. If this trend continues, it will results in affecting the progress of CBE negatively though the number of new subscribers seems to be increasing. Various researchers forwarded their opinions in favor of this argument. Churn is extremely damaging for companies(Lemmens & Croux, 2006). Existing customer's churning will likely to result in the loss of businesses and thus decline in profit(Sharma & Panigrahi, 2011). Losing customers not only leads to opportunity lost because of reduced sales, but also to an increased need for attracting new customers, which is five to six times more expensive than customer retention(Bhambri, 2012). And (Soeini & Rodpysh, 2012) also noted that churn of good customers have irrecoverable disadvantages for a famous company.

In most of the banks, there is no trend of collecting data as to why customers churn. In addition, when the number of customers gets larger and larger, it becomes difficult to understand the needs and wants of customers and to take appropriate actions accordingly. The huge size of databases makes it impossible for the organizations to analyze their databases and to retrieve useful information as per the need of the decision makers using standard database techniques (Bhambri, 2012).By applying DM tools and techniques, it is possible to extract useful knowledge from accumulated data about customers. Hence, the research questions in this study are the following:

- What transactions does a customer do before shifting to a competitor?
- What is the general behaviour of those customers who are susceptible to churn?

1.3 Objectives of the Study

1.3.1 General Objective

The general objective of this research is to apply various DM tools and techniques on an existing customers related data of CBE in order to obtain the best model which can predict the behavior of churning customers.

1.3.2 Specific Objectives

The specific objectives of the study are:

- Find out the best and appropriate sources of knowledge and data for the study.
- Analyse the business domain to determine the business objectives and DM goal
- Collect data from the sources.
- Perform data preparation tasks (such as: data selection, data cleaning, data construction and data integration) on the data for the aptness of model building process.
- Select the best and appropriate DM tools (software) and techniques for building the required models.
- Apply the selected DM tools and techniques to build models.
- Evaluate whether the determined results (models) are worth mentioning
- Provide conclusion and recommendation

1.4 Scope and Limitations of the Study

This research work focuses on applying DM techniques in supporting the CRM of CBE or specifically, into predicting the susceptible churner customers of CBE by extracting knowledge from

the available customers' data at CBE. So, the domain knowledge of the business and sample of customer related data is required. For the purpose of obtaining the required inputs, various departments in CBE headquarter such as: IT, Marketing, and other divisions working in relation to CRM are communicated as part of the study. The specific tasks to be addressed during the study are: collecting data, prepare the data, construct models of different types, and finally evaluate the models and selecting the best one in accordance with the actual facts and standards set by CBE and the DM goals. This is an academic exercise and as such the limited amount of time and resource hindered the researcher from working on a wider scope. Moreover, deploying the models is beyond the scope of the study as it needs the willingness and readiness of the organization.

1.5 Research Methodology

There are different types of standards and methodologies being used in DM researches. CRISP-DM is the most widely used and highly recommended model to be used for DM researches and projects in several industries including the banking industry. CRISP-DM evolved to become the de facto industry standard (Chapman et al., 2000; Jackson, 2002; Kulikowski, 2011; KURGAN & MUSILEK, 2006; MARISCAL, MARBAN, & FERNANDEZ, 2010). It is also stated as a neutral process model that can be used with any tool or application by any industry (Gilchrist, Mooers, Skrubbeltrang, & (Corresponding, 2012). It is vendor-independent so it can be used with any DM tool and it can be applied to solve any DM problem (Ponce & Karahoca, 2009). For these reasons, the researcher adopted the CRISP-DM model to achieve the intended outcomes of the research. Consequently, the following summarized five steps or phases of the CRISP-DM process model, which are discussed in detail and applied in CHAPTER FOUR and FIVE, are employed:

1.5.1 Business Understanding

At this first stage of the methodology the business domain is thoroughly analyzed. All the services being provided in the bank and all the related environments and situations are also analyzed from the CRM perspective. The ultimate purpose of this stage is to acquire sufficient knowledge about the business domain so that business objectives and the business success criteria for those objectives could be determined as clearly as possible. Some of the methods to be used for acquiring the intended knowledge are:

- Having thorough observations in the relevant business processes, which are expected to be affected by the study.

- Reviewing supporting documents, which are published by CBE at different times (such as Annual reports and bulletins)
- Browsing the official website of CBE.
- Having discussions with various relevant authorities of the bank (the questions prepared for the purpose of guiding the intended discussions are annexed in Appendix III)

Once the business domain is clearly understood, the business objectives and success criteria is stated and the DM goal and success criteria are also determined based on the business objectives and success criteria. At the end of this phase an initial project plan is produced that shows how the problems can be solved and also how the stated business objectives can be achieved.

1.5.2 Data Understanding

During the second phase of the research methodology, initial data is collected from the possible sources of data at CBE. In the bank there is a computerized system that can accommodate all the customers' accounts data and history of all the transactions by every customer ever since the beginning. In addition, the system is a centralized system, which can be accessed from anywhere within the bank's network. It is from this database system that all the necessary data is collected. The data is described to make clear the meanings, data types, list of possible values (if any), number of instances and missing values of each attribute. Here after, the importance or significance of each attribute to achieve the desired outcome of the research is discussed. The data is also explored to see the correlation between each attribute and to get familiar with all attributes and their corresponding values. Finally at this stage, the data quality also is verified so that any erroneous entries or missing values or measurement errors are checked to exist in the data and to propose a way for handling such data quality problems.

1.5.3 Data preparation

At this stage, the data which is needed for the experimentation is collected from the source, described and explored for better understanding. But the collected data itself cannot directly be used for modeling. This is because:

- Data might contain errors, outliers, missing values and other quality problems which need to be resolved beforehand,
- The importance of every attribute and domains of data might not be equal to achieve the intended purpose. So, data selection might become necessary.

- There might be a necessity of deriving attributes and even tables from the existing data.
- A need of integrating data from two or more tables would become mandatory, and
- The original data may not be in a format and structure, which are acceptable by the modelling tools and techniques.

For the aforementioned reasons, several steps of data processing (Data cleaning, construction, integration, and data formatting) are done in accordance with the need for making the data appropriate for the intended experimentation followed by attribute selection. For the purpose of data preparation, tools such as MS-EXCEL, MS-ACCESS are used. Ultimately after the data preparation process, a complete dataset is prepared on which the experimentation process is going to be conducted.

1.5.4 Modelling

The appropriate modeling tools and techniques for the intended objectives are selected at this stage. While selecting tools and techniques, the reasons as to why the tools and techniques are selected should be justified. Once after the tools and techniques are selected in a logical manner, a test design is generated. The test design includes:

- The way how the dataset is going to be used for building the models (as they are, or any further process should be applied)
- The way each model is tested to show the reliability of the model (training and testing set preparation, etc.) And also the way the performance of each model is evaluated so that selecting the best model could be logical.
- The steps to be followed in conducting the experimentation process for each model. Here:
    - ✓ How many repetitions of experimentation can be attempted for each modelling technique before going to another technique?
    - ✓ What specific task is going to be done in all experimentation for each specific technique?

  are discussed in detail

- After generating the test design, models are built as per the project plan prepared in the first stage of the methodology. The model is assessed to be in line with the intended purpose.

1.5.5 Evaluation of the models and the result

The models, which are built in the previous phase, need be tested to be realistic and applicable. This model evaluation process is done with the bank's representative authorities (if possible). Once the

models are found to be realistic, feasible and interesting they can be used to classify the existing data as per the model (depending on the interest and readiness of CBE).

1.6 Significance of the Study

It has been stated in Section 1.2 that customers' churn has irrecoverable damage on the overall organization's success. And conversely, retaining customers through churn prediction will contribute to the profitability and the overall success of an organization. Assuming that CBE is willing to deploy the best model of this study for predicting the churners and retain them, the organization will be benefitted in:

- Understanding the needs and wants of churning customers so that remedial actions could be taken by the management,
- Retaining those predicted customers so that an effort for acquiring new customers (which is stated to be 5 to 6 times tougher than customer retention) will be minimized,
- maximizing profit and overall success of the bank

As the bank is one of the government organizations, which are playing a catalytic role in the economic progress & development of the country, the success of CBE can be considered as the success of the government (the country).

In addition, while the susceptible churners are communicated earlier and their reason for churning is assessed and improved by the bank, customers would become happy to continue with the bank. As a result, the customers are also benefitted as they will minimize the burden of looking for other banks and subscribe as a new customer (where they don't know how well they are going to be treated)

Moreover, the output of this research can be a source of knowledge in the area of CRM in banking and used by future researchers for further studies.

1.7 Organization of the Thesis

This research paper is organized in six chapters. The first chapter covers the background of the study which presents brief introduction about the study, statement of the problem, general and specific objectives, the research methodology, the scope and limitations of the study, and significance of the study. The second chapter covers reviewed literature in relation to DM, CRM and also churn prediction in detail. In chapter three the DM techniques, which are widely used for predicting customers' churn, are discussed. Chapter four presents business understanding, data understanding,

and the process of data preparation. Chapter five presents the experimentation phase of the study. The last chapter, chapter six presents what is concluded from the study and recommendations.

CHAPTER TWO

2 CRM and DM Applications

2.1 CRM

2.1.1 CRM Overview

In the earlier days, businesses gave much focus towards their products and services. The attention given to customers was very less. About the earlier business strategy (Stanley, 2012) mentioned that customers were passive and a producer could sell his products and services in his own terms and took very little effort towards customer commitment. But, many businesses of today are shifting their attention towards customers, knowing that it is a critical success factor. The possible reasons and factors for business to become customer-centric or Relationship marketing are explained by various authors. The constant change and evolution of technology (Maroofi, Aliabadi, Fakhri, & Hadikolivand, 2013) and (Mishra & Mishra, 2009), the increase in competition, deregulation, and the internet (Fagbemi & Olowokudejo, 2011), the growing de-intermediation process in many industries due to the advent of sophisticated computer and telecommunication technologies that allow producers to directly interact with end-customers(Parvatiyar & Sheth, 2002), economic liberalization, increasing competition, high consumer choice, enlightened and demanding customer, more emphasis on quality and value for purchase (Stanley, 2012) are considered as the main factors for businesses to adopt customer-centered strategy and to the rapid development and evolution of CRM.

Many authors believe that the concept of CRM is adopted from the well-known marketing strategy called Relationship Marketing. According to (Wang & Kang, 2008), Relationship Marketing is defined as those activities that provide individualized product and services accordingly to target customers with the use of information technologies and database for the purpose of establishing relationship with customers.

There has been a misunderstanding in the meaning of CRM. (Maroofi et al., 2013) said that CRM can be understood from four different perspectives: as a business philosophy, a business strategy, a business process, or a technological tool and gave the definitions of CRM by different authors from all the four perspectives.

In his book, (Buttle, 2009) mentioned the different ways that CRM could be misunderstood, and these are:

- CRM is misunderstood as it is database marketing: - Database marketing is concerned with building and exploiting high quality customer databases for marketing purposes. Though analytical CRM has the appearance of database marketing, the scope of CRM is much wider than database marketing.
- CRM is misunderstood as it is a marketing process: - "CRM software applications are used for many marketing activities: market segmentation, customer acquisition, customer retention and customer development (cross-selling and up-selling), for example. However, operational CRM extends into selling and service functions."
- CRM is misunderstood as it is an IT issue: - Though it is true that most CRM implementations require the deployment of IT solutions, the importance of people and processes in the implementation of CRM should not be underestimated.
- CRM is misunderstood as it is about loyalty scheme: - In fact, loyalty schemes may play two roles in CRM implementation:
    - ✓ They generate data that can be used to guide customer acquisition, retention and development, and
    - ✓ May serve as an exit barrier.

  In this regard we can only say that some (but not all) CRM implementations are linked to loyalty schemes.
- CRM is misunderstood as it can be implemented by any company: - Any company can implement Strategic CRM and also possibly operational CRM. But for the implementation of Analytical Combat least data are needed to identify which customers are likely to generate most value in the future, and to identify within the customer base segments that have different requirements. So companies having no such data cannot implement analytical CRM.

Taking all the aforementioned misunderstandings into consideration, (Buttle, 2009) defined CRM as follows:

> CRM is the core business strategy that integrates internal processes and functions, and external networks, to create and deliver value to targeted customers at a profit. It is grounded on high quality customer- related data and enabled by information technology.

In most cases, CRM has been defined from marketing and IT point of view and it is the background of authors which determines the way how CRM is defined (Rai & Singh, 2012). Authors from marketing background emphasize technological side of CRM while the others consider IT perspective of CRM.

(Gupta & Aggarwal, 2012) defined CRM relatively in a more general way, as it refers to the methodologies and tools that help businesses manage customer relationships in an organized way. In other words, CRM simply means managing all customer interactions which requires using information about your customers and prospects to more effectively interact with your customers in all stages of your relationship with them.

2.1.2 Importance and Major Benefits of CRM for Businesses

As mentioned earlier, there are several factors which contribute to the growth and development of CRM and for businesses to become customer-centric. However, apart from that, many authors agree that there are several benefits business can gain by implementing CRM.

Various authors mentioned the importance of CRM systems to businesses. Some of these are:

- For selecting important and valuable customers for targeting (Parvatiyar & Sheth, 2002),
- As a value-added activity through mutual interdependence and collaboration between suppliers and customers (Parvatiyar & Sheth, 2002),
- As its implementation is positively associated with customer satisfaction (Adalikwu, 2012),
- It's important in increasing the size of customer rate and consequently improve customer retention rate and also it improves business performance by enhancing customer satisfaction and driving up customer loyalty (Buttle, 2009),
- It provides the infrastructure that facilitates long-term relationship building with customers and also reduce duplication in data entry and maintenance by providing a centralized firm-database of customer information (Mishra & Mishra, 2009).

Some authors categorize the benefits of implementing CRM into the different types of CRM and into the type of industry it is implemented. Taking the banking industry into consideration, (Bhatnagar, 2012) describe the benefits of CRM as operational, analytical and collaborative. Sales force automation, Customer service support and Enterprise marketing automation are some of the major benefits of operational CRM to banks. The major benefits of Analytical CRM to banks are Customer retention, Fraud detection, Optimizing marketing efforts as per customer life value, Credit risk analysis, Segmentation and targeting, and Development of customized new products matching the specific preferences and priorities of customers. Collaborative CRM has also benefits to banks and some of these benefits are: Providing efficient customer communication across a variety of channels, online service to reduce customer service costs, and providing access to customer data while interacting with customers.

Again from Banking business point of view (V. TAMILVENDAN & S. SWAMIDOSS, 2012)opines that CRM will make the work and activities systematic and coordinate. As a result, there will be an effective and efficient utilization of funds, reduction of operational costs and effective and efficient business operation. In addition, they mentioned the benefits of CRM as:

- It performs the function of maintaining a link with the customers, which then laid down the foundation for survival and the success of the banks.
- It enables the banks to identify potential customers for approaching them with suitable offers
- It will lead the banks to give better customer service
- It will be better stand against global competition

Taking the banking industry into consideration, the benefits of CRM are described by (SHARMA, JULKA, & BHARDWAJ, 2012) as it:

- Will result in rapid growth of business through retention of the profitable customer segment,
- Will be useful to acquire only those customers whose characteristics are known which results in increased profits and hence drives growth,
- Enables the banks (organizations) to offer the right product at right time, due to which individual customer margins can be increased,
- It decreases the cost of customer management, and
- Ease in introduction of new products as they are customer need specific and it is found that the loyal customers readily buy 50 % of the new products rather than the new customers.

2.1.3 Types and Classifications of CRM

2.1.3.1 Types of CRM

Different researchers and authors put their views about the types of CRM. In his book, (Buttle, 2009) stated that there are four types of CRM: Strategic, Operational, Analytical, and Collaborative CRM. (Bhatnagar, 2012) mentioned three types of CRM are adopted by Banks: Operational, Analytical and Collaborative. (Mishra & Mishra, 2009) focuses on two types of CRM modules (Operational and analytic CRM modules) provide the major functions of a CRM system. But to be complete, it is necessary to describe all the four types of CRM here.

1. *Strategic CRM*: - According to (Buttle, 2009), Strategic CRM is a core customer-centric business strategy that aims at winning and keeping profitable customers. For this reason, many managers would argue that customer-centric must be right for all companies. In other words,

strategic CRM can be implemented by any company; however, at different stages of market or economic development, other orientations may have stronger appeal.

2. *Operational CRM*: - Operational CRM focuses on the automation of customer-facing processes such as selling, marketing and customer service (Bhatnagar, 2012; Buttle, 2009; Mishra & Mishra, 2009). For instance banks record contact history and store valuable customer information to ensure a consistent picture of customer's relationship with the bank that can be retrieved by staff as per requirement. Such activities can be accomplished using operational CRM.
3. *Analytical CRM*: - (Bhatnagar, 2012) said that Analytic CRM is about analysing customer information to better address marketing and customer service objectives and deliver the right message to the right customer at the right time through the right channel. For analytical CRM to be implemented in businesses, having customer related data to be analysed is a mandatory issue. So, it cannot be implemented in any kind of organization.
4. *Collaborative CRM*: - involve systems facilitating customers to perform services on their own through a variety of communication and interactive channels (Bhatnagar, 2012). For instance, in the case of banks it brings people process and data together and enables channelling of data and information appropriately to bank staff for proactive decision making and enhanced informed customer service and support activities.

2.1.3.2 Classifications (Dimensions) of CRM

According to (Thanuja et al., 2011)CRM consists of four dimensions which can be seen as a closed cycle of a customer management system. They share the common goal of creating a deeper understanding of customers to maximize customer value to the organization in the long term. The four dimensions of CRM are:

1. *Customer Identification*: - This phase involves targeting the population who are most likely to become customers or most profitable to the company. Moreover, it involves analysing customers who are being lost to the competition and how they can be won back. Elements for customer identification include target customer analysis and customer segmentation. Target customer analysis involves seeking the profitable segments of customers through analysis of customers' underlying characteristics, whereas customer segmentation involves the subdivision of an entire customer base into smaller customer groups or segments, consisting of customers who are relatively similar within each specific segment.

2. *Customer Attraction: -* After identifying the segments of potential customers, organizations can direct effort and resources into attracting the target customer segments. An element of customer attraction is direct marketing. Direct marketing is a promotion process which motivates customers to place orders through various channels. For instance, direct mail or coupon distributions are typical examples of direct marketing.
3. *Customer Retention*: - This is the central concern for CRM. Customer satisfaction, which refers to the comparison of customers' expectations with his or her perception of being satisfied, is the essential condition for retaining customers. Loyalty programs involve campaigns or supporting activities which aim at maintaining a long term relationship with customers. Specifically, **churn analysis**, credit scoring, service quality or satisfaction form part of loyalty programs.
4. *Customer Development*: -This involves consistent expansion of transaction intensity, transaction value and individual customer profitability. Elements of customer development include customer lifetime value analysis, p/cross selling and market basket analysis. Customer lifetime value analysis is defined as the prediction of the total net income a company can expect from a customer. Market basket analysis aims at maximizing the customer transaction intensity and value by revealing regularities in the purchase behaviour of customers.

Different authors agree that, among the four dimensions of CRM, Customer retention has a significant role. (Hosseni & Tarokh, 2011) opine that, Customer retention is a vital part of CRM because of the high costs for identifying and attracting the new customers. It is becoming more evident that the only way to remain a leader in this industry is to not only be customer-driven but also focus on building long-term relationships (Nie, Rowe, Zhang, Tian, & Shi, 2011). Describing Customer Retention as a key component of CRM, (Karimii, Maymand, Hosseini, & Ahmadinejad, 2012) stated that many competitive businesses correctly have found that for survival in an industry it is necessary that retention of valuable customers be in the center of their management strategies.

Churn prediction and churn management are the important strategies in customer retention programs. The prevention of customer churn through customer retention is a core issue of CRM (CRM) (Prasad & Madhavi, 2012). According to (Soeini & Rodpysh, 2012), old companies are trying to keep their customers because, in today competitive market, lack of customers must be recovered by new one which has the following problems:

1. Attracting new customer is difficult and expensive,

2. High expenses of process which lead to service revoke,
3. Losing customer lead to income reduction and negative effects on

2.1.4 Customer Churn

Losing customers or customer churn is one of the major problems that companies of today face. Churning is defined by various authors mostly according to the nature of the organization to which it applies.

Churn is defined as the propensity of a customer to stop doing business with an organization and subsequently moving to some other company in a given time period (Bhambri, 2012). The customers who stop using the company's products are usually called churners (Nie et al., 2011). Churn is also called attrition and often used to indicate a customer leaving the service of one company in favor of another company (Sharma & Panigrahi, 2011). In banking domain, (Chitra & Subashini, 2011) define a churn customer as one who closes all his/her accounts and stops doing business with the bank.

2.1.4.1 Types and major reasons of Customers' Churn

According to (HADDEN, 2008), churning customers can be divided into two main groups. These are:

1. Voluntary churners and
2. Non-voluntary churners.

Non-voluntary churn is the type of churn in which the service is purposely withdrawn by the company. There are several reasons why a company could revoke a customer's service. Reasons such as abuse of service and non-payment of service are usually the main causes.

Voluntary churn is more difficult to determine. This type of churn occurs when a customer makes a conscious decision to terminate his/her service with the provider. Voluntary churn can be divided into two sub categories, incidental churn and deliberate churn. Incidental churn happens when changes in circumstances prevent the customer from further requiring the provided service. Examples of incidental churn include changes in the customer's financial circumstances so that the customer can no longer afford the service, or a move to a different geographical location where the company's service is unavailable.

Deliberate churn is the problem that most churn management solutions attempt to identify. This type of churn occurs when a customer decides to move his/her custom to a competing company due to reasons of dissatisfaction. Reasons that could lead to a customer deliberately churning include: -

1. Technology-based reasons, when a customer discovers that a competitor is offering newer technology that their existing supplier does not provide.
2. Economic reasons can also be a cause of deliberate churn such as finding the same product from a competitor at a better price.
3. Examples of other reasons for deliberate churn include quality factors like poor coverage, bad experiences with services, and consistent faults with service,

Focusing on the banking industry, (Chitra & Subashini, 2011) forwarded their opinion as to why a customer closes his/her account(s). Among the major reasons they stated for customer churn, some of them are:

- A person creates an account for a specific purpose may close it immediately after the purpose is solved.
- A person is relocated and has to move to another place and hence closes all the accounts.
- A customer may stop transacting with the bank just because of the unavailability of bank's ATMs in important places and hence close his/her accounts.

The problem here is that, in real world scenario, the bank does not always capture this kind of feedback data. Hence, no further analysis can be done and this type of churning behaviors could not be stopped.

2.1.4.2 Consequences of Customers' Churn in Businesses

The problem of customers' churn for a business has been stated by various authors and researchers. (Chitra & Subashini, 2011) opines that Churn is a major problem today in the banking sector. This is because losing the customers can be very expensive as it costs to acquire a new customer. Churn of good customers have irrecoverable disadvantages for a famous company (Soeini & Rodpysh, 2012). Losing customers not only leads to opportunity lost because of reduced sales, but also to an increased need for attracting new customers, which is five to six times more expensive than customer retention (Bhambri, 2012).While customers' churn having its own problems for businesses, reducing the number of churners has got several advantages. Identifying the churn beforehand and taking necessary steps to retain the customers would increase the overall profitability of the organization (Bhambri, 2012). Finding the churners can help companies retain their customers (Nie et al., 2011).

2.1.4.3 Managing customers' Churn

To alleviate the business problems in relation to churning, it is very important to predict the future churners and take a remedial action beforehand. For finding answers to the questions who and why is likely to churn a classification of the customers is needed. According to (Chitra & Subashini, 2011), Churn prediction deals with the identification of customers likely to churn in the near future. The basis for this is historical data, containing information about past churners. A comparison is made between these churners and existing customers. As likely churners are identified customers for which the classification suggests similarity to prior churners.

According to (Bhambri, 2012) organizations must know the following points in order to avoid and prevent churning out of the customers. These are: -

1. What is the profile, tastes, preferences and purchasing behaviour of the customer?
2. What is the transaction behaviour of various customers?
3. Which products are often purchased together by the customers of which particular profile?
4. What services and benefits would current customers likely desire?
5. Identifying the customers who are getting all types of services from your company?

(Radosavljevik, Putten, & Larsen, 2010) opine that the first step in managing churn is identifying the customers with high propensity to churn. To tackle this problem one needs to understand the behavior of customers, and classify the churn and non-churn customers, so that the necessary decisions will be taken before the churn customers switch to a competitor(Rashid, 2008).

DM can be used in customer retention applications by employing churn modeling. In a typical application, DM identifies customers who are profitable and who are likely to leave or churn. In the context of banking, churn modeling is an important DM application. Predicting churn likelihood is important to the banking industry for reducing the number of new customers who defect soon after being acquired (Ogwueleka, 2009).

2.2 DM and its Applications

2.2.1 DM Overview

It is a custom that, organizations and businesses record data of their day to day activities. Nowadays corporate and organizations are accumulating data at an enormous rate and from a very broad variety of sources such as customer transactions, credit card transactions, bank cash withdrawal to hourly weather data. A lot of relational database servers have been built to store such massive quantities of

data (ZenTut, 2013a). And (Sumathi & Sivanandam, 2006) believe that today organizations are accumulating vast and growing amounts of data in different formats and databases. According to these authors, data are any facts, numbers, or text that can be processed by a computer; and can have various formats and types. These include:

- Operational or transactional data such as sales, cost, inventory, payroll, and accounting.
- Non-operational data like industry sales, forecast data, and macroeconomic data.

But, as many authors agreed that, the data is simply being accumulated in storage device for no use. According to (Kumar & Bhardwaj, 2011), the fact lies in that data is growing at a very rapid rate, but most of data has once been stored and have never been used. This data collected from different sources if processed properly, can provide immense hidden knowledge, which can be used further for development. (ZenTut, 2013a) mentioned that, the data itself is critical to a company's growth. It contains knowledge that could lead to important business decisions that bring business to the next level. These data is never been examined in a superficial manner. It is becoming data rich but knowledge poor.

Thus, a need arises for extracting knowledge from the accumulated data. There is an eminent need for developing proper mechanisms of processing these large volumes of data and extracting useful knowledge from large repositories for better decision making (Kumar & Bhardwaj, 2011). This is considered as one of the main reasons for the DM technology to come into existence. (Han & Kamber, 2006) stated DM as it can be viewed as a result of the natural evolution of information technology. The database system industry has witnessed an evolutionary path in the development of the following functionalities:

- data collection and database creation,
- data management (including data storage and retrieval, and database transaction processing), and
- advanced data analysis (involving data warehousing and DM)

Several authors have forwarded their opinion for the reasons why DM has attracted a great deal of attention in the information industry and in society as a whole in recent years. It is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and

science exploration (Han & Kamber, 2006). (Berry & Linoff, 2004) believe that it is the convergence of the following factors:

- The data is being produced,
- The data is being warehoused,
- Computing power is affordable,
- Interest in CRM is strong, and
- Commercial DM software products are readily available.

According to (Sumathi & Sivanandam, 2006), data analysis techniques that have been traditionally used for such tasks include regression analysis, cluster analysis, numerical taxonomy, multidimensional analysis, other multivariate statistical methods, stochastic models, time series analysis, nonlinear estimation techniques, and others. These techniques have been widely used for solving many practical problems. They are, however, primarily oriented toward the extraction of quantitative and statistical data characteristics, and as such have inherent limitations. (Kumar & Bhardwaj, 2011) mentioned that the roots of DM can be traced back along three lines. These are Statistics, Artificial Intelligence & Machine Learning, and Databases. DM is also described as a convergence of three technologies by (Sahu, Shrma, & Gondhalakar, 2008). These three technologies are:

1. *Increasing Computing Power*: - As stated in Moore's law, computing power doubles every 18 months. Powerful workstations became common, and Cost effective servers (SMPs) provide parallel processing to the market.
2. *Improved Data Collection and Management*
3. *Improved Statistical and Learning Algorithms*: - Techniques have often been waiting for computing technology to catch up.

Figure 2.1 : DM as a convergence of three technologies(Sahu et al., 2008)

The term DM is defined by various authors in a different way. According to (Jackson, 2002) while the term DM is often used rather loosely, it is generally a term that's used for a specific set of activities, all of which involve extracting meaningful new information from data. However, the term DM is not new to statisticians. It is a term synonymous with data dredging or data snooping and has been used to describe the process of trawling through data in the hope of identifying patterns. In their study of Using DM techniques to increase the efficiency of CRM process, (Nejad, Nejad, & Karami, 2012) stated that DM is a technique that can help companies to move towards customer-oriented marketing.

In his study entitled as "A DM Approach for Retailing Bank Customer Attrition Analysis", (Hu, 2005) defined the term DM as follows:

> "DM is an iterative process that combines business knowledge, machine learning methods and tools and large amounts of accurate and relevant information to enable the discovery of non-intuitive insights hidden in the organization's corporate data."

According to the definition given by (Setty, T.M.Rangaswamy, & K.N.Subramanya, 2010), DM is an analytic process designed to explore data (usually large amounts of data, typically business or market related) and in search of consistent patterns and /or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data.

Though it is mentioned by various authors about the importance of DM in extracting useful knowledge, it has its own advantages, disadvantages and also there are major challenges that it is facing. (Sahu et al., 2008; ZenTut, 2013b) see the advantages of DM from Marketing/Retail, Finance/Banking, Manufacturing, and Government angles. The advantages are as stated below:

1. *Marketing/Retail*: - DM helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign…etc. Through the results, marketers will have appropriate approach to sell profitable products to targeted customers. The same is true for retail companies but in addition DM also helps the retail companies offer certain discounts for particular products that will attract more customers.
2. *Banking/Finance*: - DM gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank and financial institution can determine good and bad loans. In addition, DM helps banks detect fraudulent credit card transactions to protect credit card's owner.
3. *Manufacturing*: - By applying DM in operational engineering data, manufacturers can detect faulty equipment and determine optimal control parameters. For example semi-conductor manufacturers has a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are lot the same and some for unknown reasons even has defects. DM has been applying to determine the ranges of control parameters that lead to the production of golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.
4. *Governments*: - DM helps government agency by digging and analysing records of financial transaction to build patterns that can detect money laundering or criminal activities.

(Sahu et al., 2008; ZenTut, 2013b) believe that some of the disadvantages of DM are Privacy issue, Security issue and Misuse of Information/Inaccurate information.

As one of the new and promising field of computing technology; DM is being applied in various problem areas and becoming increasingly successful. Some authors put their views about the future challenges and opportunities of DM. According to (Venkatadri & Reddy, 2011), ever increasing technology and future application areas are always poses new challenges and opportunities for DM, the typical future trends of DM includes: Standardization of DM languages, Data pre-processing, Complex objects of data, Computing resources, Web mining, Scientific Computing, and Business data. (Sahu et al., 2008) also mentioned some of the challenges of DM are: Scalability, Dimensionality, Complex and Heterogeneous Data, Data Quality, Data Ownership and Distribution, Privacy Preservation, and Streaming Data.

2.2.2 Classification of DM

(Deshpande & Thakare, 2010) classified DM systems according to various criteria. The major classifications of DM systems are:

- *According to the type of data source mined*: This classification is according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.
- *According to the data model*: This classification based on the data model involved such as relational database, object-oriented database, data warehouse, transactional database, etc.
- *According to the kind of knowledge discovered*: This classification based on the kind of knowledge discovered or DM functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several DM functionalities together.
- *According to the mining techniques used*: This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc.

The classification can also take into account the degree of user interaction involved in the DM process such as query-driven systems, interactive exploratory systems, or autonomous systems. A comprehensive system would provide a wide variety of DM techniques to fit different situations and options, and offer different degrees of user interaction.

2.2.3 Evolution of DM and KD Process Models and Methodologies

The terms DM and KDD are mistakenly been used interchangeably. There is confusion about the terms DM, Knowledge Discovery and KDD. Many researchers and practitioners use DM as a synonym for knowledge discovery. KDP also called KDD seeks new knowledge in some application domain whereas DM is just one step of the KDP (Cios, Pedrycz, Swiniarski, & Kurgan, 2007). DM is one of the core tasks of the KDD process (Deshpande & Thakare, 2010; Jackson, 2002; Maimon & Rokach, 2010; Sumathi & Sivanandam, 2006).

KDD is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Cios et al., 2007; Fayyad, Piatetsky-shapiro, & Smyth, 1996). The process generalizes to non-database sources of data, although it emphasizes databases as a primary source of data (Cios et al., 2007).

In the early 1990s, when the KDD process term was first coined, there was a rush to develop DM algorithms that were capable of solving all problems of searching for knowledge in data (Ponce & Karahoca, 2009). The KDD process proposed by has a process model component because it establishes all the steps to be taken to develop a DM project, but it is not a methodology because its definition does not set out how to do each of the proposed tasks. It is also a life cycle. The 5 A's is a process model that proposes the tasks that should be performed to develop a DM project and was one of CRISP-DM's forerunners. Therefore, they share the same philosophy: 5 A's proposes the tasks but does not suggest how they should be performed. Its life cycle is similar to the one proposed in CRISP-DM. The Human-Centered Approach was proposed to be a people-focused DM proposal. This proposal describes the processes to be enacted to carry out a DM project, considering people's involvement in each process and taking into account that the target user is the data engineer. SEMMA is the methodology that SAS proposed for developing DM products. Although it is a methodology, it is based on the technical part of the project only. Like the above approaches, SEMMA also sets out a waterfall life cycle, as the project is developed right through to the end. The evolution of the DM and KD process models and methodologies is shown in Figure 2.2

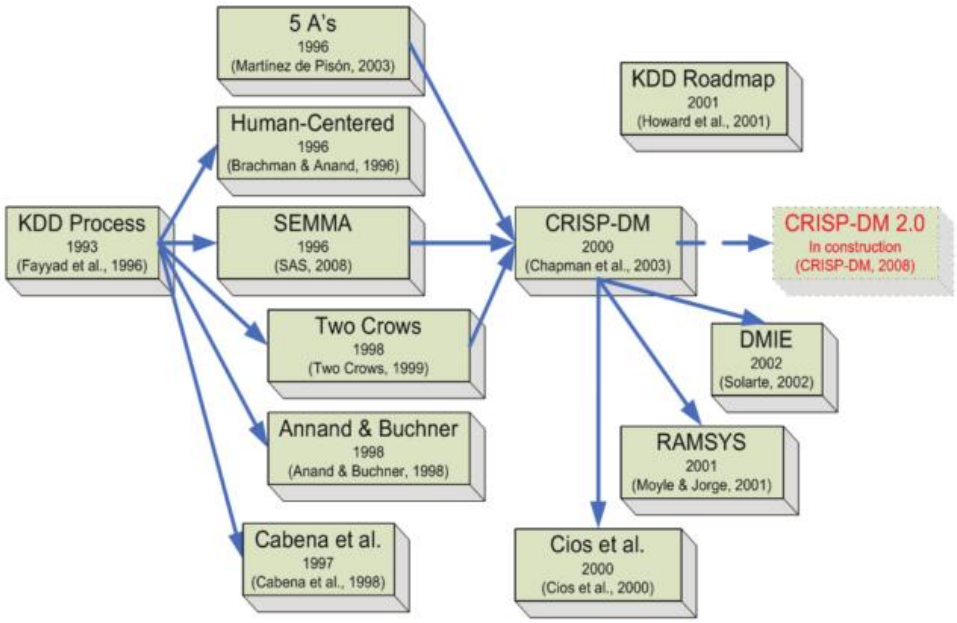


Figure 2.2: Evolution of DM and KD process models and methodologies (MARISCAL et al., 2010)

2.2.4 KDD and DM related approaches

As shown in Figure approaches such as: 5A's, Human-Centered, SEMMA, Two Crows, Annand & Buchner, and Cabenna et al. are KDD related approaches; whereas CRISP-DM 2.0, DMIE, RAMSYS, and Cios et al. are CRISP-DM related approaches. The number of steps and unique features of some of the process models in the two approaches is described as follows:

2.2.4.1 The KDD Process

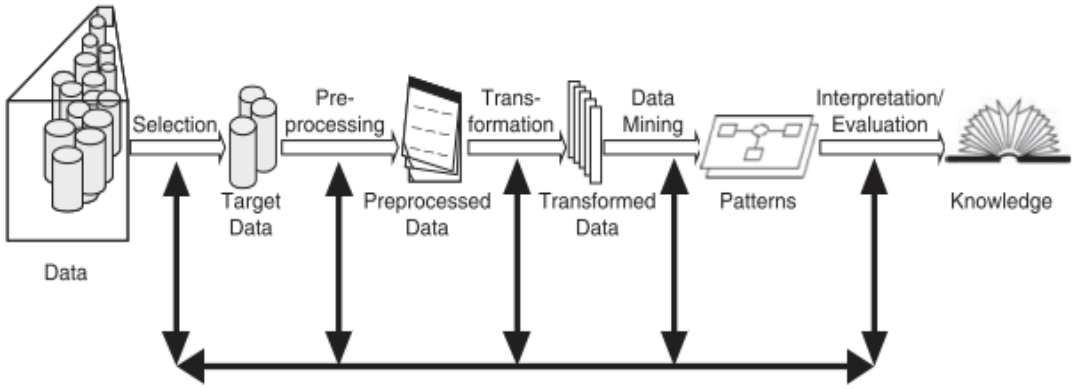


Figure 2.3 : Steps Constituting the KDD process(Fayyad et al., 1996)

According to (MARISCAL et al., 2010), the KDD process is interactive and iterative (with many decisions made by the user), involving nine steps, described from the practical viewpoint as:

1. *Learning the application domain*: It includes developing an understanding of the relevant prior knowledge and the goals of the application.
2. *Creating a target data set*: It includes selecting a data set or focusing on a subset of variables or data samples on which discovery is to be performed.
3. *Data cleaning and pre-processing*: It includes basic operations, such as removing noise or outliers if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes, as well as deciding data base management system issues, such as data types, schema and mapping of missing and unknown values.
4. *Data reduction and projection*: It includes finding useful features to represent the data, depending on the goal of the task, and using dimensionality reduction or transformation

methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

5. *Choosing the function of DM*: It includes deciding the purpose of the model derived by the DM algorithm (e.g., summarization, classification, regression and clustering).
6. *Choosing the DM algorithm*: It includes selecting method(s) to be used for searching for patterns in the data, such as deciding which models and parameters may be appropriate and matching a particular DM method with the overall criteria of the KDD process.
7. *DM*: It includes searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, clustering, sequence modelling, dependency, association rules and line analysis.
8. *Interpretation*: It includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns and translating the useful ones into terms understandable by users.
9. *Using discovered knowledge*: It includes incorporating this knowledge into the performance system, taking actions based on the knowledge or simply documenting it and reporting it to interested parties, as well as checking for and resolving potential conflicts with previously believed (or extracted) knowledge.

This model is proposed by (Fayyad et al., 1996).

2.2.4.2 SAS - THE SEMMA ANALYSIS CYCLE

SAS developed a DM analysis cycle known by the acronym SEMMA. This acronym stands for the five steps of the analyses that are generally a part of a DM project (S=Sample, E=Explore, M=Modify, M=Model, A=Assess)(Jackson, 2002).

Sample: the first step in is to create one or more data tables by sampling data from the data warehouse. Mining a representative sample instead of the entire volume drastically reduces the processing time required to obtain business information.

Explore: after sampling the data, the next step is to explore the data visually or numerically for trends or groupings. Exploration helps to refine the discovery process. Techniques such as factor analysis, correlation analysis and clustering are often used in the discovery process.

Modify: modifying the data refers to creating, selecting, and transforming one or more variables to focus the model selection process in a particular direction, or to modify the data for clarity or consistence.

Model: creating a data model involves using the DM software to search automatically for a combination of data that predicts the desired outcome reliably.

Assess: the last step is to assess the model to determine how well it performs. A common means of assessing a model is to set aside a portion of the data during the sampling stage. If the model is valid, it should work for both the reserved sample and for the sample that was used to develop the model.

2.2.4.3 SPSS - THE 5 A'S PROCESS

SPSS originally developed a DM analysis cycle called the 5 A's Process5. The five steps in the process are: ***Assess Access, Analyse, Act,*** and ***Automate*** (Jackson, 2002).

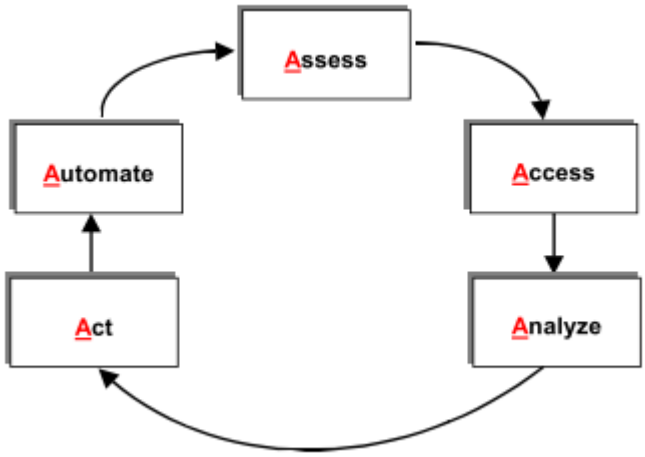


Figure 2.4 The 5A's process model(Jackson, 2002)

2.2.4.4 Human-centered approach of DM

The human-centered model emphasized the interactive involvement of a data analyst (data miner) during the process. Its basic steps are: task discovery, data discovery, data cleaning, model development, data analysis and output generation. These six steps cover the same tasks that are included in Fayyad et al. (1996c) KDD process. The main difference between both approaches is that human-centered process is focused in the tasks from the data miner viewpoint, while KDD process is more focused in data transformations. Human-centered model shows in a clearer way which decisions the user has to make (MARISCAL et al., 2010).

2.2.4.5 Cabena et al

Cabena et al. define DM (referring to the complete KDD process) as the process of extracting previously unknown, valid, and actionable information from large databases and then using the information to make crucial business decisions. The steps of DM process according to Cabena et al are: business objectives determination, data preparation (that includes data selection, data pre-processing and data transformation), DM, analysis of results, and assimilation of knowledge. There are not big differences between the DM tasks proposed by the original KDD process and Cabena et al. approach, although they structure the process in a different number of steps (MARISCAL et al., 2010).

2.2.4.6 Two Crows

The Two Crows DM process model is proposed by Two Crows Corporation in 1999. This model is derived from the previous edition of the Two Crows process model of the 1998, and also takes advantage of some insights from first versions of CRISP-DM (before CRISP-DM 1.0 is released). While the steps appear in a list, the DM process is not linear you will inevitably need to loop back to previous steps. The basic steps of Two Crows are: define business problem, build DM data base, explore data, prepare data for modeling, build model, evaluate model, and deploy model and results. Two Crows approach is very close to the original KDD process, although they use different names for similar steps (MARISCAL et al., 2010).

2.2.4.7 Anand and Buchner

Anand and Buchner have proposed a model covering the entire life cycle of an online customer, the available operational and materialized data, as well as the incorporation of marketing knowledge. The web-enabled KDP, also known as internet-enabled KDP, is an adoption of a generic process defined in earlier work adapted to web mining projects in this case. The model consists of eight steps. These are: human resource identification, problem specification, data prospecting, methodology identification, data pre-processing, pattern discovery, and knowledge post-processing. While it is true that Anand and Buchner provide a detailed analysis of initial steps of the process, unfortunately, it does not include the needed activities to use the discovered knowledge (MARISCAL et al., 2010).

2.2.4.8 CRISP-DM

In response to common issues and needs in DM project in the mid 90's, a group of organizations involved in DM (Teradata, SPSS -ISL-, Daimler-Chrysler and OHRA) proposed a reference guide to develop DM projects, named CRISP-DM (CRoss Industry Standard Process for DM) (Chapman et al.,

2000). CRISP-DM is considered the de facto standard for developing DM and knowledge discovery projects (Chapman et al., 2000; Jackson, 2002; MARISCAL et al., 2010; Ponce & Karahoca, 2009). One important factor of CRISP-DM success is the fact that CRISP-DM is industry-, tool- and application- neutral.

The CRISP-DM DM methodology is described in terms of a hierarchical process model, consisting of sets of tasks described at four levels of abstraction (from general to specific). At the top level, the DM process is organized into a number of phases; each phase consists of several second-level generic tasks. This second level is called generic, because it is intended to be general enough to cover all possible DM situations. The third level, the specialized task level, is the place to describe how actions in the generic tasks should be carried out in certain specific situations. The fourth level, the process instance, is a record of the actions, decisions and results of an actual DM engagement.

Horizontally, the CRISP-DM methodology distinguishes between the reference model and the user guide. The reference model presents a quick overview of phases, tasks, and their outputs and describes what to do in a DM project. The user guide gives more detailed tips and hints for each phase and each task within a phase and depicts how to do a DM project.

(MARISCAL et al., 2010) CRISP-DM distinguishes between four different dimensions of DM contexts. These are:

1. The application domain is the specific area in which the DM project takes place.
2. The DM problem type describes the specific classes of objectives that the DM project deals with.
3. The technical aspect covers specific issues in DM that describe different (technical) challenges that usually occur during DM.
4. The tool and technique dimension specifies which DM tool(s) and/or techniques are applied during the DM project.

The life cycle of a DM project consists of six phases, as shown in Figure 2.5. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. The outcome of each phase determines which phase, or particular task of a phase, has to be performed next. The arrows indicate the most important and frequent dependencies between phases.

The outer circle in Figure 2.5 symbolizes the cyclical nature of DM itself. DM does not end 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 DM processes will benefit from the experiences of previous ones.

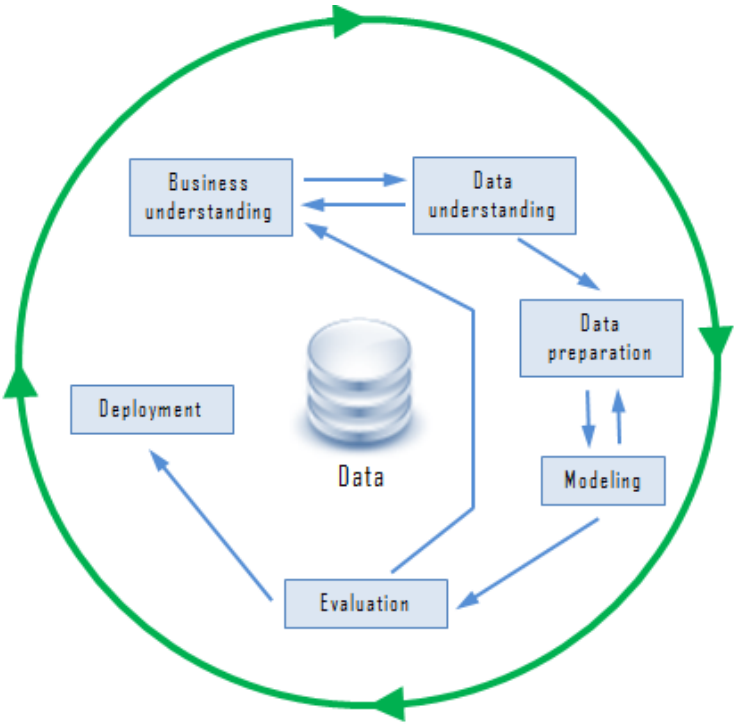


Figure 2.5: - The CRISP-DM process Model(Chapman et al., 2000)

The six phases of CRISP-DM are briefly described as follows:

1. ***Business Understanding***: - This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.
2. ***Data understanding***: - The data understanding phase starts with initial data collection and proceeds with activities that enable you to become familiar with the data, identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information.
3. ***Data preparation***: - The data preparation phase covers all activities needed to construct the final dataset [data that will be fed into the modelling 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, as well as transformation and cleaning of data for modelling tools.

4. ***Modelling***: - In this phase, various modelling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same DM problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary.
5. ***Evaluation***: - At this stage in the project, you have built a model (or models) that appear to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to thoroughly evaluate it and review the steps executed to create it, to be certain the model 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 DM results should be reached.
6. ***Deployment***: - Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. It often involves applying "live" models within an organization's decision making processes—for example, real-time personalization of Web pages or repeated scoring of marketing databases. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable DM process across the enterprise. In many cases, it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will carry out the deployment effort, it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models.

According to (MARISCAL et al., 2010) many changes have occurred in the business application of DM since CRISP-DM1.0 was published (CRISP-DM, 2007). Emerging issues and requirements include:

- The availability of new types of data (e.g., text, web and attitudinal data) along with new techniques for pre-processing, analysing and combining them with related case data.
- Integration and deployment of results with operational systems such as call centres and web sites.
- Far more demanding requirements for scalability and for deployment into real-time environments.
- The need to package analytical tasks for non-analytical end users and integrate these tasks in business workflows.

- The need to seamlessly integrate the deployment of results and closed-loop feedback with existing business processes.
- The need to mine large-scale databases in situ, rather than exporting an analytical data set.
- Organizations' increasing reliance on teams, making it important to educate greater numbers of people on the processes and best practices associated with DM and predictive analytics.

The tasks in each phase of the CRISP-DM process model is depicted in Table 2.1

Table 2.1: CRISP-DM phases and tasks

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives	Collect initial Data	Collect data	Select modeling technique	Evaluate Results	Plan deployment
Assess situation	Describe Data	Clean data	Generate test design	Review Process	Plan monitoring & maintenance
Determine DM Objectives	Explore Data	Construct data	Build model	Determine next steps	Produce final report
Produce Project Plan	Verify Data Quality	Integrate data	Assess model		Review report
		Format data			

2.2.4.9 Cios et al.

The process model of Cios et al. was first proposed in 2000 by adapting the CRISP-DM model to the needs of academic research community (MARISCAL et al., 2010). The main extensions of the latter model include a more general, research-oriented description of the steps, introduction of several explicit feedback mechanisms and a modification of the description of the last step, which emphasizes that knowledge discovered for a particular domain may be applied in other domains. Cios et al. process model is based on technologies like XML, PMML, SOAP, UDDI and OLE BD-DM.

The model consists of six steps: understanding the problem domain, understanding the data, preparation of the data, DM, evaluation of the discovered knowledge, and using the discovered knowledge. The process is iterative and interactive. Since any changes and decisions made in one of the steps can result in changes in later steps, the feedback loops are necessary.

2.2.5 DM Tasks

Authors categorize the DM tasks differently. (Han & Kamber, 2006; Singh & Chauhan, 2009) categorize DM tasks broadly into two classes. These are ***Descriptive*** and ***Predictive DM***.

Descriptive DM provides information to understand what is happening inside the data without a predetermined idea. ***Predictive DM*** allows the user to submit records with unknown field values, and the system will guess the unknown values based on previous patterns discovered form the database.

Depending on the use of DM result (Deshpande & Thakare, 2010; Padhy, Mishra, & Panigrahi, 2012) classified DM tasks into five categories:

Exploratory Data Analysis: In the repositories vast amount of information's are available .This DM task will serve the two purposes

1. Without the knowledge for what the customer is searching, then
2. It analyse the data

These techniques are interactive and visual to the customer.

Descriptive Modelling: - It describe all the data, it includes models for overall probability distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables.

Predictive Modelling: This model permits the value of one variable to be predicted from the known values of other variables.

Discovering Patterns and Rules: This task is primarily used to find the hidden pattern as well as to discover the pattern in the cluster. In a cluster a number of patterns of different size and clusters are available .The aim of this task is "how best we will detect the patterns" .This can be accomplished by using rule induction and many more techniques in the DM algorithm like (K-Means /K-Medoids). These are called the clustering algorithm.

Retrieval by Content: The primary objective of this task is to find the data sets of frequently used in audio/video as well as images. It is finding pattern similar to the pattern of interest in the data set.

2.2.6 DM Techniques & Algorithms

DM consists of many up-to-date techniques such as classification (decision trees, naive Bayes classifier, k-nearest neighbor, and neural networks), clustering (k-means, hierarchical clustering, and

density-based clustering), association (one-dimensional, multi- dimensional, multilevel association, constraint-based association) (Sumathi & Sivanandam, 2006). Adding in this regard, (Jackson, 2002) stated that DM uses the classical statistical procedures (such as LR, discriminant analysis, and cluster analysis) and machine learning techniques (such as neural networks, decision trees, and genetic algorithms).In the continuum of data analysis techniques, the disciplines of statistics and of machine learning often overlap.

The major DM techniques are discussed by many authors. The four major DM techniques are classification, clustering, association rule and information visualization (Clewley et al., 2009). The various DM techniques are association, clustering, sequence or path analysis and forecasting (Chopra, Bhambri, & Krishan, 2011). In a generalized way (ZenTut, 2013c) listed the major DM techniques and they are briefly examined as follows:

1. ***Association***: - Association is one of the best known DM techniques. In association, a pattern is discovered based on a relationship between items in the same transaction. That is the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together. Retailers are using association technique to research customer's buying habits. Based on historical sale data, retailers might find out that customers always buy crisps when they buy beers, and therefore they can put beers and crisps next to each other to save time for customer and increase sales.

   Types of association rule

   - ✓ *Multilevel association rule*
   - ✓ *Multidimensional association rule*
   - ✓ *Quantitative association rule*

2. ***Classification***: -Classification is a classic DM technique based on machine learning. 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. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all records of employees who left the company; predict who will probably leave the company in a future period." In this case, we divide the records of

employees into two groups that named “leave” and “stay”. And then we can ask our DM software to classify the employees into separate groups.

Types of classification models:

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

3. ***Clustering***: -Clustering is a DM technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. To make the concept clearer, we can take book management in library as an example. In a library, there is a wide range of books in various topics available. The challenge is how to keep those books in a way that readers can take several books in a particular topic without hassle. By using clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for entire library.

   Types of clustering methods

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

4. ***Prediction***: -The prediction, as it name implied, is one of a DM techniques that discovers relationship between independent variables and relationship between dependent and independent variables. For instance, the prediction analysis technique can be used in sale to

predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

Types of regression methods

- ✓ *Linear Regression*
- ✓ *Multivariate Linear Regression*
- ✓ *Nonlinear Regression*
- ✓ *Multivariate Nonlinear Regression*

5. ***Sequential Patterns***: -Sequential patterns analysis is one of DM technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period. In sales, with historical transaction data, businesses can identify a set of items that customers buy together at different times in a year. Then businesses can use this information to recommend customers buy it with better deals based on their purchasing frequency in the past.
6. ***Decision trees***: -Decision tree is one of the most used DM techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it.

2.2.7 Applications of DM

In their book 'Handbook of statistical analysis and DM applications' (Nisbet, Elder, & Miner, 2009) stated that DM technology can be applied anywhere a decision is made, based on some body of evidence. Accordingly, the diversity of applications in the past includes the following:

- *Sales Forecasting*: - One of the earliest applications of DM technology
- *Shelf Management*: - A logical follow-on to sales forecasting
- *Scientific Discovery*: -A way to identify which among the half-billion stellar objects are worthy of attention
- *Gaming*: - A method of predicting which customers have the highest potential for spending
- *Sports*: - A method of discovering which players/game situations have the highest potential for high scoring
- *CRM*: - Retention, cross-sell/up-sell propensity

- *Customer Acquisition*: - A way to identify the prospects most likely to respond to a membership offer

According to (Sumathi & Sivanandam, 2006), DM techniques have been applied successfully in many areas such as: database marketing, retail data analysis, stock selection, credit approval, etc. and in many fields such as: astronomy, molecular biology, medicine, geology, and many more. In addition, it has also been used in health care management, tax fraud detection, money laundering monitoring, and even sports. The details of some of the applications of DM are the following:

- *Market management*: - Target marketing, CRM, market basket analysis, cross-selling, market segmentation.
- *Risk management*: - Forecasting, customer retention, improved underwriting, quality control, competitive analysis.
- *Fraud management*: - Fraud detection.
- *Industrial-specific applications*: -
    - ✓ *Banking, finance, and securities*: - Profitability analysis (for individual officer branch, product, product group, monitoring marketing programs and channels, customer data analysis customer segmentation profiling).
    - ✓ *Telecommunications and media*: - Response scoring, marketing campaign management, profitability analysis, and customer segmentation.
    - ✓ *Health care*: - FAMS (Fraud and Abuse Management System) assisting health insurance organizations dealing with fraud and abuse (detection, investigation, settlement, prevention of recurrence)

In their journal article 'The Contribution of DM in Information Science', (Chen & Liu, 2004) mentioned that personalized environments, electronic commerce, and search engines are the main application domains in the field of information science that a DM can support. (P. Sundari & K. Thangadurai, 2010) believe that there can be a large variety of DM scenarios. But for the purpose of their study they divided the applications of DM into six major categories. These are: Science and Engineering, Business, Banking, Telecommunication, Spatial DM, and Surveillance.

2.2.7.1 DM in the Banking Industry

The application of DM techniques in the banking industry has been given emphasis by various authors. (Bhambri, 2011) stated the contribution of DM in solving business problems by finding

patterns, associations and correlations which are hidden in the business information stored in the data bases. The following points show what customer data the banking industry needs to explore and the reasons for exploring such data.

- What is the profile, taste and preferences, attitude of the customer and what is the purchasing behaviour of the customer since the time he/she is with the bank? (Used to Cross sell the products).
- What transactions does a customer do before shifting to a competitor bank? (To prevent shifting of customers)
- Which products are often purchased together by the customers of which particular profile? (For target marketing)
- What patterns in credit transactions lead to fraud? (To detect and deter fraud)
- What is the profile of a high-risk borrower? (To prevent defaults, bad loans, and improve screening)
- What services and benefits would current customers likely desire? (To increase loyalty and customer retention)
- Identifying the customers who are getting all types of services from the bank? (Identifying 'Loyal' Customers)

The contribution of DM in solving business problems by finding patterns, associations and correlations from a huge amount of data stored in banks' databases is also explained by (Moin & Ahmed, 2012). Accordingly, some of the benefits of DM in banks are the following:

- By analysing patterns and trends, bank executives can predict with increased accuracy as to:
  - ✓ how customers will react to adjustments in interest rates,
  - ✓ which customers will be likely to accept new product offers,
  - ✓ which customers will be at a higher risk for defaulting on a loan, and
  - ✓ how to make customer relationships more profitable
  - ✓ which are most profitable credit card customers or high-risk loan applicants

(Moin & Ahmed, 2012) also give the following examples on the areas where the banking industry has been effectively utilizing DM.

1. *Marketing*: - One of the most widely used areas of DM for the banking industry is marketing. The bank's marketing department can use DM to analyse customer databases. DM carries various analyses on collected data to determine the consumer behaviour with reference to

product, price and distribution channel. The reaction of the customers for the existing and new products can also be known based on which banks will try to promote the product, improve quality of products and service and gain competitive advantage. Bank analysts can also analyse the past trends, determine the present demand and forecast the customer behaviour of various products and services in order to grab more business opportunities and anticipate behaviour patterns.

2. *Risk Management*: - DM is widely used for risk management in the banking industry. Bank executives need to know whether the customers they are dealing with are reliable or not. Banks provide loan to their customers by verifying the various details relating to the loan such as: amount of loan, lending rate, repayment period, and type of property mortgaged, demography, income, and the credit history of the borrower. Customers with bank for longer periods, with high income groups are likely to get loans very easily. Bank executives by using DM technique can also analyse the behaviour and reliability of the customers while selling credit cards too.
3. *Fraud Detection*: - Another popular area where DM can be used in the banking industry is in fraud detection. Being able to detect fraudulent actions is an increasing concern for many businesses; and with the help of DM more fraudulent actions are being detected and reported. Two different approaches have been developed by financial institutions to detect fraud patterns. In the first approach, a bank taps the data warehouse of a third party and use DM programs to identify fraud patterns. The bank can then cross-reference those patterns with its own database for signs of internal trouble. In the second approach, fraud pattern identification is based strictly on the bank's own internal information.

2.2.7.2 DM to support CRM

According to (Moin & Ahmed, 2012), DM can be useful in all the three phases of a customer relationship cycle: Customer Acquisition, Increasing value of the customer and Customer retention. Customer acquisition and retention are very important concerns for any industry, especially the banking industry. Today customers have wide range of products and services provided by different banks. Hence, banks have to cater the needs of the customer by providing such products and services which they prefer. This will result in customer loyalty and customer retention.DM techniques help to analyze the customers who are loyal from those who shift to other banks for better services. If the customer is shifting from his bank to another, reasons for such shifting and the last transaction performed before shifting can be known which will help the banks to perform better and retain its customers.

The use of DM in supporting the CRM of various industries is also described by (Thanuja et al., 2011). They listed several industries including banking, finance, retail, insurance, and telecommunications that can be assisted by DM for better understanding their business, to be able to serve their customers and to increase their effectiveness in the long run. The following industries are some of the examples to show how they can use DM mainly to support CRM processes:

1. *The banking industry*: - DM tools for CRM are widely used by leading banks for customer segmentation and profitability, credit scoring and approval, predicting payment default, marketing, detecting fraudulent transactions, cash management and forecasting operations, optimizing stock portfolios, and ranking investment.
2. *The Retail industry*: - The retail industry is also realizing that it is possible to gain competitive advantage utilizing DM. Retailers have been collecting enormous amount of data throughout the years. For retailers, DM can be used to provide information on product sales trends, customer buying habits and preferences, supplier lead times and delivery performance, seasonal variations, customer peak traffic periods, and similar predictive data for making proactive decisions. The most widely used areas of DM for retail industry is marketing, risk management, fraud detection, customer acquisition and retention.
3. *The Telecom industry*: -There are several reasons for applying DM techniques for CRM in telecommunications:
4. *Competitive market*: - After years of being a monopoly market, the telecommunications market is now highly competitive. A monopoly does not change much, but competitive markets change constantly. Customers are able to switch providers easily, because there are many of them available. For this reason telecommunications companies explore DM solutions to achieve competitive advantage. By understanding the demographic characteristics and customers' behaviour, telecommunications companies can successfully tailor their marketing strategies to reach those most likely to use their services, to increase customer loyalty and improve customer profitability.
5. *High churn rates*: - Churn refers to the monthly or the annual turnover of the customer base. Competitive climate naturally results in high churn rates. Initially, growth in the telecommunications market was exponential, and since many new customers arrived, the churn was not a problem.

Many organizations are using DM to help manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers (Two Crows Corporation, 2005). By determining characteristics of good customers (profiling), a company can target prospects with similar characteristics. By profiling customers who have bought a particular product it can focus attention on similar customers who have not bought that product (cross-selling). By profiling customers who have left, a company can act to retain customers who are at risk for leaving (reducing churn or attrition), because it is usually far less expensive to retain a customer than acquire a new one.

(Chopra et al., 2011) mentioned that DM techniques can significantly improve the customer conversion rate by more focused marketing. Following are the applications of various techniques of DM to CRM:

- Expanding the Customer Base by acquiring new and profitable customers. To expand the customer base, DM can answer questions like:
  - ✓ Which new market the organization can intrude into?
  - ✓ Which kind of customers would you like to acquire?
  - ✓ Which kind of customers will drive your growth in future?
  - ✓ Which new customers are likely to be interested in your products?
- Customer acquisition is the number one issue for every company. As corporation increase budgets to attract and obtain new customers, DM becomes a critical tool for profiling good customers, performing market segmentation, and improving the results of direct-marketing campaigns. In almost all cases the organization needs to cover large number of areas in a given time. DM tools can solve these problems to a greater extent. These tools can provide the organization the relevant data which can be used by the organization to carry on its marketing campaign to acquire new territories and new customers. Instead of mass pitching a certain "hot" customer service representatives product, the can be equipped with customer profiles enriched by DM that help them to identify which products and services are most relevant to callers, so instead of calling a large set of people, the focus will be on a particular set of people, which will reduce the cost of calling to large extent and probability of maturing the calls will be more. The DM techniques which can be applied in this phase are Clustering and associations.
- Lengthening the relationship with your top customers. To lengthen the customer relationships, DM can answer these questions:

- ✓ Which customers in particular do you want to keep?
- ✓ Which customers will drive most of your profits?
- ✓ Which customers might switch to your competitors and why?
- ✓ Which customers are dissatisfied with your services and products?

- By reducing the customer attrition, the organization will increase its profits. Finding the new customers and exploring new territories need huge investment, so it is better for the company to satisfy the existing customers and they should try to strengthen the relationships with them. Customer turnover is a difficult problem to manage because it usually occurs without warning, because once the customer has decided to switch over to the competitor then the organization has left with almost no option to retain the customer. DM introduces a major paradigm shift to manage the shifting of customers by adding predictive capabilities. Data-mining tools can be used to model the patterns of past churning customers by examining billing histories, demographic information, and other customer data. Then, the same model can be used to predict other good customers who are likely to leave in the near future. Armed with this information, the marketer can proactively instigate campaigns to keep their customer, rather than fighting to get them back later.
- Increasing Customer Delight through customized services: Intensifying and deepening customer relationships is also the need of the hour for the organizations. The concept of customer satisfaction has been shifted to customer delight, which can be enhanced by providing more customized services. The company needs to explore:
  - ✓ Which customers are likely to give you more business?
  - ✓ Which products and services interest a particular customer?
  - ✓ Which products are typically bought together and by which set of customers?
  - ✓ What cross selling opportunities should you consider?
- Reaching to the heart of the customers by providing more customized services can again be a success mantra for the organizations to survive. Unlike increasing market share, which focuses on obtaining a greater number of customers, increasing customer share refers to the notion when the customer avails more services from the same organization. Two common methods for this are customized product-launch campaigns, and cross-selling. The following tools of DM can be used: Clustering, Sequence analysis and forecasting

2.2.8 Related Works

Various works have been done by researchers and authors in relation to DM, though only few researches have been conducted in relation to CRM by local researchers. In a global context a lot has been done in the area of churn prediction for various industries. So, the related works are presented in the following two sections in two categories. The first section focuses on global works related to churn prediction and the second section is about the local DM researches.

2.2.8.1 Related Works on Churn Prediction (Global Context)

In their paper, (Nie et al., 2011), applied two DM algorithms to build a churn prediction model using credit card data collected from a real Chinese bank. They select the certain variables from the perspective of not only correlation but also economic sense. In addition to the accuracy of analytic results, they design a misclassification cost measurement by taking the two types error and the economic sense into account, which they believe to be more suitable to evaluate the credit card churn prediction model. The algorithms used in their study include LR and decision tree which are proven mature and powerful classification algorithms. The test result shows that regression performs a little better than decision tree.

In their conference proceeding paper "Customer Retention in Banking Sector using Predictive DM Technique", (Chitra & Subashini, 2011) have made a solution for the churn problem in banking sector using DM technique. They use four sets of data variables: customer behavior, customer perceptions, customer demographics and macro environment variables. They have used Classification and Regression Trees to yield a better overall classification rate.

In their study of "Modeling partial customer churn: On the value of first product-category purchase sequences",(Miguéis, Poel, Camanho, & Falcão, 2012) used LR as the classification technique. They took real sample of approximately 75,000 new customers from the data warehouse of a European retail company and use it to test their proposed models. They used the area under the receiver operating characteristic curve and 1%, 5% and 10% percentiles lift to assess the performance of the partial-churn prediction models. The empirical results of their study revealed that both proposed models outperform the standard RFM model.

(Sharma & Panigrahi, 2011) proposes a neural network (NN) based approach to predict customer churn in subscription of cellular wireless services. The results of their experiments indicate that neural network based approach can predict customer churn with accuracy more than 92%. Further, they

observed that medium sized NNs perform best for the customer churn prediction when different neural network's topologies were experimented.

In their study on modeling purchasing behavior of bank customers in Indian scenario, (Prasad & Madhavi, 2012) applied predictive DM techniques. They experimented with 2 classification techniques namely CART, and C 5.0. The prediction success rate of Churn class by CART is quite high but C 5.0 had shown poor results in predicting churn customers. However, they found that the prediction success rate of Active class by C 5.0 is more effective than the other technique. The related works in a global context are summarized in Table 2.2

Table 2.2 Summary of related works (global)

Area	Business	Techniques Applied	Results	Reference
Churn Prediction using credit card data	Banking	LR and Decision Tree for classifying the classes	LR performs a little better than Dec. Tree	(Nie et al., 2011)
Customer Retention using predictive technique	Banking	Classification and Regression Tree (CART)	CART yields better classification rate	(Chitra & Subashini, 2011)
Modeling Partial churn for product-category sequences	Retail	LR as the classification technique	The proposed models outperform the standard one	(Miguéis et al., 2012)
Proposes an approach to predict customer churn	cellular wireless services	a neural network (NN) based approach	The approach can predict customer churn with accuracy more than 92%	(Sharma & Panigrahi, 2011)
Modeling purchasing behavior	Bank	2 classification techniques namely CART, and C 5.0 to predict churn	The success rate of Churn class by CART is quite high	(Prasad & Madhavi, 2012)

2.2.8.2 Local DM Research works

Many DM and related research works have been done by local researchers; though in my observation no research works in churn prediction subject. In fact, there are few researches in CRM area, which covers the other dimensions of CRM (Stated in Section 2.1.3.2). In this section, the local works, which are somehow related to this study, are presented.

(Henock, 2002)conducted his study targeted at testing the application of data mining techniques to support CRM activities at Ethiopian Airlines. He considers the Ethiopian Airlines' frequent flyer

program's database, which contains individual flight activity and demographic information of more than 22,000 program members. He applied the K-means clustering algorithm was used to segment individual customer records into clusters with similar behaviors and then the decision tree classification techniques were employed to generate rules that could be used to assign new customer records to the segments. The study revealed encouraging results of DM techniques in supporting CRM. (DENEKEW, 2003) has also conducted his study in the same area to fill the gaps, which are left as a further study by (Henock, 2002).

(Kumneger, 2006) conducted his study on the application of DM techniques to support CRM for the case of the Ethiopian Shipping Lines. He used customer profile file of ESL having more than 20,000 records. He applied K-Means clustering algorithm to segment individual customer records in to clusters with similar behaviors and then decision tree classification techniques were employed to generate rules that could be used to assign new customer record to the segments. The study showed encouraging results in the applicability of DM techniques in supporting CRM.

In the banking and insurance domain (TARIKU, 2011)tried to develop models that can detect and predict fraud in insurance claims in the case of Africa Insurance Company applying clustering algorithm followed by classification techniques for developing the predictive model. He has tried to apply first the K-Means clustering algorithm followed by classification techniques the J48 decision tree and Naïve Bayes algorithms for developing the predictive model. The J48 model shows a better accuracy of 97.19% in classifying new insurance datasets as fraud and non-fraud suspicious claims.(LUEL, 2011) worked the application of DM technology in the expansion of electronic transaction at Dashen Bank and used decision tree J48, ANN classification algorithms, and K- means clustering. The ANN model has shown the best classification accuracy.(Tesfaye, 2002) also tried to develop predictive modeling to support insurance risk assessment of motor insurance policies. He used Decision tree and neural network modeling techniques. The local related works are summarized in Table 2.3

Table 2.3 Summary of Local related works

Area	Business	Techniques Applied & Results	Reference
Application of DM techniques to support CRM	Ethiopian Airlines	Apply K-Means clustering and Decision Tree classification techniques to classify customers	(Henock, 2002)
>>	>>	Similar with (Henock, 2002) but to fill the gap in the earlier research. He applied J48 technique	(DENEKEW, 2003)

Area	Business	Techniques Applied & Results	Reference
>>	Ethiopian Shipping Lines	Apply K-Means clustering and Decision Tree classification techniques to classify customers	(Kumneger, 2006)
Develop models that can detect and predict fraud	Insurance (Africa Insurance Company)	Clustering algorithm (K-Means) followed by classification techniques (J48 and Naïve Bayes) for classifying Fraud and Non-Fraud instances. J48 shows better classification accuracy.	(TARIKU, 2011)
Application of DM technology in the expansion of electronic transaction	Banking (Dashen Bank)	Applied decision tree (J48), ANN classification algorithms, and K- means clustering. The ANN model has shown the best classification accuracy.	(LUEL, 2011)
Develop predictive modeling to support insurance risk assessment process	Insurance	Applied Decision tree and neural network modeling techniques.	(Tesfaye, 2002)

In this research, models are built that depicts churn behavior of bank customers and predict churners, the case of CBE. Decision tree, LR and Bagging modeling techniques are applied and the algorithm with the best predictive performance is to be selected. Though the aforementioned researches have substantial contribution for showing the directions in conducting this research, this area (churn prediction) is the first time to be researched locally. So, the research fills the knowledge gap as to how churners can be predicted from existing historical data.

CHAPTER THREE

3 DM Techniques for Churn Prediction

Various DM techniques are being used for churn prediction related problems. According to (Bhambri, 2012) some of the various techniques of DM which can be applied in the prediction of churn behavior of customers are Association, Clustering, Forecasting and Classification. The contribution of each technique in predicting customers' churn is briefly stated as follows:

1. ***Association***: - Association and correlation is used to find frequently used data items in the large data sets. It is the technique of finding patterns where one event is connected to another event.
2. ***Clustering***: - For the identification of similar classes of objects. This is the technique of combining the transactions with similar behaviour into one group, or the customers with same set of queries or transactions into one group.
3. ***Forecasting***: - Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more dependent and independent variables.
4. ***Classification***: - Classification is the most commonly applied DM technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large.

(Sharma & Panigrahi, 2011) opines that statistical and DM techniques have been utilized to construct the churn prediction models. They believe that neural networks, support vector machines and LR models are among the popular techniques to predict customer churn. In the same context, (Ogwueleka, 2009)also stated that some of the DM techniques which are mostly used and appropriate to construct churn prediction models are LR, neural networks and decision trees. Assessing the three prediction models, decision trees are the most interpretable in that they can be translated into decision rules. In the case of neural networks, decision trees can be used to model complex non- linear and interaction relationships

In their literature review of churn prediction techniques, (Ballings & Poel, 2012) noted that the two most commonly used analytical techniques are LR and classification trees. The performance of these two techniques is very similar and depends on a multitude of factors such as the normality of the data, the number of categorical variables, the size of the training sample and the signal-to-noise ratio. In

addition, (Berry & Linoff, 2004) stated that binary outcome churn models can be built with any of the usual tools for classification including LR, decision trees, and neural networks.

In the next few sections, the DM modeling techniques, which are stated by various researchers as the most widely used and the most appropriate for churn prediction, are discussed. In fact, these are the modeling techniques to be used in this study.

3.1 Classification Technique for Prediction

(Jantan, Hamdan, & Othman, 2010) Classification and prediction are among the methods that can produce intelligent decision. Currently, many classification and prediction methods have been proposed by researchers in machine learning, pattern recognition, and statistics. Classification and prediction in DM are two forms of data analysis that can be used to extract models to describe important data classes or to predict future data trends.

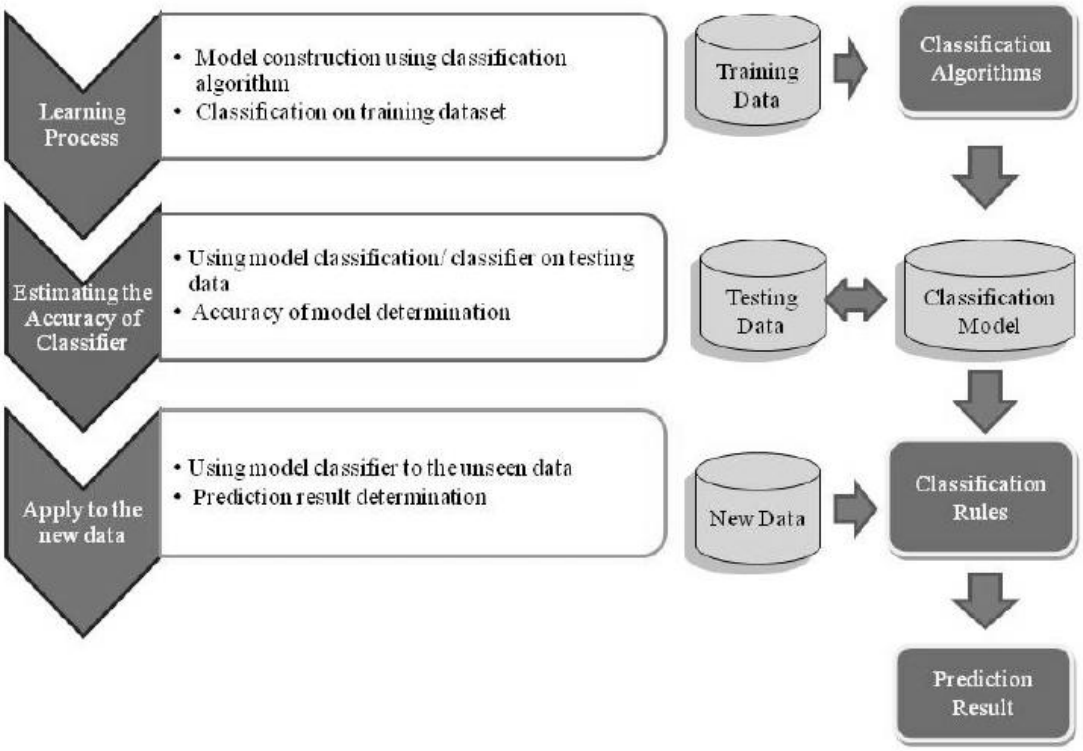


Figure 3.1: Classification Process in DM(Jantan et al., 2010)

As shown in Figure 3.1, the classification process has two phases; the first phase is learning process where the training data are analyzed by the classification algorithm. Learned model or classifier is represented in the form of classification rules. The second phase is classification process, where the

test data are used to estimate the accuracy of classification model or classifier. If the accuracy is considered acceptable, the model can be applied to the new data to know the prediction result.

There are many techniques that can be used for classification such as decision tree, Bayesian methods, Bayesian network, rule-based algorithms, neural network, support vector machine, association rule mining, k-nearest-neighbor, LR, genetic algorithms, Bagging & Boosting and fuzzy logic. Different authors proposed their best modeling techniques for predicting customers' churn in different criterion. (Nie et al., 2011)believe those LR and decision trees are mature DM algorithms to build models and predict customers' churn. Again, from a business perspective,(Radosavljevik et al., 2010) stated that the output of either decision trees or LR algorithms (the model) is very easy to interpret and communicate to parties that do not have extensive experience with DM (e.g. business managers). This quality is even more evident in the case of decision trees, which have a very intuitive graphical representation. In this study, the discussion focuses on the three main classification techniques i.e. decision tree (J48), LR and Bagging.

3.1.1 Decision Tree

The decision tree technique is one of the most intuitive and popular DM methods, especially as it provides explicit rules for classification and copes well with heterogeneous data, missing data and non-linear effects (Tufféry, 2011).

The decision tree technique is used in classification to detect criteria for dividing the individuals of a population into n predetermined classes (in many cases, n=2). We start by choosing the variable which, by its categories, provides the best separation of the individuals in each class, thus providing sub-populations, called nodes, each containing the largest possible proportion of individuals in a single class; the same operation is then repeated on each new node obtained, until no further separation of the individuals is possible or desirable (according to criteria which depend on the type of tree). The construction is such that each of the terminal nodes (the leaves) mainly consists of the individuals of a single class. An individual is assigned to a leaf, and therefore to a certain class, with a reasonably high probability, when it conforms to all the rules for reaching this leaf. The set of rules for all the leaves forms the classification model as it is shown in Figure 3.2

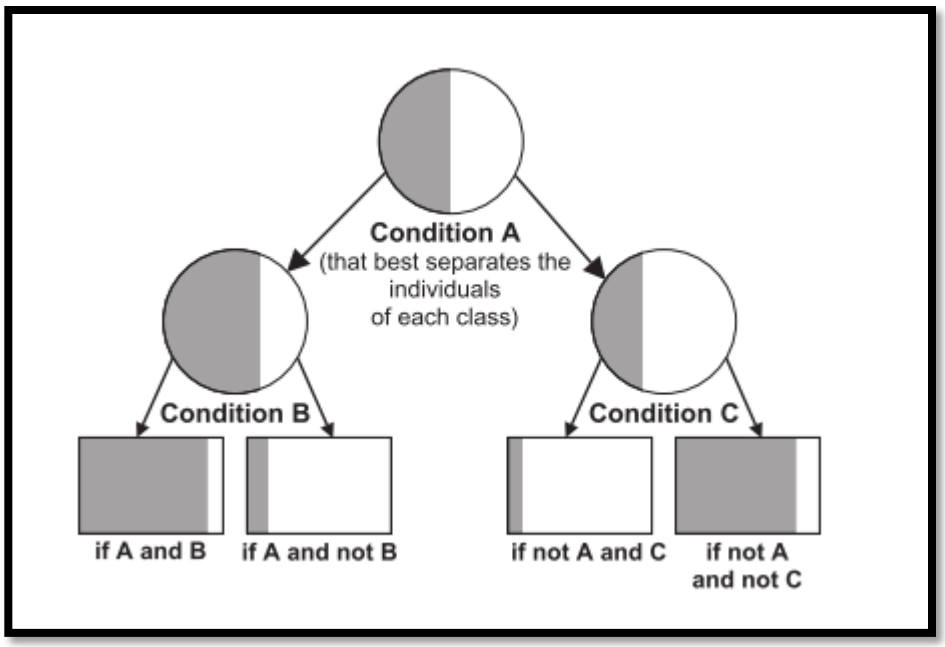


Figure 3.2: Decision Tree(Tufféry, 2011)

In the decision-tree method, each internal node tests a feature, each branch corresponds to a feature value, and each leaf node assigns a classification (Dua & Du, 2011). The methodology for using decision tree is described as follows:

Step 1: - Split a variable at all of its split points. Sample sections into multiple nodes at each split point.

Step 2: - Select the best split in the variable in terms of splitting criterion.

Step 3: - Repeat Steps 1 and 2 for all variables at the root node.

Step 4: - Rank the best splits and select the variable that achieves the highest purity at the root.

Step 5: - Assign classes to the nodes according to a rule that minimizes misclassification costs.

Step 6: - Repeat Steps 1–5 for each non-terminal node.

Step 7: - Grow a large tree until each leaf is pure.

Step 8: - Prune and choose the final tree using the cross validation (CV).

In the above steps, the splitting criterion plays a critical role in the feature selection process for splitting. The process employs a feature selection measure, such as information gain (IG).

Decision tree construction process is top- down, divide-and-rule. Essentially it is a greedy algorithm (Jain & Upendra, 2012). Starting from root node, for each non-leaf node, firstly choose an attribute to test the sample set; Secondly divide training sample set into several sub-sample sets according to testing results, each sub-sample set constitutes a new leaf node; Thirdly repeat the above division

process, until having reached specific end conditions. In the process of constructing decision tree, selecting testing attribute and how to divide sample set are very crucial. Different decision tree algorithm uses different technology. In practice, because the size of training sample set is usually large, the branches and layers of generated tree are also more. In addition, abnormity and noise existed in training sample set will also cause some abnormal branches, so we need to prune decision tree. One of the greatest advantages of decision tree classification algorithm is that: It does not require users to know a lot of background knowledge in the learning process.

Decision tree can produce a model with rules that are human-readable and interpretable. The classification task using decision tree technique can be performed without complicated computations and the technique can be used for both continuous and categorical variables (Jantan et al., 2010). This technique is suitable for predicting categorical outcomes and less appropriate for application with time series data. Decision tree classifiers are quite popular techniques because the construction of tree does not require any domain expert knowledge or parameter setting, and is appropriate for exploratory knowledge discovery. It is believed by many researchers that, the decision tree is among the powerful classification algorithms. A Decision Tree is one of the most popular classification algorithms in current use in DM and Machine Learning (Endo, Shibata, & Tanaka, 2008). Decision tree technology is a common; intuitionist and fast classification method (Jain & Upendra, 2012). Some of decision tree classifiers are C4.5/C5.0/J4.8, NBTree, SimpleCart, REPTree, BFTree and others. But here, the focus of the discussion is J48 as it is one of the algorithms to be used for modeling in this study.

3.1.1.1 J.48 Algorithm

Decision tree J48 implements Quinlan’s C4.5 algorithm for generating a pruned or un-pruned C4.5 tree. C4.5 is an extension of Quinlan's earlier ID3 algorithm. J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets.

The C4.5 technique is one of the decision tree families that can produce both decision tree and rule-sets; and construct a tree (Rajput, Aharwal, Dubey, Saxena, & Raghuvanshi, 2011). Besides that, C4.5 models are easy to understand as the rules that are derived from the technique have a very straightforward interpretation. J48 classifier is among the most popular and powerful decision tree classifiers. C5.0 and J48 are the improved versions of C4.5 algorithms. WEKA toolkit package has its own version known as J48. J48 is an optimized implementation of C4.5.

J48 examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data (Aruna, Rajagopalan, & Nandakishore, 2011). To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling to choose that class. But it can also happen that none of the features give any information gain. In this case J48 creates a decision node higher up in the tree using the expected value of the class.

J48 can handle both continuous and discrete attributes, training data with missing attribute values and attributes with differing costs (Aruna et al., 2011). Further it provides an option for pruning trees after creation

3.1.2 LR

(Hlosta, Stríž, Kup, Zendulka, & Hruška, 2013)LR (LR) is a machine learning model for binary classification. The method can handle both numeric and categorical variables. Given a learned model, the value of the output variable is computed by applying the logistic function to linear combination of attribute values and weight vector. The function is defined as follows:

$$P(Y_i = 1 \mid x_i, w) = \frac{1}{1 + e^{-w^T x_i}}$$

The logistic function converts the input value to interval [0, 1]. The result describes a confidence value for a given case being of the class 1. Typically, threshold t = 0.5 is applied to determine whether an examined example belongs to class 0 or 1.

Regression is the analysis, or measure, of the association between a dependent variable and one or more independent variables(B.K. & Srivatsa, 2011). This association is usually formulated as an equation in which the independent variables have parametric coefficients that enable future values of the dependent variable to be predicted. Two of the main types of regression are: linear regression and LR. In linear regression the dependent variable is continuous and in logistic it is either discrete or categorical. For LR to be used, the discrete variable must be transformed into a continuous value that is a function of the probability of the event occurring. Regression is used for three main purposes:

1. Description,
2. control and
3. Prediction.

LR is also called as logistic model or logit model, is a type of predictive model which can be used, when the target variable is a categorical variable with two categories - for example active or inactive, healthy or unhealthy, win or loss, purchase product or doesn't purchase product etc. LR is used for the prediction of the probability of occurrence of an event by fitting the data into a logistic curve. Like many forms of regression analysis, it makes use of predictor variables; variables may be either numerical or categorical. For example, the probability that a person has a heart attack in a specified time that might be predicted from the knowledge of person's age, sex and body mass index. LR is used extensively in the medical and social sciences as well as in marketing applications such as prediction of customer's propensity to purchase a product or cease a subscription. The response, Y, of a subject can take one of two possible values, denoted by 1 and 0 (for example, Y=1 if a disease is present; otherwise Y=0). Let $X=(x_1, x_2, \ldots, x_n)$ be the vector of explanatory variables. The LR model is used to explain the effects of the explanatory variables in the form of binary response.

$$\text{Logit}\{Pr(Y=1|x)\} = \log\{Pr(Y=1|x) / (1- Pr(Y=1|x))\} = \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\ldots+\beta_kx_k$$

Where β_0 is called "the intercept" and $\beta_1+\beta_2+\beta_3$, and so on are called the "regression coefficients" of $x_1+x_2+x_3$respectively. Each of the regression coefficients describes the size of the contribution of the risk factor. A positive regression coefficient means that the risk factor increases the probability of outcome, where as a negative regression coefficient means that the risk factor decreases the probability of outcome, a large regression coefficient means that the risk factor strongly influences the probability of that outcome, a non-zero regression coefficient means that the risk factor has little influence on the probability of outcome. The logistic function is given by: $P=1/ (1+e^{-\text{logit}(p)})$

A graph of the function is shown in Figure 3.3. The logistic function is useful because it can take an input any value from negative infinity to positive infinity, whereas the output is confined to values between 0 and 1.

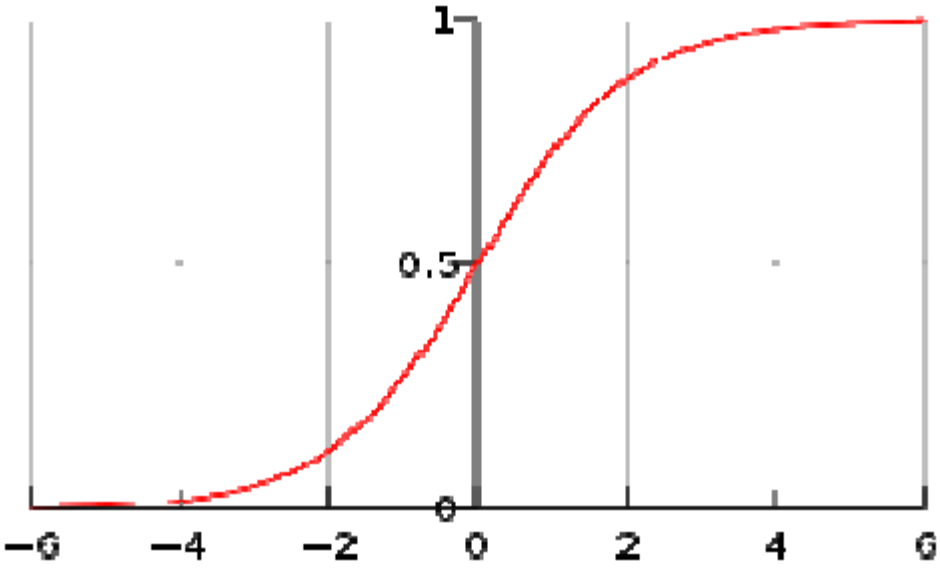


Figure 3.3: A graph of LR function(B.K. & Srivatsa, 2011)

3.1.3 Bagging Algorithms

Bagging Algorithms can be constructed in such a way that by combining the properties of two or more algorithms.(Witten, Frank, & Hall, 2011) stated that combining the decisions of different models means amalgamating the various outputs into a single prediction. The simplest way to do this in the case of classification is to take a vote (perhaps a weighted vote); in the case of numeric prediction it is to calculate the average (perhaps a weighted average). Bagging and boosting both adopt this approach, but they derive the individual models in different ways. In bagging the models receive equal weight, whereas in boosting weighting is used to give more influence to the more successful ones—just as an executive might place different values on the advice of different experts depending on how successful their predictions were in the past. To introduce bagging, suppose that several training datasets of the same size are chosen at random from the problem domain. Imagine using a particular machine learning technique in order to build a decision tree for each dataset. You might expect these trees to be practically identical and to make the same prediction for each new test instance. But, surprisingly, this assumption is usually quite wrong, particularly if the training datasets are fairly small. This is a rather disturbing fact and seems to cast a shadow over the whole enterprise! Slight changes to the training data may easily result in a different attribute being chosen at a particular node, with significant ramifications for the structure of the sub-tree beneath that node. This automatically implies that there are test instances for which some of the decision trees produce correct predictions and others do not.

We can combine trees by having them vote on each test instance. If one class receives more votes than any other, it is taken as the correct one. Generally, the more the merrier: Predictions made by voting become more reliable as more votes are taken into account. Decisions rarely deteriorate if new training sets are discovered, trees are built for them, and their predictions participate in the vote as well. In particular, the combined classifier will seldom be less accurate than a decision tree constructed from just one of the datasets. (Improvement is not guar- anteed, however. It can be shown theoretically that pathological situation exists in which the combined decisions are worse.)

(Bauer & KOHAV, 2004)The Bagging algorithm (Bootstrap aggregating) by Breiman votes classifiers generated by different bootstrap samples (replicates). Figure 3.4shows the algorithm. A Bootstrap sample is generated by uniformly sampling m instances from the training set with replacement. T bootstrap samples B_1, B_2,..., B_T are generated and a classifier C_i is built from each boot- strap sample B_i. A final classifier C^* is built from C_1,C_2,...,C_T whose output is the class predicted most often by its sub-classifiers, with ties broken arbitrarily. For a given bootstrap sample, an instance in the training set has probability $1-(1-1/m)^m$ of being selected at least once in the m times instances are randomly selected from the training set. For large m, this is about $1-1/e$ =63.2%, which means that each bootstrap sample contains only about 63.2% unique instances from the training set. This perturbation causes different classifiers to be built if the inducer is unstable (e.g., neural networks, decision trees) and the performance can improve if the induced classifiers are good and not correlated; however, Bagging may slightly degrade the performance of stable algorithms (e.g., k-nearest neighbor) because effectively smaller training sets are used for training each classifier. The Bagging algorithm is shown in Figure 3.4

Input: training set S, Inducer $\mathcal{I}$, integer T (number of bootstrap samples).

1. for $i = 1$ to T {
2. $S' =$ bootstrap sample from S (i.i.d. sample with replacement).
3. $C_i = \mathcal{I}(S')$
4. }
5. $C^*(x) = \arg\max_{y \in Y} \sum_{i:C_i(x)=y} 1$ (the most often predicted label y)

Output: classifier C^*.

Figure 3.4 : The Bagging Algorithm(Bauer & KOHAV, 2004)

3.2 Handling Class Imbalance

Churn is obviously a rare event. Customer churn is often a rare event in service industries, but of great interest and great value(Burez & Poel, 2009). According to (Bekkar & Alitouche, 2013), when a model is trained on an imbalanced data set, it tends to show a strong bias to the majority class, since classic learning algorithms intend to maximize the overall prediction accuracy. In addition, (Batista, Prati, & Monard, 2004) opine that learning from imbalanced data sets is often reported as being a difficult task.

Generally, there are two ways of handling class imbalance problems. These are oversampling the minority class and under-sampling the majority class. Both of these techniques are reported to have their own drawbacks. Under-sampling may throw out potentially useful data, while over sampling increases the Training Set size and hence the time to train a classifier. According to (Weiss, 2004), advanced sampling methods may use intelligence when removing/adding examples or combine under-sampling and oversampling techniques. One under-sampling strategy only removes majority-class examples that are redundant with other examples or border regions with minority-class examples, figuring that they may be the result of noise. SMOTE on the other hand, over-samples by introducing new, non-replicated minority-class examples. Minority-class examples are generated by adding examples from the line segments that join the k minority-class nearest neighbours (SMOTE uses k=5). This causes additional generalization, as opposed to the specialization that may arise from exactly replicating examples.

The SMOTE method, which is an advanced method of over-sampling, aims to make the decision borders of minority class more general, and thus turned the issue over-fitting with basic over-sampling (Bekkar & Alitouche, 2013). In order to handle the class imbalance of the training sets SMOTE sampling technique can be applied so that the proportion of the two classes to become equivalent (50% each).

CHAPTER FOUR

4 The Business Domain and the Data

4.1 General Understanding of the Business Domain (CBE)

As it is stated in section 1.5.1, thorough understanding of the business domain under investigation is a very important step for clearly stating the business objectives, the DM goal, the resources needed for conducting the project, the associated risks and major constraints, and for planning each phases of the project. The knowledge about the business domain (CBE) is acquired through different techniques such as:

- Visiting and browsing the official website of CBE,
- Referring to various printed medias of the bank (Annual reports, Magazines, and others)
- Having successive discussions with the relevant authorities and experts of the bank.

The list of questions prepared in order to guide the discussions with the relevant personnel of the CBE is enclosed in Appendix III

4.1.1 The Business Objectives of CBE

CBE is engaged in a commercial banking service. For a better management of the overall tasks, there are about fourteen districts (four in Addis Ababa and the rest ten in regions) to coordinate the day-to-day activities of each branch allocated to them.

The bank is operating under the top level supervision of a board of directors which is composed of a board chairman and nine board members. Reporting to the board of directors, there is a Process Council in CBE that manages the bank's core activities at a higher managerial level. The process council members of the bank include the following top managerial positions:

1. The President,
2. Vice President – Credit Management,
3. Vice President – Trade Services,
4. Vice President – Credit Appraisal and Portfolio Management,
5. Vice President – Human Resource Management,
6. Vice President – Information Systems,
7. Vice President – Customer Account Transaction Service (CATS),

8. Vice President – Finance,
9. Vice President – Facilities Management,
10. Chief Business Development Officer,
11. Chief Risk and Compliance Management Officer,
12. Chief Internal Auditor,
13. Chief Legal and Loan Recovery, and
14. Director Strategy Management,

The bank is self-financed from the profits generated from the services being provided in each branch. As a government owned bank, one of the major objectives of CBE is to avail banking service to citizens all over the country and abroad (in some places). For that reason, the bank is opening branches in the remotest part of the country even for no profit. Obviously, high profits are obtained from the branches in the urban areas and this also significantly contributes to the overall profitability of the bank. According to the annual report of the bank (CBE, 2011), the 2003 E.C budget year net profit and the total assets of CBE in millions are 2,862.90 and 114,265.10 respectively.

The CRM of CBE is not organized within a specified department; rather it is dispersed in different working sections. The Credit related CRM activities of the bank is led by the Credit Management Vice President and is organized into three directorates in accordance with the amount of credit and the customer type. These are: Business Corporate CRM, Commercial CRM, and Consumer Loan Relationship Management. The Trade Services CRM is also similarly led by Trade Services Vice President and having central organization to manage the CRM process. The case of CRM in Deposit services is different from the aforementioned two services. In every branch there is a CRM Deputy Branch Manager and also a Customer Relationship Officer assigned to perform all the CRM related activities. The Customer Account Transaction Service (CATS), which is led by a Vice President, is responsible for each and every account transactions made in all the branches all over the country.

The main focus of this research is in the CRM related activities of the private sector deposit service. Hence, the CATS division is the one to which this research work is intended to be conducted. This is the division to be affected after the accomplishment of the DM project, provided that the outcome of the research is accepted by the bank officials and the bank management is willing to deploy the models.

This study is initiated to alleviate CRM related problems observed in CBE. The problems are identified thorough discussions with various relevant authorities of the bank. These authorities also

show their interest that a DM project needs to be conducted to propose solutions for the observed problems by using the existing data. But they also expressed that deploying the outcomes of the DM project (the final model) has its prerequisites. It needs to be accepted by the management as well as the board. So, they are not ready to deploy the model as soon as the research work is done. It needs discussion among the management and the board before such outcomes are to be deployed. The reason for this is that DM solutions are new for the bank and such initiatives are usually undertaken only from the bank's business development itself.

Customers' churn is the main CRM problem observed in CBE. It is observed that customers are closing their accounts for unknown reasons. The bank doesn't have an instrument for collecting feedback from churners as to why they are closing their account. Currently, to reduce the number of churners, the CRM officials assigned at each branch try to convince customers to change their mind while the customers are requesting account closing service. Though it is a good trend its success rate is very low because the bank is trying to convince a customer who has already decided to close his/her account. It is believed that the attempt might be successful if the customer had been communicated earlier before he/she made a decision.

CBE, as its customer's retention objective, would like to have a long-term relationship with all its customers and doesn't want to lose even a single customer. Thus this study is commissioned as part of the objective of keeping current customers by predicting when they are prone to move to a competitor (reduce the churn of customers in CBE).

As stated in section 2.1.4.3, predicting customers who are prone to move to a competitor is only the part of Customers' churn management process but not the ultimate solution to reduce churn. The bank would like to reduce at least 85% of the expected churners through Customers' churn management process. For this business objective to succeed, customers who are prone to move to a competitor should be predicted with an accuracy of a bit greater than 85% (assuming contingencies in other churn management processes).

4.1.2 Assessment of the Existing Situation

The basic data and knowledge required for the project is customer transactional data accumulated in the database (from which the knowledge is going to be extracted) and knowledge regarding the whole organizational structure of the bank, the CRM processes, the existing problems in relation to CRM and the likes. As the bank is currently using enhanced banking database system, the customer related data can be obtained from the system and it is the only and sufficient source of data. In order to respect the

privacy of customers' data, attributes such as the customer's name (individual, group, or company), telephone numbers, account number, and other sensitive information are not required for the project. The knowledge can be obtained from the official website of the bank, from different documentations (such as: The Annual Reports of the bank, Bulletins and others), and through discussions with the concerned experts and managers of the bank. The researcher, who is a domain expert of Information Science, is the main personnel resource to conduct the DM project. In addition, the experts in the bank are also assumed to have significant contribution, as they will be involved in providing the required data for the research.

The bank, as a policy, does not expose the privacy of its customers in any way. So, fields such as: Name, Address, and Telephone Number … are to be filtered by the bank's data experts themselves. Such data do not have contribution for the outcomes of the research too. As this study is basically an educational research, it is expected to deliver a good quality result & accomplished within the specified academic time frame of the institute. However, for a better results achievement and successful accomplishment of the research work, it is assumed that a good quality data (which includes the features required for the study) is provided on time. As stated earlier, the bank authorities would like to see the outcomes at the end of the project. They might deploy the results in the future provided that the obtained results are interesting and only accepted by the top level management of the bank.

Budget is considered as the major constraint of this project. As no budget is allotted for DM tools purchase, future arousal of unexpected need of DM tools, which couldn't be procured for free, might affect the quality of the result. Other than the risk in relation to budget, unexpected situations might occur regarding the schedule and the data. If the data is of poor quality and not delivered in line with the project plan, the quality of the result and the project schedule are going to be affected. In case such conditions happen, the researcher is expected to explain to the concerned data experts of the bank about the quality of data expected for the project and wait patiently till the expected data is obtained. To compensate the wasted time, it might become obligatory to work for additional hours during the remaining project time or else additional time to finish the project will be requested otherwise.

For this project, a laptop computer having a specification of 3GB RAM, 300GB hard disk, Core2Duo 2.53GB speed laptop with Windows7 operating system is used.

4.1.3 DM Goals

The DM project is conducted in order to reduce the number of churner customers in CBE and also identify potential churners based on the patterns from the customers' data so that earlier churn management precautions could be taken by the management (such as communicating the predicted churner before he/she comes to close his/her account). Though churn management processes, which are the next steps undertaken after churn prediction, are not the scope of this study. Transactions in the final month (status month) by customers is not going to be considered for modeling; assuming that this period will be used for conducting other churn management steps.

The major theme of this study is to identify the customers, which are prone to churn out (close their accounts), out of a combination of active and churned customers; based on the patterns shown in their historical data. The problem is a classification DM problem type (classifying churning and active customers based on their pattern similarity). Hence, the DM goal of the project to be conducted as part of this research is to build a model using available customers' data in order to predict those customers who are going to close their accounts.

Models, which are intended to predict churners, are built using the existing customers' data by applying various DM tools and techniques. The models are assessed on their performance in predicting churners. As the business objective is to reduce at least 85% of churners, the models should correctly predict more than 85% of the churners. An accuracy of about 90% is, according to (Lazarov & Capota, 2007), sufficient for a classifier to predict churning. So, the DM is judged as success if 90% of churners can be predicted correctly.

4.1.4 Project Plan to Meet the DM Objectives

The DM project is conducted phase by phase. For each phase the average time required, and the associated risks might happen are depicted in Table 4.1

Table 4.1: DM Project Plan

PHASES	ALLOTED TIME	Methodology Applied (Steps)	RISKS
Business Understanding	1 week	Reviewing printed medias, visit website, and having discussions with relevant personnel	Resource availability
Data Understanding	2 weeks	Data description, exploration, & verify quality	Tools
Data Preparation	4 weeks	Data selection, cleaning, integration…	Tools
Modeling	2 weeks	Tools & tech, selection, test design	Tools
Evaluation	1and1/2 week	Evaluating the result from the best model	

The DM tools and techniques to be used in data preparation and experimentation processes are selected based on the following criterion:

1. The tools should be free as no budget is allotted for procuring DM tools,
2. The tools should have the capability of performing the modelling processes.
3. The techniques should be among the most widely used classification modelling techniques, which are supported by many authors as appropriate for churn prediction.

Based on the aforementioned criterion WEKA, MS-EXCEL and MS-ACCESS are used data preparation processes. WEKA tool is to be used for modeling. And the techniques to be employed are Decision Tree (J48), LR, and Bagging.

4.2 DATA UNDERSTANDING

4.2.1 Description and Exploration of the Data Collected from CBE

To conduct the DM project successfully, it is expected to get a customer related data from CBE. These data should be of quality standard for predicting customer churn. The data should contain:

- Demographic information of customers (age, sex, type of work, monthly income …)
- The date the account has been created,
- The initial amount saved,
- The transaction details of at least two years (this means how much amount is debited or credited and the balance after each transaction)
- The current status of the account, and the like

Data such as the name of the customer, telephone number, address (woreda, kebele, house number, etc.) are not needed for the experiment.

Though all the customers provide every detail during subscription, the bank records in the system only those fields, which they assume to be the most important for the day-to-day transaction. The other details such as the demographic information about customers are left in the papers (forms) and are not recorded in the system. As a result, the demographic information, which could have significant contribution for churn prediction, couldn't be obtained from the bank for the DM project.

CBE has a privacy issue while providing data about their customers. To keep the privacy of customers a data expert is assigned to extract the intended data from the systems and to filter all the attributes which can expose the personal privacy of customers. Currently the bank is in a process to migrate the

data from the earlier SMART system to the new CORE-Banking system. So, for the purpose of consistency of the data, the earlier (SMART) system has been used to extract the data for the DM project. The data is extracted by taking sample from three branches of CBE by the data expert himself.

The data is prepared in two parts. The first part, which is the Customers' Account detail, contains data about the customers' account (fixed in nature). The second part is the corresponding transactional data of those customers which are selected in the first part. Three set of samples are taken from three different branches by the data expert himself in a tab delimited text format, then imported to excel (.xlsx) format. The data is delivered after being filtered for privacy reason (The Customer's Name, Telephone Number, etc. are deleted and the Account Number is altered in both tables in a similar fashion to keep the consistency).

MS-EXCEL, MS-ACCESS, and WEKA 3.7.10 (Hall et al., 2013) tools are used for data exploration and data preparation purposes. The description of the data is presented separately for the two types of data as follows:

4.2.1.1 The Customers' Account Data:

This table contains Account information of 13171 customers of various types in 9 attributes. The types of each attribute and the corresponding values, descriptions, and related information is summarized in Table 4.2

Table 4.2: Description of the Attributes in the Customers' Accounts table

Attribute	Type	Value	Description & Basic Statistics	Count	Missing Values
AcctNo	Alpha-numeric		To Identify each Customer Uniquely	13171	None
AcctType	Alpha-numeric	S01	SAVINGS A/C - PRIVAE SECTOR	12142	None
		S02	COOPERATIVES	69	
		S03	SPECIAL DEMAND DEPOSIT	953	
		S05	SAVINGS A/C- PUBLIC AGENCIES	7	
Comp/Ind	Alpha-numeric	C	COMPANY CUSTOMER	64	None
		I	INDIVIDUAL CUSTOMER	13107	
CurrCode	text	ETB	Ethiopian Birr	13171	None
AcctStatus	text	ACT	ACTIVE	9701	None
		INA	INACTIVE	3019	
		NEW	NEW	19	
		CLS	CLOSED	415	

Attribute	Type	Value	Description & Basic Statistics	Count	Missing Values
		CIA	Cash Indemnity A/C	7	
		BLK	Blocked A/C	9	
		UNC	UNCLAIMED	1	
CurrentBalance	double		The balance of the customer at the date of status mentioned. Min=0.00, Max=1732139.61 Mean=6393.90 StdDev=39643.10	13171	None
DateAcctOpened	Date	mm/dd/yy	The date at which the account is opened Ranges b/n 11/30/07 & 05/24/13	13171	None
DateAcctClosed	Date	mm/dd/yy	The date at which the account is closed (if any) Ranges b/n 09/26/09 &05/22/13	394	12777
DateOfStatus	Date	mm/dd/yy	The date at which the account status of the customer is taken Ranges b/n 09/26/09 & 05/22/13	3711	9460

4.2.1.2 The Customers' Transactions Data:

For the customers' accounts stated before, the day-to-day transactional information is included in this table. The table contains about 628634 transactional values and 10 attributes. The detail of the customers' transactional data in shown in Table 4.3

Table 4.3: Description of the data in the Customers' Transaction table

Attribute	Type	Value	Description	Count	Missing Values
AcctNo	Alpha-numeric		To Identify each Customer Uniquely	628634	None
AcctType	Alpha-numeric	S01	SAVINGS A/C - PRIVAE SECTOR	619751	None
		S02	COOPERATIVES	801	
		S03	SPECIAL DEMAND DEPOSIT	8031	
		S05	SAVINGS A/C- PUBLIC AGENCIES	5051	
Comp/Ind	Alpha-numeric	C	COMPANY CUSTOMER	1282	None
		I	INDIVIDUAL CUSTOMER	627352	
CurrCode	Text	ETB	Ethiopian Birr	628634	None

Attribute	Type	Value	Description	Count	Missing Values
DateOfTxn	Date		Specifies the date of transaction Ranges b/n 09/26/09 & 05/24/13	628634	None
Dr/CrFlag	Text	D	Specifies that the transaction is a Debit Transaction	348917	
		C	Specifies that the transaction is Credit Transaction	279717	
OperCode	Text	CCO	COMMISSION ON MT,TT,DD -LOCAL	1853	None
		CRE	CREDIT THE ACCOUNT	23720	
		DEB	DEBIT THE ACCOUNT	943	
		DEP	DEPOSITS	41836	
		IAC	INTEREST ACCRUAL-CREDIT	131481	
		ICC	INTEREST CAPITALIZATION-CREDIT	118026	
		IAD	INTEREST ACCRUAL-DEBIT	118037	
		IBS	INWARD BLOCKING SAVINGS A/C	765	
		ITX	INCOME TAX	118016	
		LDD	LOAN DISBURSEMENT BY A/C - CR	3	
		MRC	MT RECD AND CREDITING THE CUST. A/C	1	
		OBL	OUTWARD BLOCKING CURRENT A/C	1	
		OPC	OPENING BALANCE-CREDIT	1104	
		TEL	TELEPHONE CHARGES	1852	
		TRC	TT RECD AND CREDITING THE CUST. A/C	2468	
		TSD	TT SENT BY DEBITING CUSTOMER A/C – DR	2252	
		WDL	WITHDRAWALS	66276	

Attribute	Type	Value	Description	Count	Missing Values
DrForeignAmt	double		Debited amount in foreign currency Min=0.00, Max=5377568.61, Mean=1197.27, StdDev=18420.65	628634	None
DrLocalAmt	double		Debited amount in local currency Min=0.00, Max=5377568.61, Mean=1197.27, StdDev=18420.65	628634	None
CrForeignAmt	double		Credited amount in foreign currency Min=0.00, Max=5985978.17, Mean=1324.90, StdDev=39643.10	628634	None
CrLocalAmt	double		Credited amount in local currency Min=0.00, Max=5985978.17, Mean=1324.90, StdDev=39643.10	628634	None

The Customers' Accounts table contains data which are relatively fixed in nature. This means that the majority of the attributes holds information about customers, which are not supposed to be changed through time. Beyond the description in Table 4.2, the properties of each attribute in the Customers' Accounts table are elaborated in the next paragraphs.

1. The "AcctNo" attribute, which is the combination of alphabets and numerals, is used to identify each customer uniquely. It is due to the values in this attribute that the relationship between each customer with its transaction in the Customers' Transaction table is established. For this reason, this attribute is significantly important for the experimentation.
2. The "AcctType" attribute shows the type of saving account to which the customer belongs. As stated earlier there are four different types of saving account types included in the table. In fact, only the private sector saving accounts ("S01") are the focus of this research and the rest account types ("S02", "S03", and "S05") are not going to be considered for further analysis. Once the "S01" values are selected, the "AcctType" attribute has no significant importance and going to be deleted.
3. The "Comp/Ind" attribute states whether the customer is an individual "I" or a company "C". Since the focus of the research to predict the churn of individual private sector saving accounts, the attribute is no more be important once the individual customers are selected.

4. The "CurrCode" attribute stated the type of currency used by the customer. As shown in Table 4.2 all the values are "ETB" so the attribute is not necessary and is going to be deleted as it no more differentiate among multiple currencies.
5. The "AcctStatus" attribute, which specifies the current status of a given customer account, is a very important one for the churn prediction experimentation. Though there are about seven statuses given as a value of this attribute, only the two values "ACT" (active accounts) and "CLS" (churned customers or closed accounts) are to be used for comparison in the prediction process. In fact, this is the attribute to be used as a class attribute in the prediction process of the experimentation.
6. The "DateAcctOpened" attribute specifies the date at which the account was created or the customer has subscribed at the bank. This attribute is significantly important and is going to be used for further analysis as the duration of each customer with the bank is going to be derived from these dates.
7. The "DateAcctClosed" attribute, as shown in the table has date values only for 394 customers and the rest 12777 values are missing. This is because; the attribute stands for those customers who already quit transacting with the bank (churned customers). All the 394 values stated in this attribute are repeated in the "DateOfStatus" attribute of corresponding customers. As a result, this attribute is not important at all and it is only a redundant attribute.
8. The "DateOfStatus" attribute states the date at which the customers' status and current balance is updated. Unfortunately, 9460 values are missing (not supplied by CBE) in this attribute and there are only 3711 values. In fact, there is a way of substituting the missing values from the date of the last transaction of each customer in the Customers' Transaction table. As the case of the "DateAcctOpened" attribute, the importance of this attribute in order to derive the customer's duration is very significant and is used for further analysis.
9. The "CurrentBalance" attribute, which states the balance of each customer at the date of status, has no importance for further analysis. This is because; using the balance of customers whom the date of status is unknown is meaningless.

In the same way, the attributes of the Customers' Transaction table, which is shown in Table 4.3 is elaborated as follows:

1. The "AcctNo" attribute has similar meaning with the corresponding attribute in the Customers' Account table. But for this attribute, there are many repetitions of similar account numbers unlike the case of the Customers' Account table. This is because, a customer can transact with

the bank many times in a certain period of time. The major use of this attribute is to establish integrity with the Customers' Account table and it is going to be used for the next steps in the dataset preparation process.

2. The "AcctType" and "Comp/Ind" attributes are just having similar meaning and function with their corresponding attributes in the Customers' Account table.
3. The "DateOfTxn" attribute, which states the dates at which the customer transact with the bank, is very important for the dataset preparation. Its major importance is to find out the number of transaction had in some selected months.
4. The "DrCrFlag" attribute is a very important attribute as it shows what kind of transaction a customer had (either debit or credit) in each transaction. This attribute is going to be used for further analysis to count the amount of debited and credited transactions of each customer within a given period of time.
5. The "OperCode" attribute specifies the kind of operation performed in each debited or credited transaction. There are seventeen types of operation codes (as stated in Table) and each operation belongs to either debit or credit transaction. The operations "CCO", "DEB", "IAD", "IBS", "ITX", "TSD", "TEL" and "WDL" are belongs to debited transaction and the rest operations "CRE", "DEP", "IAC", "ICC", "LDD", "MRC", "OBL", "OPC", and "TRC" are belongs to credited transaction. For the purpose of the experimentation only selected operation types are going to be considered. These are "DEB", "IBS", "TSD", and "WDL" from the debited transaction operation types and "CRE", "DEP", "OPC", and "TRC" from the credited transaction. The importance of this attribute continues until the specified operation types are selected and the rest are discarded.
6. The attributes "DrLocalAmt" and "CrLocalAmt" specifies the amount that is either debited or credited respectively in each transaction. These attributes are important to derive the monthly average debited or credited amount of each customer in the final dataset; so they are going to be used for further analysis.
7. The attributes "DrForeignAmt" and "CrForeignAmt" are not important due to the fact that all transactions are in local currency and the same figure is stated in their corresponding local amounts stated in No. 6 above.

The dataset is going to be prepared by analyzing the two tables independently and together, and merging the two tables and finally performing further analysis steps. Many of the attributes exist in the

current table are not directly the part of the final dataset. In other words, the attributes in the final dataset are the outcomes of the existing attributes through multiple analytical steps.

The Customers' Transaction table contains the transactional history of those 13172 customers listed in the Customers' Account table. Unfortunately the transaction of 1178 customers is not included. The reason for this might be those customers didn't transact with the bank during the period for which the transaction data is selected.

As a matter of the fact that a single customer can transact with the bank many times within a certain period of time, the number of data in the Customers' Transaction table (628634) is by far bigger than the number of data in the Customers' Accounts table (13172). As a result, the data exploration process for the Customers' Transaction table is done by using MS-EXCEL and MS-ACCESS tools. WEKA couldn't handle such a big amount of data with the hardware resources available.

4.2.2 Verification of Data Quality

Issues that can generate a data quality problem and found in the two tables are presented as follows:

- Out of the 415 closed accounts, 394 values are supplied in the "DateAcctClosed" attribute of the Customers' Accounts table. The dates at which the accounts closed for the rest 21 customers is missing.
- 9460 values are missing in the "DateOfStatus" attribute of the Customers' Accounts and only 3711 values are filled with appropriate dates of status. In fact, the "DateAcctClosed" which is stated in (1) above is part of the "DateOfStatus" (i.e. all the 394 values exist as a date of status in the "DateOfStatus" attribute). So the solution to handle the problem of the missing values is that to use the last date of transaction of each customer from the Customers' Transaction table. This is an appropriate way because the date of status is nearly the same as the date at which the customer make the last transaction.
- In the "DrCrFlag" attribute of the Customers' Transaction table is mistakenly filled as a debit ("D") for 13085 corresponding "IAC" values in the "OperCode" attribute while it is a known fact that "IAC" is an operation of credited transaction. This can simply be depicted from the amount credited for the same transactions in the same table. This problem can simply be handled by replacing the error D's by C's in the "DrCrFlag" attribute.
- The other problem is that the transaction of 1178 customers is not available in the Customers' Transactions table. The solution for this is just to exclude those customers from the Customers' Accounts table.

4.3 DATA PREPARATION

4.3.1 Rationale for Data Selection

As per the data obtained from CBE for this DM project, the final dataset to be prepared at the end of this section looks like to contain the following information.

- Three months of transactions of customers aggregated in a monthly basis needs to be prepared. For the length of time to be considered for churn prediction, (Berry & Linoff, 2004)stated that binary outcome churn models usually have a fairly short time horizon such as 60 or 90 days. In addition, (Madhavi, 2012) opines that in bank's context the last three months of transactions are adequate for the purpose of churn prediction. But assuming that there will be other churn management steps to be undertaken once after the churners are predicted; the last month, at which the account status is taken, is not going to be considered. In the same way, the transaction in a month at which the customer subscribed is not going to be considered for the reason of avoiding incomplete monthly transactions. So for the selected three months of transactions each customer:
  - ✓ The number of debit transactions of each month,
  - ✓ The number of credit transactions of each month,
  - ✓ The average debited amount of each month,
  - ✓ The average credited amount of each month
  - ✓ The number of debit transactions in the three months
  - ✓ The number of credit transactions in the three months
  - ✓ The sum of debited amount in the three months and
  - ✓ The sum of credited amount in the three months are to be prepared as an attribute
- In addition to the above, few attributes are also going to be included. These are:
  - ✓ The number of days that each customer stays with the bank (duration), and
  - ✓ The Account Status (whether ACTIVE or CHURN) as a class attribute.

Most of the attributes listed above does not exist directly in the original data which are shown in Table 4.2 and Table 4.3, rather they are to be generated through the successive data preparation processes.

Some part of the data has no contribution for the purpose of the experimentation. For this reason a data selection is required to create a meaningful dataset for modeling. The domains which are to be excluded from both tables are presented as follows.

4.3.1.1 In the Customers' Accounts Table:

Since the objective of this DM project is to predict the individual private saving account customers' churn behavior the following data needs to be eliminated first.

1. In the "AcctType" attribute values "S02", "S03", & "S05" and from the "Comp/Ind" attribute values with "C" are not to be considered because these are out of the domain of the study.
2. Those customers having "AcctStatus" values other than "ACT" and "CLS" are to be deleted. This is because for the churn prediction experimentation it is needed only to compare the behaviour of active and churned customers. The others, for instance "BLK" or blocked accounts are stay without transaction for a defined period of time with a reason. So, considering such domains might mislead the prediction result.
3. Customers having less than three months of duration (excluding the first and the last month) as a customer with the bank are not going to be considered. This is because; the experimentation is based on three months of transactional data of customers. In fact, this step is performed after the missing value in the "DateOfStatus" is filled and the duration of each customer is known.
4. Those customers who have got no transaction data in the Customers' Transaction table are also to be excluded.

Once after the above domains of data are eliminated from the Customers' Accounts table the following attributes needs to be deleted:

- ✓ The attributes "AcctType", "Comp/Ind" and "CurrCode" are no more be useful because all holds a single-type value; "S01" for "AcctType", "I" for "Comp/Ind" and "ETB" for "CurrCode" . So they need to be deleted.
- ✓ The attribute "DateAcctClosed" is not important and should be deleted because all the 394 values it holds are repeated in the "DateOfStatus" attribute.
- ✓ The "CurrentBalance" attributes cannot be used as the "DateOfStatus" of 9460 customers is unknown. So the values are representing the balance of customers for unknown date of status.
- ✓ The attributes "DateAcctOpened" and "DateOfStatus" are important until the duration of each customer with the bank is calculated and the last three months of transactions of each customer is identified using the values in these attributes. Then after, the two attributes are no more be important for the upcoming steps and need to be excluded.

4.3.1.2 In the Customers' Transactions table

Here also, there is a domain of data not to be included. In fact, some attributes with no significant importance for the experiment are also deleted.

1. The transaction data of all customers excluded for different reasons from the Customers' Accounts table should also be excluded from this table.
2. Those transactions having values other than "CRE", DEB", "DEP", IBS", "OPC", "TRC", "TSD" and "WDL" in the "OperCode" attribute are not going to be used and should be deleted. The reason for this is that most of them are not from the direct transactional involvement of customers. For instance, "ITX" or income tax, "IAC" & "IAD" or interests, "CCO" or commissions, "TEL" or telephone charges are derived from the monthly accounting activities of the bank or from other transactions of customers. So considering these values might exaggerate the number of transactions of a customer for which he/she has not been involved.
3. The attribute "AcctType" "CurrCode" and "Comp/Ind" are to be deleted for the reason specified earlier in Customers' Accounts table. The attributes "DrForeignAmt" and "CrForeignAmt" are also have no use as all the transactions are performed in local currency and the same amounts are specified in "DrLocalAmt" and "CrLocalAmt" respectively.

4.3.2 Data Cleaning

Data cleaning is done for those values having data quality problem (as stated in section4.2.2). This means the missed date of status values are filled with the last date at which the customers transact with the bank. And this will resolve the problem because even most of the filled dates of status values are the same or differ by two or three days from the last date of transaction of those customers. The other data quality problem is simple to resolve and fortunately that part of the data is not going to be considered for further analysis (as stated in section 4.3.1)

4.3.3 Data Construction Process

At this stage, the important domain of the data and the attributes are selected in a logical manner and the data cleaning process is also done. Here a few attributes and a table are going to be constructed based on the existing data in the two tables.

In the Customers' Accounts table four derived attributes are needed. These are:

- "CustDuration" attribute, which is derived from the updated "DateOfStatus" attribute and the "DateAcctOpened" attribute just by arithmetically determining the number of days between these date values. The need for deriving the "CustDuration" attribute is for two reasons.
  - ✓ The first reason is to identify the customers having less than three months of time as a customer with the bank and to delete them. Since, the transaction of at least three full months from each customer is needed for the experiment, such customers are no more be important.
  - ✓ The second reason is that the "CustDuration" attribute itself is needed as part of the dataset.
- The other three attributes are "Mon3", "Mon2" and "Mon1" which hold the last three months of transaction for each customer. One of the difficult parts of the data preparation process is that the last three months of transaction is not the same period for all customers. If one customer closes his account in June 2013, the last three months of transactions for this customer is going to be March 2013 up to May 2013 but a customer who churned in January 2013 the last three months are going to be from October 2012 to December 2012. These attributes are derived from the "DateOfStatus" attribute just by taking the three consecutive months just before the date of status of each customer.

Again, in the Customers' Transaction table two attributes are derived from the "DrCrFlag" attribute. This is because counting of the number of debited and credited transaction in a given month is necessary for the preparation of the final dataset. The two derived attributes are "DrTxn" and "CrTxn". The value of "DrTxn" is 1 whenever the corresponding value in "DrCrFlag" is "D" and 0 for "C". The "CrTxn" attribute has got the same purpose but its values are 1 for credit transactions and 0 for debit.

A new table is also derived for the purpose of simplifying the querying process while merging the Customers' Account and the Customers' Transaction tables in MS-ACCESS. This table, namely "MonthsOfTxns" is derived from the Customers' Accounts table and has two attributes. The first attribute is the "AcctNo" attribute which exists in both tables. The second attribute is "MonthsOfTxns" which is derived from the attributes "Mon3", "Mon2" and "Mon1" and holds the three selected months of transactions of each customer in a column. Unlike the case of the Customers' Accounts table, each value in "AcctNo" attribute is repeated three times for the corresponding last three months of transactions.

After these attributes and tables are derived, the “Mon3”, “Mon2”, “Mon1”, “DateAcctOpened” and “DateOfStatus” attributes in the Customers’ Accounts table and the “DrCrFlag” attribute in the Customers’ Transaction table are deleted for the reason stated earlier.

After the data construction process is completed, the data and also the number of attributes in the tables are decreased and only the relevant data is left. The attributes left in the three tables together with the description and number of data is shown in the following tables.

Table 4.4: Data left in Customers' Accounts table after data construction

Attribute	Type	Value	Description & Basic Stat	Count
AcctNo	Alpha-numeric		To Identify each Customer Uniquely	6045
AcctStatus	Text	ACTIVE	ACTIVE	5720
		CHURN	INACTIVE	325
CustDuration	Integers		The number of days of each customer with the bank Min=96 , Max=2002 , Mean=431 , StdDev= 388.2,	6045

Table 4.5: Data left in Months of Transaction table after data construction

Attribute	Type	Value	Description & Basic Stat	Count
AcctNo	Alpha-numeric		To Identify each Customer Uniquely	18135
MonOfTxns	Date	mm/dd/yy	To give the last three months of transactions of each customer	18135

Table 4.6: Data in Customers' Transaction table after data construction

Attribute	Type	Value	Description & Basic Stat	Count
AcctNo	Alpha-numeric		To Identify each Customer Uniquely	139364
DateOfTxn	Date		The date at which the customer transact	139364
DrTxn	Integer	0 and 1	1 for debit transaction and 0 for credit	139364
CrTxn	Integer	0 and 1	1 for debit transaction and 0 for credit	139364
DrAmt	Double		The debited amount	139364
CrAmt	Double		The Credited amount	139364

4.3.4 Data Integration

Here, the two tables (Customers' Accounts and Customers' Transactions) are merged through using the Months' of Transaction table as intermediary. This process is done by importing the three MS-EXCEL tables into MS-ACCESS for ease of querying purpose. The query is shown in Figure below. Once the tables are merged and become a single table, it is exported back as MS-EXCEL document for further processing until the final dataset is produced.

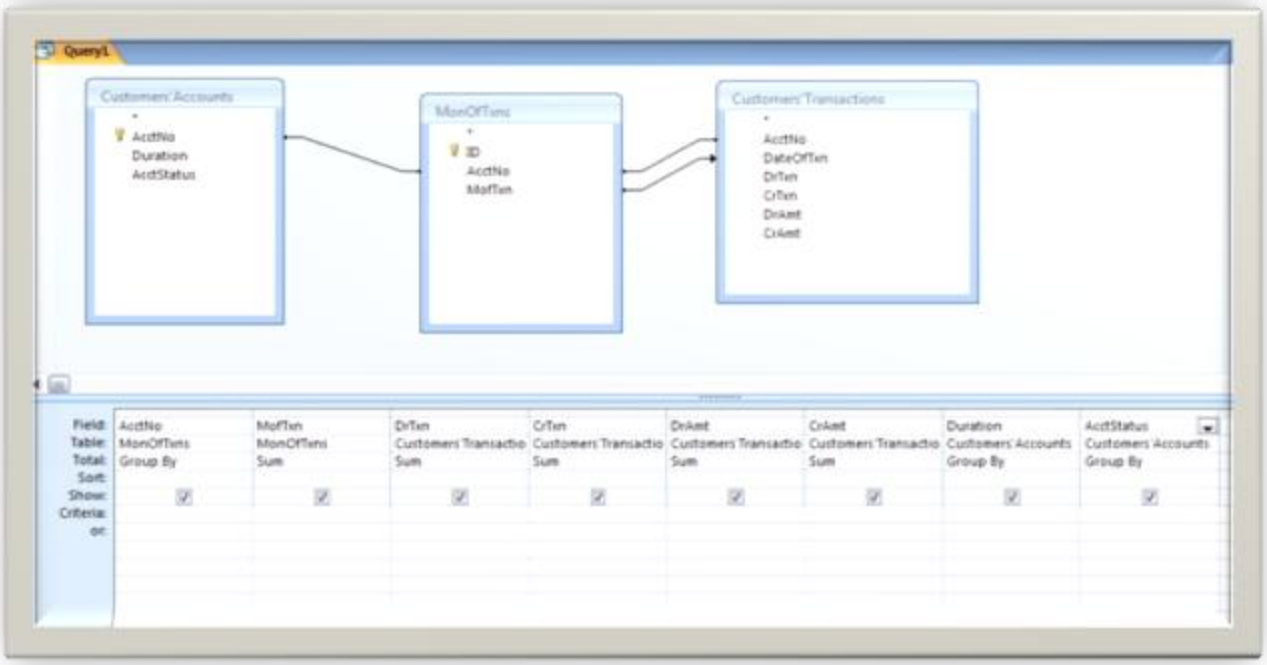


Figure 4.1: Data Integration in MS-ACCESS

4.3.5 Data Formatting

The table which is obtained by merging the two tables through MS-ACCESS tool contains attributes such as:

- The account numbers,
- The customer's duration
- The customer's status (ACTIVE/CHURN)
- For the selected last three months of each customer:
  - ✓ The numbers of both debited and credited transactions in each month
  - ✓ The amounts that are both debited and credited in each month.

But in the final dataset, it has been intended to include the monthly average debited/credited amounts (not simply the amounts) and also the sum of the number of all debited transactions and credited transactions in the three months, and the sum of all the debited amounts and the credited amounts in the three months. For this reason, several arithmetical operations have been performed in MS-EXCEL and the final dataset is produced. The description of each attribute in the final dataset is shown in Table 4.7. In addition, some algorithms expect the attributes to be arranged in such a way that the most important ones to become first, the less significant attributes to become next, and the dependent attribute (the attribute to be used as a class) to become last. For that reason, the attributes are arranged from the third month ("M3") to the earlier month ("M1") as shown in Table 4.7

Table 4.7: The Structure of the final Dataset

Attribute	Type/Value	Description
AcctNo	Alpha-numeric	To identify each Customer uniquely
DrTxnM3	integer	The number of debited transactions in month 3
AvgDrAmtM3	double	The average debited amount in month 3
CrTxnM3	integer	The number of credited transactions in month 3
AvgCrAmtM3	double	The average credited amount in month 3
DrTxnM2	integer	The number of debited transactions in month 2
AvgDrAmtM2	double	The average debited amount in month 2
CrTxnM2	integer	The number of credited transactions in month 2
AvgCrAmtM2	double	The average credited amount in month 2
DrTxnM1	integer	The number of debited transactions in month 1
AvgDrAmtM1	double	The average debited amount in month 1
CrTxnM1	integer	The number of credited transactions in month 1
AvgCrAmtM1	double	The average credited amount in month 1
TotDrTxn	integer	The total number of debited transactions in 3 months
TotDrAmt	double	The total debited amount in three months
TotCrTxn	integer	The total number of credited transactions in 3 months
TotCrAmt	double	The total credited amount in three months
CustDuration	integer	The duration of the customer with the bank
AcctStatus	ACTIVE	Active customer accounts
	CHURN	Closed customer accounts

Here the values of the account status are changed to "ACTIVE" and "CHURN" instead of "ACT" and "CLS" respectively just for clarity purpose. Even though the recent WEKA versions include a package called "WEKAExcel" which can directly open and Explore MS-EXCEL data as it is, the researcher preferred to prepare the most appropriate data format for WEKA that is ARFF file. This is because

some incompatibility issues are observed while directly opening the MS-EXCEL file in WEKA. The number of instances is observed to be reduced by one. So, the dataset is first converted to CSV format, and then opened by WEKA and saved as ARFF format. The sample of the dataset in ARFF format is shown in Figure 4.2

```
@relation FinalDS-weka.filters.unsupervised.attribute.Remove-R1

@attribute DrTxnM3 numeric
@attribute AvgDrAmtM3 numeric
@attribute CrTxnM3 numeric
@attribute AvgCrAmtM3 numeric
@attribute DrTxnM2 numeric
@attribute AvgDrAmtM2 numeric
@attribute CrTxnM2 numeric
@attribute AvgCrAmtM2 numeric
@attribute DrTxnM1 numeric
@attribute AvgDrAmtM1 numeric
@attribute CrTxnM1 numeric
@attribute AvgCrAmtM1 numeric
@attribute TotDrTxn numeric
@attribute TotDrAmt numeric
@attribute TotCrTxn numeric
@attribute TotCrAmt numeric
@attribute CustDuration numeric
@attribute AcctStatus {CHURN,ACTIVE}

@data
2,1300,0,0,5,614.6,0,0,1,300,0,0,8,5973,0,0,155,CHURN
0,0,0,0,1,1700,1,1717.5,2,915,1,1717.5,3,3530,2,3435,858,ACTIVE
```

Figure 4.2The ARFF format of the Final Dataset

4.3.6 Attribute Selection

At this point the dataset has 6045 instances and 19 attributes. The majority of the attributes are equally important for the experimentation purpose except the “AcctNo” attribute, which has been significantly important for identifying each customer uniquely. This attribute is no more important for the modeling purpose. Hence, it is removed by using the “Remove” functionality in WEKA explorer.

CHAPTER FIVE

5 Experimentation

5.1 Selection of Modelling Techniques and Tools

According to (Gibert, Sànchez-Marrè, & Codina, 2010),the main goal of the problem to be solved and the structure of the available data are the main parameters taken into account by humans to choose the proper DM technique in a real application. Three modeling techniques are going to be used for the experimentation of the DM project. These are J48, LR and Bagging. Several factors have been taken into consideration while selecting these modeling techniques. Some of the factors are:

1. As the problem is a classification DM problem, the modelling techniques are selected from the classification modelling techniques,
2. The modelling techniques J48 and LR are opined by various authors and researchers as the most widely used and relatively with a better success rate in predicting churn in different industries. And the Bagging technique is chosen to check the impact of combinations of classifiers in predicting churners.
3. The specific requirements of each model have been observed and ascertained that all the selected techniques can adequately support nominal classes and numeric attributes (which is the case of the existing dataset property)
4. In addition, the ease of understanding the models and presentation capabilities are also taken into account.

In selecting the tool to be used for modeling, the factors taken into consideration are:

1. The cost of the tool: - Since no budget is allotted in this research for the procurement of DM tool, an open source (free tool) is needed.
2. Applicability: - The ability of the tool to run the modelling techniques (algorithms)
3. Performance improvement: - The ability of the tool in improving its performance in different situations.

WEKA is one of the open source DM tools having several functionalities. In their study of DM tools comparison, (Wahbeh, Al-radaideh, Al-kabi, & Al-shawakfa, 2011) stated that WEKA toolkit:

- Has achieved the highest applicability.

- Achieved the highest improvements (when moving from the Percentage Split test mode to the Cross Validation test mode)
- Is the best tool in terms of the ability to run the selected classifiers, and
- Can better handle the problem of multiclass data sets

For these reasons, WEKA is selected as a modeling tool in this study

5.2 Test Design

In this section, it is discussed how the samples are prepared for modelling, how the predicting accuracy of each modelling is evaluated and the major tasks to be conducted at each experiment for the selected algorithms.

5.2.1 Preparing samples from dataset

In classification modelling, models are built using a training set for learning purpose and the predicting performance of models is tested using the corresponding test sets. In order to check whether the predicting performance of models are affected by the percentages of partitioning the dataset into training and test sets, the application of SMOTE to oversample the minority class in the training set, and the usage of all or selected attributes in the training set; the following procedure is followed in preparing different set of samples:

1. The dataset is systematically partitioned in to three pairs of disjoint sets of training sets and test sets (for validating the reliability of the models):
    a. 66% Training set and 34% Test set,
    b. 70% Training set and 30% Test set, and
    c. 75% Training set and 25% Test set
2. As the CHURN class is a minority class (which accounts only 5.38% of the total dataset) and due to the fact that the predicting performance of some modelling techniques is affected by class imbalance, another set of training and testing sets are prepared by applying SMOTE on the training sets of the aforementioned partitioned sets (stated in a, b, and c). SMOTE is applied to oversample the minority class (CHURN class) and consequently the number of instances in both classes becomes equivalent (almost 50% each)
3. WEKA has an attribute selection facility, which suggests the most relevant attributes for predicting the classes. The selected attributes on each training set before and after SMOTE has been applied are different. For all the training sets (66%, 70%, and 75%), the attributes selected as

relevant for predicting the class attribute are: DrTxnM3, CrTxnM3, CrTxnM2, CrTxnM1, TotCrTxn, and CustDuration. After SMOTE has been applied on all the training sets (66%, 70%, and 75%), the attributes selected as relevant for predicting the class attribute are: DrTxnM3, DrTxnM2, DrTxnM1, CrTxnM3, CrTxnM2, CrTxnM1, and CustDuration. So another set of training and tests are prepared with the corresponding selected attributes the samples prepared on (1) and (2).

5.2.2 Testing and Evaluation Criteria

As the main objective of this study is to predict customers at CBE with higher possibility of churning (closing their accounts), the predicting performance of each model should be judged as to how well it can predict the CHURN class. Different suggestions are provided by authors as to how a churn prediction model can be evaluated. According to (Burez & Poel, 2009), Precision, Recall, and Accuracy(or MER) are often used to measure the classification quality of binary classifier. In addition, (Lazarov & Capota, 2007) used sensitivity, specifity and accuracy to measure the quality of churn prediction models.

In each step of modeling process, the classification quality or the performance of a model in predicting the CHURN class is validated by the separate test sets. As the DM success criteria is stated as predicting at least 90% of the churners, the recall of each model can be taken to check whether each model fulfils the success criteria. In addition, the F-Measure, which is the harmonic mean of the Recall and Precision, is used to select the best model. As both recall and precision are important measures of predicting performance, their harmonic mean (F-Measure) is used to measure the performance of the models.

5.2.3 Modelling Procedure

Three experiments are conducted for building several models by applying each of the three selected modeling techniques. The rationale for selecting three experiments for each technique is to address the model building for the default value, for different number of folds of cross validation, and parameter setting procedures. The major tasks included in each of three experiments are stated as follows:

Experiment 1: - In the first experiment of each modelling technique, learning models are built using different training sets for the default cross validation (K=10). As stated in Section 5.2.1, the three training sets (66%, 70%, and 75%) are used in four different ways for each. These are:

1) When all attributes are used (All_Attr),
2) When the selected attributes are used (Sel_Attr),
3) Oversampled by SMOTE and all attributes are used (OvS_All_Attr)

4) Oversampled by SMOTE and selected attributes are used (OvS_Sel_Attr)

The Recall and Precision of each model is recorded and the model with better performance is selected and assessed using the separate test set.

Experiment 2: - Models are built for different number of folds (K=5, 15, 20, 25, 30, and 35) of cross validation test option using all the different training sets stated in Experiment 1. For each case the number of folds applied (K), the Recall and Precision of the better models is recorded. The best of these entire models is selected and assessed using the separate test set.

Experiment 3: - Considering the model, which is selected as the best in Experiment 2, attempts to enhance the predicting performance of the model is done by changing the values of the different parameters of the algorithm and also the number of folds (K) of cross validation test option. After all the possible attempts has been done, the model with better result is considered as the best model of the specific modelling technique being used and assessed on the corresponding test set

5.3 Modelling

In this step models are built by the three different techniques selected in Section 5.1 according to the procedure stated in the test design section.

5.3.1 Modelling Using the J48 Decision Tree modelling Technique

5.3.1.1 Experiment 1

Twelve models are built for different training set samples applying the default cross validation test option (K=10). The recall and precision of each model are shown in Table 5.1.

Table 5.1 Outputs of the J48 models for the default Cross Validation (K=10)

THE WAY TS USED	USING 66% TS		USING 70% TS		USING 75% TS	
	RECALL	PREC.	RECALL	PREC.	RECALL	PREC.
All_Attr.	0.870	0.926	0.894	0.906	0.905	0.928
Sel_Attr	0.870	0.949	0.868	0.947	0.881	0.918
OvS_All_Attr	**0.987**	**0.990**	0.990	0.990	0.992	0.991
OvS_Sel_Attr	0.988	0.989	**0.990**	**0.994**	**0.991**	**0.994**

As shown in Table 5.1, better results are observed in all the three cases when the training sets are oversampled by SMOTE. It seems from the results that the model with best performance is obtained when the 75% oversampled training set is used with the selected attributes, But the model built using the 70% oversampled training set with the selected attributes outperforms the others while tested with

its corresponding test set. In this model the number of correctly classified instances is 7945 (99.226%) and the number of incorrectly classified instances is 62 (0.774%). Considering the CHURN class, the Recall, Precision, and F-Measure of this model are 0.990, 0.994, and 0.992 respectively. The confusion matrix of this model is summarized in Table 5.2

Table 5.2 Confusion matrix of the better J48 learning model when K=10

ACTUAL	PREDICTED		TOTAL	CORRECTLY CLASSIFIED
	CHURN	ACTIVE		
CHURN	3964	39	4003	99.026%
ACTIVE	23	3981	4004	99.426%

This learning model is assessed on the corresponding 30% test set with selected attributes. While tested the number of correctly classified instances is 1804 (99.449%) and the number of incorrectly classified instances is 10 (0.551%). Considering the CHURN class, the Recall, Precision, and F-Measure of this model are 0.939, 0.958, and 0.948 respectively. The confusion matrix is summarized in Table 5.3

Table 5.3 Confusion matrix of the testing result of J48 model for K=10

ACTUAL	PREDICTED		TOTAL	CORRECTLY CLASSIFIED
	CHURN	ACTIVE		
CHURN	92	6	98	93.878%
ACTIVE	4	1712	1716	99.767%

5.3.1.2 Experiment 2

By changing the number of folds (K=5, 15, 20, 25, 30, and 35) of cross validation test option, several models are built for different training set samples. The number of folds (K), the recall and precision of the models with relatively better performances are shown in Table 5.4

Table 5.4 Better results of J48 model for K=5, 15, 20, 25, 30, and 35

THE WAY TS USED	USING 66% TS			USING 70% TS			USING 75% TS		
	K	RECALL	PREC.	K	RECALL	PREC.	K	RECALL	PREC.
All_Attr.	15	0.875	0.945	25	0.903	0.899	5	0.909	0.929
Sel_Attr	30	0.880	0.955	20	0.881	0.939	20	0.893	0.927
OvS_All_Attr	20/30	0.990	0.991	25	0.993	0.991	**20**	**0.993**	**0.993**
OvS_Sel_Attr	**20**	**0.992**	**0.989**	**25**	**0.991**	**0.995**	30	0.991	0.995

As shown inTable 5.4 better results are observed in all the three cases when the training sets are oversampled by SMOTE. Again here, unlike the results shown in the table the model built using the 70% oversampled training set with the selected attributes and K=20, outperforms the others while tested with its corresponding test set. In this model the number of correctly classified instances is 7950(99.288%) and the number of incorrectly classified instances is 57 (0.712%). Considering the CHURN class, the Recall, Precision, and F-Measure of this model are 0.991, 0.995, and 0.993 respectively. The confusion matrix of this model is summarized in Table 5.5

Table 5.5 Confusion matrix of the better J48 learning model when K=20

ACTUAL	PREDICTED		TOTAL	CORRECTLY CLASSIFIED
	CHURN	ACTIVE		
CHURN	3967	36	4003	99.101%
ACTIVE	21	3983	4004	99.476%

This learning model is assessed on the corresponding 30% test set with selected attributes. While tested the number of correctly classified instances is 1804 (99.449%) and the number of incorrectly classified instances is 10 (0.551%). Considering the CHURN class, the Recall, Precision, and F-Measure of this model are 0.939, 0.958, and 0.948 respectively. The confusion matrix is summarized in Table 5.6

Table 5.6 Confusion matrix of the testing result of J48 model with K=20

ACTUAL	PREDICTED		TOTAL	CORRECTLY CLASSIFIED
	CHURN	ACTIVE		
CHURN	92	6	98	93.878%
ACTIVE	4	1712	1716	99.767%

As it can be seen from the testing results of Experiment 1 and Experiment 2, the predicting performance when K=10 and K=20 is the same. While conducting Experiment 3, either of the two best models in Experiment 1 and Experiment 2 can be used.

5.3.1.3 Experiment 3

Considering the better resulting models in the previous two experiments, the values of the J48 modeling technique is changed so that a learning model with better predicting performance can be obtained. But no positive change in predicting performance is observed by changing all the values. One of the two better learning models in the previous experiments can be taken as the best J48 model. Since the result in Experiment 1 is with less number of folds, it is taken as the best J48 model.

5.3.2 Modelling Using the LR Algorithm

5.3.2.1 Experiment 1

Here also, twelve models are built for different training set samples applying the default cross validation test option (K=10). The recall and precision of each model are shown in Table 5.7

Table 5.7 Outputs of the LR models for the default Cross Validation (K=10)

THE WAY TS USED	USING 66% TS		USING 70% TS		USING 75% TS	
	RECALL	PREC.	RECALL	PREC.	RECALL	PREC.
All_Attr.	0.907	0.942	0.912	0.928	0.905	0.940
Sel_Attr	0.875	0.955	0.881	0.962	0.872	0.955
OvS_All_Attr	**0.976**	**0.963**	**0.980**	**0.966**	**0.976**	**0.969**
OvS_Sel_Attr	0.962	0.959	0.962	0.962	0.956	0.964

As shown in Table 5.7, in all the three training sets, better results are achieved when the training sets are oversampled by SMOTE and all the attributes are used (rather than the suggested attributes). Though the better predicting accuracy is observed in the 70% training set, the model built using the 66% oversampled training set using althea attributes outperforms the others while being tested on its corresponding test set. In this model the number of correctly classified instances is 7391(96.931%) and the number of incorrectly classified instances is 234 (3.069%). Considering the CHURN class, the Recall, Precision, and F-Measure of this model are 0.976, 0.963, and 0.969 respectively. The confusion matrix of this model is summarized in Table 5.8

Table 5.8 Confusion matrix of the better LR learning model when K=10

ACTUAL	PREDICTED		TOTAL	CORRECTLY CLASSIFIED
	CHURN	ACTIVE		
CHURN	3719	93	3812	97.560%
ACTIVE	141	3672	3813	96.302%

This learning model is assessed on the corresponding 34% test set with all the attributes. While the number of correctly classified instances is 1952 (96.825%) and the number of incorrectly instances is 64 (3.175%). Considering the CHURN class, the Recall, Precision, and F-Measure of model are 0.963, 0.636, and 0.766 respectively. The confusion matrix is summarized in

Table 5.9

Table 5.9 Confusion matrix of the testing result of LR model for K=10

ACTUAL	PREDICTED		TOTAL	CORRECTLY CLASSIFIED
	CHURN	ACTIVE		
CHURN	105	4	109	96.330%
ACTIVE	60	1847	1907	96.854%

5.3.2.2 Experiment 2

By changing the number of folds (K=5, 15, 20, 25, 30, and 35) of cross validation test option, several models are built for different training set samples. The number of folds (K), the recall and precision of the models with relatively better performances are shown in Table 5.10

Table 5.10 Better results of LR model for K=5, 15, 20, 25, 30, and 35

THE WAY TS USED	USING 66% TS			USING 70% TS			USING 75% TS		
	K	RECALL	PREC.	K	RECALL	PREC.	K	RECALL	PREC.
All_Attr.	15	0.921	0.909	20	0.916	0.929	5	0.909	0.936
Sel_Attr	15	0.875	0.950	20	0.881	0.957	20	0.872	0.955
OvS_All_Attr	**20**	**0.975**	**0.963**	**20**	**0.980**	**0.966**	**5**	**0.977**	**0.969**
OvS_Sel_Attr	25	0.961	0.959	30	0.962	0.962	15	0.956	0.964

As shown in Table 5.10better results are observed in all the three cases when the training sets are oversampled by SMOTE. Again here, unlike the results shown in the table the, model built using the 66% oversampled training set using all the attributes and K=20, outperforms the others while tested with its corresponding test set. In this model the number of correctly classified instances is 7385(96.853%) and the number of incorrectly classified instances is 240 (3.148%). Considering the CHURN class, the Recall, Precision, and F-Measure of this model are 0.975, 0.963, and 0.969 respectively. The confusion matrix of this model is summarized in Table 5.11

Table 5.11 Confusion matrix of the better LR learning model when K=20

ACTUAL	PREDICTED		TOTAL	CORRECTLY CLASSIFIED
	CHURN	ACTIVE		
CHURN	3715	97	3812	97.456%
ACTIVE	143	3670	3813	96.250%

This learning model is assessed on the corresponding 34% test set with all the attributes. While tested the number of correctly classified instances is 1952 (96.825%) and the number of incorrectly classified

instances is 64 (3.175%). Considering the CHURN class, the Recall, Precision, and F-Measure of this model are 0.963, 0.636, and 0.766 respectively. The confusion matrix is summarized in Table 5.12

Table 5.12 Confusion matrix of the testing result of LR model for K=20

ACTUAL	PREDICTED		TOTAL	CORRECTLY CLASSIFIED
	CHURN	ACTIVE		
CHURN	105	4	109	96.330%
ACTIVE	60	1847	1907	96.854%

The predicting performances of the better models in Experiment 1 and 2 are the same while tested on the corresponding test sets. So both can be considered for the parameter setting process in Experiment3

5.3.2.3 Experiment 3

Here the values of the parameters in LR algorithm are changed and the improvement in the prediction accuracy of the better results of the first two experiments is checked. Only the ridge parameter shows a difference in the predicting accuracy while its value is changed. But no positive change is observed. It is therefore, the better learning model obtained in the first experiment is selected as the best LR model as it has relatively better in its recall and precision than the other.

5.3.3 Modelling Using the Bagging Algorithm

5.3.3.1 Experiment 1

With the same procedure, twelve models are built for different training set samples applying the default cross validation test option (K=10). The recall and precision of each model are shown in Table 5.13

Table 5.13 Outputs of the Bagging models for the default Cross Validation (K=10)

THE WAY TS USED	USING 66% TS		USING 70% TS		USING 75% TS	
	RECALL	PREC.	RECALL	PREC.	RECALL	PREC.
All_Attr.	0.875	0.955	0.885	0.944	0.905	0.965
Sel_Attr	0.884	0.965	0.881	0.957	0.885	0.947
OvS_All_Attr	0.990	0.992	0.990	0.990	**0.993**	**0.991**
OvS_Sel_Attr	**0.992**	**0.990**	**0.989**	**0.991**	0.990	0.993

In all the three training sets, better results are achieved when the training sets are oversampled by SMOTE. Unlike the case of the other two algorithms, the Bagging model built by using the 75%

oversampled training set with all attributes, which is having a better result, also outperforms the others while tested on the corresponding test set. In this model the number of correctly classified instances is 8509(99.184%) and the number of incorrectly classified instances is 70 (0.816%). Considering the CHURN class, the Recall, Precision, and F-Measure of this model are 0.993, 0.991, and 0.992 respectively. The confusion matrix of this model is summarized in Table 5.14

Table 5.14 Confusion matrix of the better Bagging learning model when K=10

ACTUAL	PREDICTED		TOTAL	CORRECTLY CLASSIFIED
	CHURN	ACTIVE		
CHURN	4258	31	4289	97.561%
ACTIVE	39	4251	4290	99.091%

This learning model is assessed on the corresponding 25% test set with all the attributes. While tested the number of correctly classified instances is 1502 (99.339%) and the number of incorrectly classified instances is 10 (0.661%). Considering the CHURN class, the Recall, Precision, and F-Measure of this model are 0.976, 0.909, and 0.941 respectively. The confusion matrix is summarized in Table 5.15

Table 5.15 Confusion matrix of the testing result of the Bagging model for K=10

ACTUAL	PREDICTED		TOTAL	CORRECTLY CLASSIFIED
	CHURN	ACTIVE		
CHURN	80	2	82	97.561%
ACTIVE	8	1422	1430	99.441%

5.3.3.2 Experiment 2

By changing the number of folds (K=5, 15, 20, 25, 30, and 35) of cross validation test option, several models are built for different training set samples. The number of folds (K), the recall and precision of the models with relatively better performances are shown in Table 5.16

Table 5.16 Better results of Bagging models for K=5, 15, 20, 25, 30, and 35

THE WAY TS USED	USING 66% TS			USING 70% TS			USING 75% TS		
	K	RECALL	PREC.	K	RECALL	PREC.	K	RECALL	PREC.
All_Attr.	25	0.889	0.965	**35**	**0.907**	**0.972**	15	0.905	0.965
Sel_Attr	20	0.884	0.960	15/30	0.885	0.962	15	0.905	0.948
OvS_All_Attr	30	0.991	0.992	15	0.991	0.991	**15**	**0.994**	**0.992**
OvS_Sel_Attr	**15**	**0.993**	**0.993**	25	0.990	0.994	20	0.991	0.994

In the two training sets, better results are achieved when the training sets are oversampled by SMOTE and in one training set better result is observed when SMOTE is not applied. The model built by using the 75% oversampled training set with all attributes and K=15, which is having a better result, also outperforms the others while tested on the corresponding test set. In this model the number of correctly classified instances is 8518(99.289%) and the number of incorrectly classified instances is 61 (0.711%). Considering the CHURN class, the Recall, Precision, and F-Measure of this model are 0.994, 0.992, and 0.993 respectively. The confusion matrix of this model is summarized in Table 5.17

Table 5.17 Confusion matrix of the better Bagging learning model when K=15

ACTUAL	PREDICTED		TOTAL	CORRECTLY CLASSIFIED
	CHURN	ACTIVE		
CHURN	4264	25	4289	99.417%
ACTIVE	36	4254	4290	99.161%

The model is assessed validated by the separate 25% test set. While tested the number of correctly classified instances is 1502 (99.339%) and the number of incorrectly classified instances is10 (0.661%). Considering the CHURN class, the Recall, Precision, and F-Measure of this model are 0.976, 0.909, and 0.941 respectively. The confusion matrix is summarized in Table 5.18

Table 5.18 Confusion matrix of the testing result of the Bagging model for K=15

ACTUAL	PREDICTED		TOTAL	CORRECTLY CLASSIFIED
	CHURN	ACTIVE		
CHURN	80	2	82	97.561%
ACTIVE	8	1422	1430	99.441%

The two better models selected in Experiment 1 and 2 show similar performances while tested on the corresponding test sets. But, the learning model built as a better model in the second experiment gives a better Recall and Precision as compared to the other. So, it is going to be considered in the next experimentation.

5.3.3.3 Experiment 3

Among the parameters in Bagging, only the classifier parameter (which assigns an algorithm as a base classifier) results in a better prediction performance.

Table 5.19 Parameter in bagging algorithm that can improve predicting performance

Parameter	Default Value	Description
Classifier	REPTree	-- The base classifier to be used.

When the RandomTree algorithm is selected as a base classifier of bagging and the number of folds of cross validation is set to 15, the number of correctly classified instances becomes 8539 (99.534 %) and the number of incorrectly classified instances becomes 40 (0.466 %). The confusion matrix of this model is summarized in Table 5.20

Table 5.20 The Bagging learning model after parameter setting for K=15

ACTUAL	PREDICTED		TOTAL	CORRECTLY CLASSIFIED
	CHURN	ACTIVE		
CHURN	4275	14	4289	99.674%
ACTIVE	26	4264	4290	99.394%

This learning model is assessed on the corresponding 25% test set using all the attributes. While tested the number of correctly classified instances is 1502 (99.339%) and the number of incorrectly classified instances is 10 (0.661%). Considering the CHURN class, the Recall, Precision, and F-Measure of this model are 0.939, 0.939, and 0.939 respectively. The confusion matrix is summarized in Table5.21

Table5.21 Confusion matrix of the testing result of the bagging model after parameter setting

ACTUAL	PREDICTED		TOTAL	CORRECTLY CLASSIFIED
	CHURN	ACTIVE		
CHURN	77	5	82	93.902%
ACTIVE	5	1425	1430	99.650%

Hence, this model is considered as the best bagging model.

5.3.4 Assessment of the models built by the three algorithms

As it is stated in the test design section of this chapter, the basic criteria for comparing the models are its recall should be greater than 90% (as a DM success criteria) and then the best model is selected by the highest value of F-Measure (to consider both precision and recall)

In order to simplify the comparison process, the predicting performance of the best of each of the three modeling techniques on their learning model and while re-evaluated on their corresponding test sets is summarized in the tables below.

Table 5.22 Predicting performance of the best learning models

MODELLING TECHNIQUE EMPLOYED	PREDICTING PERFORMANCE		
	RECALL	PRECISION	F-MEASURE
J48	0.991	**0.995**	0.993
LR	0.975	0.963	0.969
BAGGING	**0.997**	0.994	**0.995**

Considering the best learning models built by using the three modeling techniques, all the modeling techniques fulfill the DM success criteria as the recall of each model is greater than 90%. The Bagging algorithm shows better recall as compared to the other two and J48 shows better precision. LR model shows the least predicting performance in both. The F-Measure of the bagging algorithm is the highest of all showing that its predicting performance in the learning model outperforms the other two modeling techniques. But what matters most in evaluating the models is the predicting performances the models while they are re-evaluated on their corresponding test set (as shown in Table 5.23)

Table 5.23 Predicting performance of the best models re-evaluated on test sets

MODELLING TECHNIQUE EMPLOYED	PREDICTING PERFORMANCE		
	RECALL	PRECISION	F-MEASURE
J48	0.939	**0.958**	**0.948**
LR	**0.963**	0.636	0.766
BAGGING	0.939	0.939	0.939

From the DM success criteria point of view all the modeling techniques achieve successful predicting performance. This is because; in all the models the CHURN class is predicted with 90% and above accuracy (or recall is above 90%). But it is shown that LR resulted in the highest recall in this case. While the best learning models of the three algorithms validated on their corresponding test set, the J48 model resulted in a recall (0.939), precision (0.958) and consequently the best F-Measure (0.948) values. Bagging algorithm, which shows the best predicting performance in the learning model, shows similar result in its recall (0.939) and less result in its precision (0.939) as compared to J48. LR exhibits the highest recall results (0.963) but the least precision (0.636) and consequently the least F-Measure (0.766).

The J48 modeling technique gives the best results in predicting the CHURN class as it can be seen its F-Measure value (0.948) is the highest as compared to the other two. The Bagging modeling technique, which shows almost equivalent predicting performance with J48, is the second best modeling technique for predicting the CHURN class. Only a slight difference in performance is observed between J48 and Bagging model. Hence, the J48 model is the best model to predict the CHURN class of this specific study followed by the Bagging model and LR model respectively

Apart from its highest result in predicting the CHURN class, the J48 modeling technique is the easiest to understand for individuals who are not domain experts. This is because the outcomes of the models in J48 are given in a form of a tree that anyone can simply extract rules out of it). In addition, the

attributes used for classification are: DrTxnM3, DrTxnM2, DrTxnM1, CrTxnM3, CrTxnM2, CrTxnM1, and CustDuration. This shows that a dataset for modeling can be generated with less effort of data preparation process. Anyone can simply prepare a dataset applying a tool such as MS-ACCESS for aggregating the number of debit and credit transactions of each customer in each month and the total number of days the customer stays with the bank. It is therefore the J48 modeling technique is having a reasonable potential of deployment.

5.4 Evaluation of the Outcome

Before conducting the modeling process, it has been stated in different sections that there are some factors which can affect the performance of the models. Some of these are:

- The demographic information about customers, which might serve as useful component for churn prediction, couldn't be obtained for this study,
- The data has some missing values and other quality problems, and
- It has been used a free DM tool, which has limited capability in handling enormous amount of data,

But having all these constraints, 93.9% of the total churners can be predicted correctly and the overall performance of the model is 94.8%. This shows that by improving the aforementioned constraints, there is a possibility of enhancing the performance of the model. In fact, the obtained result itself is also beyond what is expected as a business success criterion.

The impact of the result of the J48 model in accomplishing the business objective is promising. In other words 93.8% of the churners can be predicted correctly in the churn prediction step of the customers' churn management process. As the business objective is in reducing 85% of the churners in CBE, about 8.8% of extra churners are predicted as a contingency for the next churn management steps. This will make the result more reliable.

The following rules, which classify the highest number of instances of both classes, are extracted from the pruned tree of the best J48 model. Out of a total of 37 rules generated, 12 rules with greater number of instances classified (written inside the brackets at the end of each rule) are selected.

Rule 1: If (DrTxnM3 <= 2.99AND CustDuration <= 259 AND CrTxnM3 <= 0.99), THEN CLASS= CHURN (3308)

Rule 2: If(DrTxnM3 <= 2.99AND CustDuration <= 259 AND CrTxnM3 >0.99 AND DrTxnM1 <= 0.98 AND DrTxnM2 <= 1.00), THEN CLASS= CHURN(379)

Rule 3: If(DrTxnM3 <= 2.99AND CustDuration <= 259 AND CrTxnM3 >0.99 AND DrTxnM1 >0.98 AND CrTxnM3 >1.03 AND DrTxnM1 >1.03 AND DrTxnM1 <= 3.99), THEN CLASS= CHURN(120)

Rule 4: If(DrTxnM3 <= 2.99AND CustDuration <= 259 AND CrTxnM3 >0.99 AND DrTxnM1 >0.98 AND CrTxnM3 <= 1.03 AND DrTxnM1 <= 2.87 AND DrTxnM1 > 1 AND DrTxnM1 <= 2.00), THEN CLASS= CHURN(41)

Rule 5: If(DrTxnM3 >2.99AND CrTxnM2 >1.00 AND CrTxnM1 >0.96 AND DrTxnM2 >2.96 AND CrTxnM2 <= 2.02), THEN CLASS= ACTIVE(3221)

Rule 6: If(DrTxnM3 <= 2.99AND CustDuration > 259), THEN CLASS= ACTIVE(218)

Rule 7: If(DrTxnM3 >2.99AND CrTxnM2 >1.00 AND CrTxnM1 >0.96 AND DrTxnM2 >2.96 AND CrTxnM2 >2.02 AND CustDuration >216.79), THEN CLASS= ACTIVE(193)

Rule 8: If(DrTxnM3 >2.99AND CrTxnM2 >1.00 AND CrTxnM1 >0.96 AND DrTxnM2 <= 2.96 AND DrTxnM2 <= 2.08AND CrTxnM1 <= 1), THEN CLASS= ACTIVE(88)

Rule 9: If(DrTxnM3 <= 2.99AND CustDuration <= 259 AND CrTxnM3 >0.99 AND DrTxnM1 >0.98 AND CrTxnM3 <= 1.03 AND DrTxnM1 >2.87), THEN CLASS= ACTIVE(59)

Rule 10: If(DrTxnM3 <= 2.99AND CustDuration <= 259 AND CrTxnM3 >0.99 AND DrTxnM1 >0.98 AND CrTxnM3 <= 1.03 AND DrTxnM1 <= 2.87 AND DrTxnM1 <= 1 AND DrTxnM3 <= 1.10), THEN CLASS= ACTIVE(56)

Rule 11: If(DrTxnM3 >2.99AND CrTxnM2 >1.00 AND CrTxnM1 <= 0.96 AND CrTxnM1 <= 0.11 AND CrTxnM3 >0.99), THEN CLASS= ACTIVE(51)

Rule 12: If(DrTxnM3 >2.99AND CrTxnM2 >1.00 AND CrTxnM1 >0.96 AND DrTxnM2 >2.96 AND CrTxnM2 >2.02 AND CustDuration <= 216.79 AND CrTxnM2 >2.72 AND DrTxnM1 <= 4.08), THEN CLASS= ACTIVE(40)

As it can be shown from the rules above, out of the 4003 churn instances in the training set, 3848 (96.13%) churn instances are classified by the first four rules (Rule1 – Rule4). And these rules have a common property that DrTxnM3<=2.99 AND CustDuration<=259. So, customers having 2.99 or less debited transaction in the third month from back and having less than 259 days of with the bank have the highest possibility of being churned out.

The only problem observed in the J48 modeling is that it uses fractional numbers for making rules on those attributes, which can only be expressed as an integer. For instance DrTxnM3 denotes the number of debited transactions in month 3 cannot be a fraction. Its value is given as an integer (0, 1, 2, etc.). In general, the model built using the J48 modeling technique is found to be realistic, easy to learn and

having the best performance in predicting the churners in this specific study. So, it is plausible to consider the J48 model as the best and final model of this study.

The novelty of this research is that it is the first local research in the area of churn prediction and it will be used as a reference by future researchers who are interested in conducting their research on predicting churners in different sectors and industries. So, it will at least fill the knowledge gap in the area of churn prediction. The research shows the ways how the business should be understood, how the data should be prepared, the ways of selecting tools and techniques, and the modeling procedures to get the best J48 model that can predict churners with good predicting accuracy. The CBE can use the model to apply it on their existing data and benefitted from the outcome of this research or else they can conduct a new DM project following the procedures of this research; minimizing the stated limitations in order to get a model of even better performance.

CHAPTER SIX

6 Conclusion and Recommendations

6.1 Conclusion

DM is alleviating different problems in various sectors and industries. One of the major applications of DM is Churn Prediction. Customers' churn has been the main challenge for industries, which are operating under fierce competition, such as Telecom (in global context), Banks, Insurances, and Other Retail Industries. As stated in the literature part of this study, customers' churn resulted in irrecoverable damage especially in banks.

In this study, an attempt has been made to predict customers at CBE, which are having the possibility of closing their accounts (churners), based on the existing customers' data stored in database. The CRISP-DM methodology, which is believed to be the De-Facto Industry standard of DM, has been employed to conduct the study. The major steps followed are: Business domain and data understanding, Data Preparation, Modeling, and Evaluation of the output.

Sample data of Customers' Accounts and transactions, which is collected from CBE, is used to conduct the study. For the purpose of making the data appropriate for modeling, several data preparation tasks have been undertaken. The major components of the datasets are three months of transactions of customers and other information such as duration of stay of customers with the bank.

As the major goal of the study is predicting churning customers, classification DM techniques are selected for modeling. The modeling techniques used are: J48, LR, and Bagging. Pairs of Training and Testing sets of varying percentages are prepared. In order to handle the impact of class imbalance in predicting performance, SMOTE has been applied on all the training sets. The modeling process is conducted applying the cross validation test option in WEKA tool on the training sets in two ways. These are applying/not applying SMOTE on the training sets and with all/selected attributes of the training set. The modeling process (experimentation) is conducted in three steps: first by 10 fold cross validation, the second step is finding better result by changing the number of folds, and third is improving the predicting performance of the obtained models by changing the specific parameter values of each modeling algorithm. The performance of the best models of each algorithm is validated on separate test sets to show the reliability of the models.

The model which has been built applying J48 algorithm, which shows the highest predicting performance of 94.8%, is selected as the best model satisfying the entire success and test criterion followed by Bagging and LR modeling techniques respectively.

In conclusion, the results obtained from this study are very promising and show that churners can really be predicted from the patterns in the historical data of customers by applying DM tools and techniques.

6.2 Recommendations

This research work is basically conducted for an academic reason. But it is found that the results obtained out of the study are promising and can be applied to address the problems of customers' churn in CBE or other local banks. So, it is believed that the study can have a significant contribution for researches to be conducted in relation to customers' churn especially in the banking sector. It is observed in the research work that the J48 modeling technique shows good performance in predicting the churners. The technique can also be applied to give good result in other areas where there is a need to classify a binary outcomes such as Fraud/Non-Fraud, Strong/Weak, etc.

In this research, the demographic information about customers couldn't be included and the data used has some quality problems. It is believed that the result would have been better if the demographic data has been included and larger amount of data having a good quality has been used. In addition, had there been ample time and resources, the researcher would have included:

- Predicting churners by assigning the probability of each customer of being churned out
- Predicting churners along with their values to the bank (as all customers are not equally valuable to the bank & churning of high-valued customers has a damaging consequences)

So, future researches can be conducted in these areas.

Hence, based on the findings of this study, the researcher would like to forward the following recommendations:

- The bank needs to have a data warehouse which can accommodate valuable information about their customers so that future DM researches can easily be conducted without any limitation of resources.
- Churn prediction researches including the points mentioned earlier can be conducted in a broader context with ample data and using proprietary DM tools of better capabilities so that models of improved predicting performances could be built and deployed.

REFERENCES

Adalikwu, C. (2012). Customer relationship management and customer satisfaction. *African Journal of Business Management*, *6*(22), 6682–6686. doi:10.5897/AJBM12.634

Aruna, S., Rajagopalan, S. P., & Nandakishore, L. V. (2011). AN EMPIRICAL COMPARISON OF SUPERVISED LEARNING ALGORITHMS IN DISEASE DETECTION. *International Journal of Information Technology Convergence and Services (IJITCS)*, *1*(4), 81–92.

B.K., R., & Srivatsa, S. K. (2011). Evaluation of Logistic Regression and Neural Network Model With Sensitivity Analysis on Medical Datasets. *International Journal of Computer Science and Security (IJCSS)*, *5*(5), 503–511.

Ballings, M., & Poel, D. Van Den. (2012). The Relevant Length of Customer Event History for Churn Prediction : How long is long enough ? The Relevant Length of Customer Event History for Churn Prediction : How long is long enough ?

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *Sigkdd Explorations*, *6*(1), 20–29.

Bauer, E., & KOHAV, R. (2004). An Empirical Comparison of Voting Classification Algorithms : Bagging , Boosting , and Variants. *Machine Learning*, *38*(1998), 1–38.

Bekkar, M., & Alitouche, T. A. (2013). IMBALANCED DATA LEARNING APPROACHES REVIEW. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, *3*(4), 15–33.

Beniwal, S., & Arora, J. (2012). Classification and Feature Selection Techniques in Data Mining. *International Journal of Engineering Research & Technology (IJERT)*, *1*(6), 1–6.

Berry, M. J. A., & Linoff, G. S. (2004). *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management* (2nd ed., pp. 1–643). Indianapolis: Wiley Publishing, Inc.

Bhambri, V. (2011). Application of Data Mining in Banking Sector. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY*, *2*(2), 199–202.

Bhambri, V. (2012). Data Mining as a Tool to Predict Churn Behavior of Customers. *International Journal of Computer & Organization Trends*, *2*(3), 85–89.

Bhasin, M. L. (2006). Data Mining: A Competitive Tool in the Banking and Retail Industries. *The Chartered Accountant*, (October), 588–594.

Bhatnagar, S. (2012). Customer Relationship Management in Banking: Need for a strategic approach. *compendium*, 199–206.

Boulding, W., Staelin, R., Ehret, M., Johnston, W. J., Berry, L., Deighton, J., … Bolton, R. N. (2005). A Customer Relationship Management Roadmap : What Is Known , Potential Pitfalls , and Where to Go. *Journal of Marketing*, *69*(October), 155–166.

Burez, J., & Poel, D. Van Den. (2009). Expert Systems with Applications Handling class imbalance in customer churn prediction. *Expert Systems With Applications*, *36*(3), 4626–4636. doi:10.1016/j.eswa.2008.05.027

Buttle, F. (2009). *Customer Relationship Management Concepts and Technologies* (2nd ed., pp. 1–500). Burlington: Elsevier Ltd.

CBE. (2011). *Commercial Bank of Ethiopia - Annual Report 2010/11* (pp. 1–61). Addis Ababa.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *Crisp-dm 1.0 Step-by-step data mining guide* (pp. 1–73). SPSS Inc.

Chen, S. Y., & Liu, X. (2004). The Contribution of Data Mining in Information Science. *Journal of Information Science*, 1–20.

Chitra, K., & Subashini, B. (2011). Customer Retention in Banking Sector using Predictive Data Mining Technique. In *ICIT 2011 The 5th International Conference on Information Technology*.

Chopra, B., Bhambri, V., & Krishan, B. (2011). Implementation of Data Mining Techniques for Strategic CRM Issues. *Int. J. Comp. Tech. Appl.*, *2*(August), 879–883.

Cios, K. J., Pedrycz, W., Swiniarski, R. W., & Kurgan, L. A. (2007). *Data Mining A Knowledge Discovery Approach* (pp. 1–606). New York: Springer Science+Business Media, LLC.

Clewley, N., Chen, S. Y., Surfhvv, W., Plqlqj, D. W. D., Dq, L. V, Ri, H. D., … Lw, L. (2009). Applications for Data Mining Techniques in Customer Relationship Management.

Commercial Bank of Ethiopia. (2012). Commercial Bank of Ethiopia, CBE.

DENEKEW, A. (2003). *THE APPLICATION OF DATA MINING TO SUPPORT CUSTEMER RELATIONSHIP MANAGEMENT AT ETHIOPIAN AIRLINES*. ADDIS ABABA UNIVERSITY.

Deshpande, S. P., & Thakare, V. M. (2010). DATA MINING SYSTEM AND APPLICATIONS : A REVIEW. *International Journal of Distributed and Parallel systems (IJDPS) Vol.1,*, *1*(1), 32–44.

Dua, S., & Du, X. (2011). *Data Mining and Machine Learning in Cybersecurity*. New York: Taylor and Francis Group.

Endo, A., Shibata, T., & Tanaka, H. (2008). Comparison of Seven Algorithms to Predict Breast Cancer Survival. *Biomedical Soft Computing and Human Sciences*, *13*(2), 11–16.

Fagbemi, A. O., & Olowokudejo, F. F. (2011). A Comparison of the Customer Relationship Management Strategies of Nigerian Banks and Insurance Companies. *Int. J. Manag. Bus. Res*, *1*(3), 161–170.

Fayyad, U., Piatetsky-shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*, 37–54.

Gibert, K., Sànchez-Marrè, M., & Codina, V. (2010). Choosing the Right Data Mining Technique : Classification of Methods and Intelligent Recommendation. *International Environmental Modelling and Software Society (iEMSs)*, (Kdnuggets 2006).

Gilchrist, M., Mooers, D. L., Skrubbeltrang, G., & (Corresponding, F. V. (2012). Knowledge Discovery in Databases for Competitive Advantage. *Journal of Management and Strategy*, *3*(2), 2–15. doi:10.5430/jms.v3n2p2

Gupta, G., & Aggarwal, H. (2012). Improving Customer Relationship Management Using Data Mining. *International Journal of Machine Learning and Computing*, *2*(6), 874–877.

HADDEN, J. (2008). *A Customer Profiling Methodology for Churn Prediction*. CRANFIELD UNIVERSITY, SCHOOL OF APPLIED SCIENCES.

Hall, M., National, H., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2013). The WEKA Data Mining Software : An Update. SIGKDD Explorations. Retrieved from http://www.cs.waikato.ac.nz/ml/weka/

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (2nd ed., pp. 1–743). San Francisco: Morgan Kaufmann Publishers.

Henock, W. (2002). *APPLICATION OF DATA MINING TECHNIQUES TO SUPPORT CUSTOMER RELATIONSHIP MANAGEMENT AT ETHIOPIAN AIRLINES*. ADDIS ABABA UNIVERSITY.

Hlosta, M., Stríž, R., Kup, J., Zendulka, J., & Hruška, T. (2013). Constrained Classification of Large Imbalanced Data by Logistic Regression and Genetic Algorithm. *International Journal of Machine Learning and Computing*, *3*(2), 214–218. doi:10.7763/IJMLC.2013.V3.305

Hosseni, M. B., & Tarokh, M. J. (2011). Customer Segmentation Using CLV Elements. *Journal of Service Science and Management*, *4*(September), 284–290. doi:10.4236/jssm.2011.43034

Hu, X. (2005). A Data Mining Approach for Retailing Bank Customer Attrition Analysis. *Applied Intelligence*, *22*, 47–60.

Jackson, J. (2002). DATA MINING : A CONCEPTUAL OVERVIEW. *Communications of the Association for Information Systems (Volume*, *8*, 267–296.

Jain, Y. K., & Upendra. (2012). An Efficient Intrusion Detection Based on Decision Tree Classifier Using Feature Reduction. *International Journal of Scientific and Research Publications,*, *2*(1), 1–6.

Jantan, H., Hamdan, A. R., & Othman, Z. A. (2010). Human Talent Prediction in HRM using C4 . 5 Classification Algorithm. *International Journal on Computer Science and Engineering*, *02*(08), 2526–2534.

Karimii, O., Maymand, M. M., Hosseini, M. H., & Ahmadinejad, M. (2012). Customer Switching Behavior : Developing model in the Iranian Retail Banking Industry. *Journal of Basic and Applied Scientific Research*, *2*(12), 11984–11991.

Kulikowski, J. L. (2011). KNOWLEDGE MINING FROM DATA : METHODOLOGICAL PROBLEMS AND DIRECTIONS FOR DEVELOPMENT. *TASK QUARTERLY 15 No 2*, (2), 227–233.

Kumar, D., & Bhardwaj, D. (2011). Rise of Data Mining : Current and Future Application Areas. *International Journal of Computer Science Issues*, *8*(5), 256–260.

Kumneger, F. (2006). *Application of Data Mining Techniques to support Customer Relationship Management (CRM) for Ethiopia Shipping Lines (ESL)*. Addis Ababa University.

KURGAN, L. A., & MUSILEK, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, *21*(1), 1–24. doi:10.1017/S0269888906000737

Lazarov, V., & Capota, M. (2007). Churn Prediction.

Lemmens, A., & Croux, C. (2006). BAGGING AND BOOSTING CLASSIFICATION TREES TO PREDICT CHURN. *Journal of Marketing Research*, *43*(2).

LUEL, B. (2011). *THE ROLE OF DATA MINING TECHNOLOGY IN ELECTRONIC TRANSACTION EXPANSION AT DASHEN BANK S.C*. ADDIS ABABA UNIVERSITY.

Madhavi, S. (2012). THE PREDICTION OF CHURN BEHAVIOUR AMONG INDIAN BANK CUSTOMERS : AN APPLICATION OF DATA MINING TECHNIQUES. *International Journal of Marketing, Financial Services & Management Research*, *1*(2), 11–19.

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook*. (O. Maimon & L. Rokach, Eds.) (2nd ed., pp. 1–1285). Springer Science+Business Media, LLC.

MARISCAL, G., MARBAN, O., & FERNANDEZ, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, *25*(2), 137–166. doi:10.1017/S0269888910000032

Maroofi, F., Aliabadi, B. M., Fakhri, H., & Hadikolivand. (2013). Effective Factors on CRM Development. *Asian Journal of Business Management*, *5*(1), 52–59.

Miguéis, V. L., Poel, D. Van Den, Camanho, A. S., & Falcão, J. (2012). Modeling partial customer churn : On the value of first product-category purchase sequences. *Expert Systems with Applications*, *39*, 11250–11256. doi:10.1016/j.eswa.2012.03.073

Mishra, A., & Mishra, D. (2009). Customer Relationship Management : Implementation Process Perspective. *Acta Polytechnica Hungarica Vol.*, *6*(4), 83–99.

Moin, K. I., & Ahmed, Q. B. (2012). Use of Data Mining in Banking. *International Journal of Engineering Research and Applications (IJERA)*, *2*(2), 738–742.

National Bank of Ethiopia. (2012). Nbebank.com. Retrieved March 13, 2013, from http://www.nbe.gov.et/

Nejad, M. B., Nejad, E. B., & Karami, A. (2012). Using Data Mining Techniques to Increase Efficiency of Customer Relationship Management Process. *Research Journal of Applied Sciences, Engineering and Technology*, *4*(23), 5010–5015.

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, *38*, 15273–15285. doi:10.1016/j.eswa.2011.06.028

Nisbet, R., Elder, J., & Miner, G. (2009). *HANDBOOK OF STATISTICAL ANALYSIS AND DATA MINING APPLICATIONS* (pp. 1–824). Burlington: Elsevier Inc.

Ogwueleka, F. N. D. (2009). Potential Value of Data Mining for Customer Relationship Marketing in the Banking Industry. *Advances in Natural and Applied Sciences*, *3*(1), 73–78.

Ombati, T. O., Magutu, P. O., Nyamwange, S. O., & Nyaoga, R. B. (2010). TECHNOLOGY AND SERVICE QUALITY IN Importance and Performance of Various Factors Considered In the. *African Journal of Business & Management (AJBUMA)*, *1*, 151–164.

P. Sundari, & K. Thangadurai. (2010). An Empirical Study on Data Mining Applications. *Global Journal of Computer Science and Technology*, *10*(5), 23–27.

Padhy, N., 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), 43–58.

Parvatiyar, A., & Sheth, J. N. (2002). Customer Relationship Management : Emerging Practice , Process , and Discipline. *Journal of Economic and Social Research*, *3*(2), 1–34.

Ponce, J., & Karahoca, A. (2009). *Data Mining and Knowledge Discovery in Real Life Applications*. (J. Ponce & A. Karahoca, Eds.) (pp. 1–438). Vienna: In-Teh.

Prasad, U. D., & Madhavi, S. (2012). PREDICTION OF CHURN BEHAVIOR OF BANK CUSTOMERS USING DATA MINING TOOLS. *Business Intelligence Journal*, *5*(1), 96–101.

Radosavljevik, D., Putten, P. Van Der, & Larsen, K. K. (2010). The Impact of Experimental Setup in Prepaid Churn Prediction for Mobile Telecommunications : What to Predict , for Whom and Does the Customer Experience Matter ? *Transactions on Machine Learning and Data Mining*, *3*(2), 80–99.

Rai, R., & Singh, R. P. (2012). CRM in Banking : Trends & Dynamics. *GIAN JYOTI E-JOURNAL*, *1*(2).

Rajput, A., Aharwal, R. P., Dubey, M., Saxena, S. P., & Raghuvanshi, M. (2011). J48 and JRIP Rules for E-Governance Data. *International Journal of Computer Science and Security (IJCSS)*, *5*(2), 201–207.

Rashid, T. (2008). Classification of Churn and non-Churn Customers for Telecommunication Companies. *International Journal of Biometrics and Bioinformatics*, *3*(5), 82–89.

Sachdev, S. B., & Verma, H. V. (2004). RELATIVE IMPORTANCE OF SERVICE QUALITY DIMENSIONS: A MULTISECTORAL STUDY. *Journal of Services Research*, *4*(1).

Sahu, H., Shrma, S., & Gondhalakar, S. (2008). A Brief Overview on Data Mining Survey. *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*, *1*(3), 114–121.

Scott, J. W., & Arias, J. C. (2011). Banking profitability determinants. *Business Intelligence Journal*, *4*(2), 209–230.

Setty, D. V., T.M.Rangaswamy, & K.N.Subramanya. (2010). A Review on Data Mining Applications to the Performance of Stock Marketing. *International Journal of Computer Applications (0975 – 8887)*, *1*(3), 24–34.

SHARMA, A., JULKA, T., & BHARDWAJ, S. (2012). Customer relationship management: a growth catalyst for hdfc bank. *ZENITH International Journal of Business Economics & Management Research*, *2*(2), 149–166.

Sharma, A., & Panigrahi, P. K. (2011). A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services. *International Journal of Computer Applications*, *27*(11), 26–31.

Singh, Y., & Chauhan, A. S. (2009). Neural networks in data mining. *Journal of Theoretical and Applied Information Technology*, 37–42.

Soeini, R. A., & Rodpysh, K. V. (2012). Applying Data Mining to Insurance Customer Churn Management. In *2012 IACSIT Hong Kong Conferences IPCSIT* (Vol. 30, pp. 82–92). Singapore.

Stanley, S. (2012). NEW PERSPECTIVES IN THE BANKING SECTOR – THE CRM WAY. *International Journal of Marketing, Financial Services & Management Research*, *1*(11), 19–24.

Sumathi, S., & Sivanandam, S. N. (2006). *Introduction to Data Mining and its Applications*. *Studies in Computational Intelligence , Volume 29* (pp. 1–828). Berlin: Springer-Verlag Berlin Heidelberg.

TARIKU, A. (2011). *MINING INSURANCE DATA FOR FRAUD DETECTION: THE CASE OF AFRICA INSURANCE SHARE COMPANY*. ADDIS ABABA UNIVERSITY.

Tesfaye, H. (2002). *PREDICTIVE MODELING USING DATA MINING TECHNIQUES IN SUPPORT OF INSURANCE RISK ASSESSMENT*. ADDIS ABABA UNIVERSITY.

Thanuja, V., Venkateswarlu, B., & Anjaneyulu, G. S. G. N. (2011). Applications of Data Mining in Customer Relationship Management. *Journal of Computer and Mathematical Sciences*, *2*(3), 423–433.

Tufféry, S. (2011). *DATA MINING AND STATISTICS FOR DECISION MAKING*. (P. Giudici, G. H. Givens, & B. K. Mallick, Eds.) (pp. 1–689). West Sussex: John Wiley & Sons, Ltd Registered.

Two Crows Corporation. (2005). *Introduction to Data Mining and Knowledge Discovery* (3rd ed., pp. 1–36). Two Crows Corporation.

V. TAMILVENDAN, & S. SWAMIDOSS. (2012). A study on customer relationship management in banking industry. *Journal of Exclusive Management Science*, *1*(2), 1–4.

Venkatadri, M., & Reddy, L. C. (2011). A Review on Data mining from Past to the Future. *International Journal of Computer Applications*, *15*(7), 19–22.

Wahbeh, A. H., Al-radaideh, Q. A., Al-kabi, M. N., & Al-shawakfa, E. M. (2011). A Comparison Study between Data Mining Tools over some Classification Methods. *International Journal of Advanced Computer Science and Applications*, (Special Issue on Artificial Intelligence,), 18–26.

Wang, M., & Kang, K.-W. (2008). Applying Customer Relationship Management on the New Recruiting Force Project -An Example of Ministry of National Defense. *The Journal of Human Resource and Adult Learning*, *4*(2), 185–189.

Weiss, G. M. (2004). Mining with Rarity : A Unifying Framework. *SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets*, *6*(1), 7–19.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools and Techniques* (3rd ed., pp. 1–629). Burlington: Elsevier Inc.

ZenTut. (2013a). Data Mining Tutorial. *data mining*. Retrieved April 06, 2013, from http://www.zentut.com/data-mining/

ZenTut. (2013b). Advantages and Disadvantages of Data Mining. *data mining*. Retrieved April 09, 2013, from http://www.zentut.com/data-mining/advantages-and-disadvantages-of-data-mining/

ZenTut. (2013c). Data Mining Techniques. *data mining*. Retrieved April 09, 2013, from http://www.zentut.com/data-mining-

APPENDIXES

Appendix I Output of the J48 Best Model

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: 70_Perc_TS-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.supervised.instance.SMOTE-C0-K5-P1663.87-S1-weka.filters.unsupervised.attribute.Remove-R2,4,6,8,10,12-16

Instances: 8007

Attributes: 8

DrTxnM3

CrTxnM3

DrTxnM2

CrTxnM2

DrTxnM1

CrTxnM1

CustDuration

AcctStatus

Test mode: 20-fold cross-validation

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

J48 pruned tree

DrTxnM3 <= 2.995798

| CustDuration <= 259

| | CrTxnM3 <= 0.99: CHURN (3308.0)

| | CrTxnM3 >0.99

| | | DrTxnM1 <= 0.98

| | | | DrTxnM2 <= 1.00: CHURN (379.0/3.0)

| | | | DrTxnM2 >1.00

| | | | | DrTxnM1 <= 0.003068

| | | | | | CrTxnM1 <= 0.940471: CHURN (13.0/1.0)

| | | | | | CrTxnM1 > 0.940471: ACTIVE (9.0/1.0)

| | | | | DrTxnM1 > 0.003068: CHURN (24.0)

| | | DrTxnM1 >0.98

| | | | CrTxnM3 <= 1.03

| | | | | DrTxnM1 <= 2.87

| | | | | | DrTxnM1 <= 1

| | | | | | | DrTxnM3 <= 1.10: ACTIVE (56.0/3.0)

| | | | | | | DrTxnM3 >1.10

| | | | | | | | DrTxnM3 <= 1.894512: CHURN (6.0)

| | | | | | | | DrTxnM3 > 1.894512: ACTIVE (10.0/1.0)

| | | | | | DrTxnM1 > 1

| | | | | | | DrTxnM1 <= 2.00: CHURN (41.0)

| | | | | | | DrTxnM1 >2.00

| | | | | | | | DrTxnM3 <= 0.762097: CHURN (6.0)

| | | | | | | | DrTxnM3 > 0.762097: ACTIVE (16.0/3.0)

| | | | | DrTxnM1 >2.87: ACTIVE (59.0/1.0)

| | | | CrTxnM3 >1.03

| | | | | DrTxnM1 <= 1.03: ACTIVE (8.0/1.0)

| | | | | DrTxnM1 >1.03

| | | | | | DrTxnM1 <= 3.99: CHURN (120.0)

| | | | | | DrTxnM1 >3.99

| | | | | | | CustDuration <= 197.5402: CHURN (18.0/2.0)

| | | | | | | CustDuration > 197.5402: ACTIVE (2.0)

| CustDuration > 259: ACTIVE (218.0)

DrTxnM3 > 2.995798

| CrTxnM2 <= 1.00

| | DrTxnM2 <= 0.074065: ACTIVE (5.0)

| | DrTxnM2 > 0.074065

| | | CrTxnM1 <= 0.949495: CHURN (26.0)

| | | CrTxnM1 > 0.949495

| | | | CrTxnM3 <= 0.798188: CHURN (3.0)

| | | | CrTxnM3 > 0.798188: ACTIVE (3.0)

| CrTxnM2 >1.00

| | CrTxnM1 <= 0.96

| | | CrTxnM1 <= 0.11

| | | | CrTxnM3 <= 0.99

| | | | | CrTxnM3 <= 0.099193

| | | | | | DrTxnM2 <= 1.490773: ACTIVE (4.0)

| | | | | | DrTxnM2 > 1.490773: CHURN (2.0)

| | | | | CrTxnM3 > 0.099193: CHURN (7.0)

| | | | CrTxnM3 >0.99: ACTIVE (51.0/2.0)

| | | CrTxnM1 >0.11: CHURN (9.0)

| | CrTxnM1 >0.96

| | | DrTxnM2 <= 2.96

| | | | DrTxnM2 <= 2.075059

| | | | | CrTxnM1 <= 1: ACTIVE (88.0)

| | | | | CrTxnM1 > 1

| | | | | | CrTxnM1 <= 1.990647: CHURN (8.0)

| | | | | | CrTxnM1 > 1.990647: ACTIVE (15.0/1.0)

| | | | DrTxnM2 > 2.075059: CHURN (11.0)

| | | DrTxnM2 >2.96

| | | | CrTxnM2 <= 2.02: ACTIVE (3221.0/3.0)

| | | | CrTxnM2 >2.02

| | | | | CustDuration <= 216.79

| | | | | | CrTxnM2 <= 2.72: CHURN (7.0)

| | | | | | CrTxnM2 >2.72

| | | | | | | DrTxnM1 <= 4.08: ACTIVE (40.0)

| | | | | | | DrTxnM1 >4.08

| | | | | | | | CrTxnM3 <= 2.440051: ACTIVE (10.0)

| | | | | | | | CrTxnM3 > 2.440051

| | | | | | | | | DrTxnM1 <= 5.958918: CHURN (4.0)

| | | | | | | | | DrTxnM1 > 5.958918: ACTIVE (7.0/1.0)

| | | | | CustDuration >216.79: ACTIVE (193.0)

Number of Leaves: 37

Size of the tree: 73

Time taken to build model: 0.14 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	7950	99.2881 %
Incorrectly Classified Instances	57	0.7119 %
Kappa statistic	0.9858	
Mean absolute error	0.0095	
Root mean squared error	0.082	
Relative absolute error	1.9058 %	
Root relative squared error	16.3922 %	
Coverage of cases (0.95 level)	99.4879 %	
Mean rel. region size (0.95 level)	50.6682 %	
Total Number of Instances	8007	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.991	0.005	0.995	0.991	0.993	0.986	0.996	0.996	CHURN
	0.995	0.009	0.991	0.995	0.993	0.986	0.996	0.990	ACTIVE
Weighted Avg.	0.993	0.007	0.993	0.993	0.993	0.986	0.996	0.993	

=== Confusion Matrix ===

a	b	<-- classified as
3967	36	a = CHURN
21	3983	b = ACTIVE

Appendix II Output of the Best LR Model

=== Run information ===

Scheme: weka.classifiers.functions.Logistic -R 0.01 -M -1

Relation: 66_Perc_TS-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.supervised.instance.SMOTE-C0-K5-P1665.27-S1

Instances: 7625

Attributes: 18

DrTxnM3

AvgDrAmtM3

CrTxnM3

AvgCrAmtM3

DrTxnM2

AvgDrAmtM2

CrTxnM2

AvgCrAmtM2

DrTxnM1

AvgDrAmtM1

CrTxnM1

AvgCrAmtM1

TotDrTxn

TotDrAmt

TotCrTxn

TotCrAmt

CustDuration

AcctStatus

Test mode: 10-fold cross-validation

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

LR with ridge parameter of 0.01

Coefficients...

Variable	Class CHURN
DrTxnM3	-0.7625
AvgDrAmtM3	0.0001
CrTxnM3	-0.0985
AvgCrAmtM3	0
DrTxnM2	-0.6106
AvgDrAmtM2	0
CrTxnM2	-0.3831
AvgCrAmtM2	0.0001
DrTxnM1	-0.0649
AvgDrAmtM1	0
CrTxnM1	0.2916
AvgCrAmtM1	0
TotDrTxn	-0.2772
TotDrAmt	0
TotCrTxn	-0.1011
TotCrAmt	0

CustDuration -0.0187

Intercept 7.7547

Odds Ratios...

Class

Variable CHURN

==========

DrTxnM3 0.4665

AvgDrAmtM3 1.0001

CrTxnM3 0.9062

AvgCrAmtM3 1

DrTxnM2 0.543

AvgDrAmtM2 1

CrTxnM2 0.6818

AvgCrAmtM2 1.0001

DrTxnM1 0.9372

AvgDrAmtM1 1

CrTxnM1 1.3386

AvgCrAmtM1 1

TotDrTxn 0.7579

TotDrAmt 1

TotCrTxn 0.9039

TotCrAmt 1

CustDuration 0.9815

Time taken to build model: 0.74 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	7391	96.9311 %
Incorrectly Classified Instances	234	3.0689 %
Kappa statistic	0.9386	
Mean absolute error	0.0503	
Root mean squared error	0.1585	
Relative absolute error	10.0682 %	
Root relative squared error	31.7098 %	
Coverage of cases (0.95 level)	99.1344 %	
Mean rel. region size (0.95 level)	57.141 %	
Total Number of Instances	7625	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.976	0.037	0.963	0.976	0.969	0.939	0.990	0.987	CHURN
	0.963	0.024	0.975	0.963	0.969	0.939	0.990	0.989	ACTIVE
Weighted Avg.	0.969	0.031	0.969	0.969	0.969	0.939	0.990	0.988	

=== Confusion Matrix ===

a	b	<-- classified as
3719	93	a = CHURN
141	3672	b = ACTIVE

Appendix III Guiding Questions prepared for the discussions held with relevant personnel during business understanding

==========

Q1: - Among the major CRM related problems stated below, which problems are the most critical problems observed in CBE? (Select one or many)

- ☐ Credit Default
- ☐ Fraud
- ☐ Segmenting customers based on Loyalty
- ☐ Customers' churn
- ☐ Others please specify _____

  _____

Q2: - For the problem(s) you selected in Q1, how worse it is?

Q3: - What precaution measures are being taken currently and how effective the measures are in alleviating the problem?

Q4: - By how much these problem should be reduced in order to get a better change/success in the business?

==========

Appendix IV Letter of Cooperation written from AAU to CBE

አዲስ አበባ ዩኒቨርሲቲ
የተፈጥሮ ሳይንስ ኮሌጅ
የኢንፎርሜሽን ሳይንስ ት/ቤት

Addis Ababa University
College of Natural Sciences
School of Information Science

05 APR 2013

Ref: SIS/035/2013
Date: March 20, 2013

To: Commercial Bank of Ethiopia
Addis Ababa

Subject: Request for Cooperation

Dear Sir / Madam

Mr. Kassahun Gebremeskel Mebratu (ID: GSE/0969/03) is a Graduate student in Information

He is currently conducting research on Application of Data Mining Techniques to Support Customer Relationship Management at Commercial Bank of Ethiopia.

I would like to thank you in advance for all the assistance that you would provide to the student.

With regards

SCHOOL OF INFORMATION SCIENCE ADDIS ABABA UNIVERSITY

Solomon Teferra (PhD)
Head, School of Information Science

aau
✉: 1176 ☎: +251-(11)-122-91-71 ☎:251- (11)-122- 91-00 FAX: 251-(11)-123-972 ☎:2122- 91-92

Appendix V Letter of Cooperation written from CBE, HRM directorate to relevant divisions

የኢትዮጵያ ንግድ ባንክ
COMMERCIAL BANK OF ETHIOPIA
Inter Departmental Memorandum

DATE ቀን	:	April 24, 2013
TO ለ	:	Director – CATS CPC
FROM ከ	:	A/Director – Human Resource Development
SUBJET ጉዳዩ	:	Request for assistance and cooperation

Addis Ababa University, College of Natural Science, under its letter Ref. SIS/035/2013 dated March 20, 2013 has requested our bank to assist and cooperate Ato Kassahun Gebremskel Mebratu to undertake his research project on the topic "Application of Data Mining Techniques to Support Customer Relationship Management at CBE".

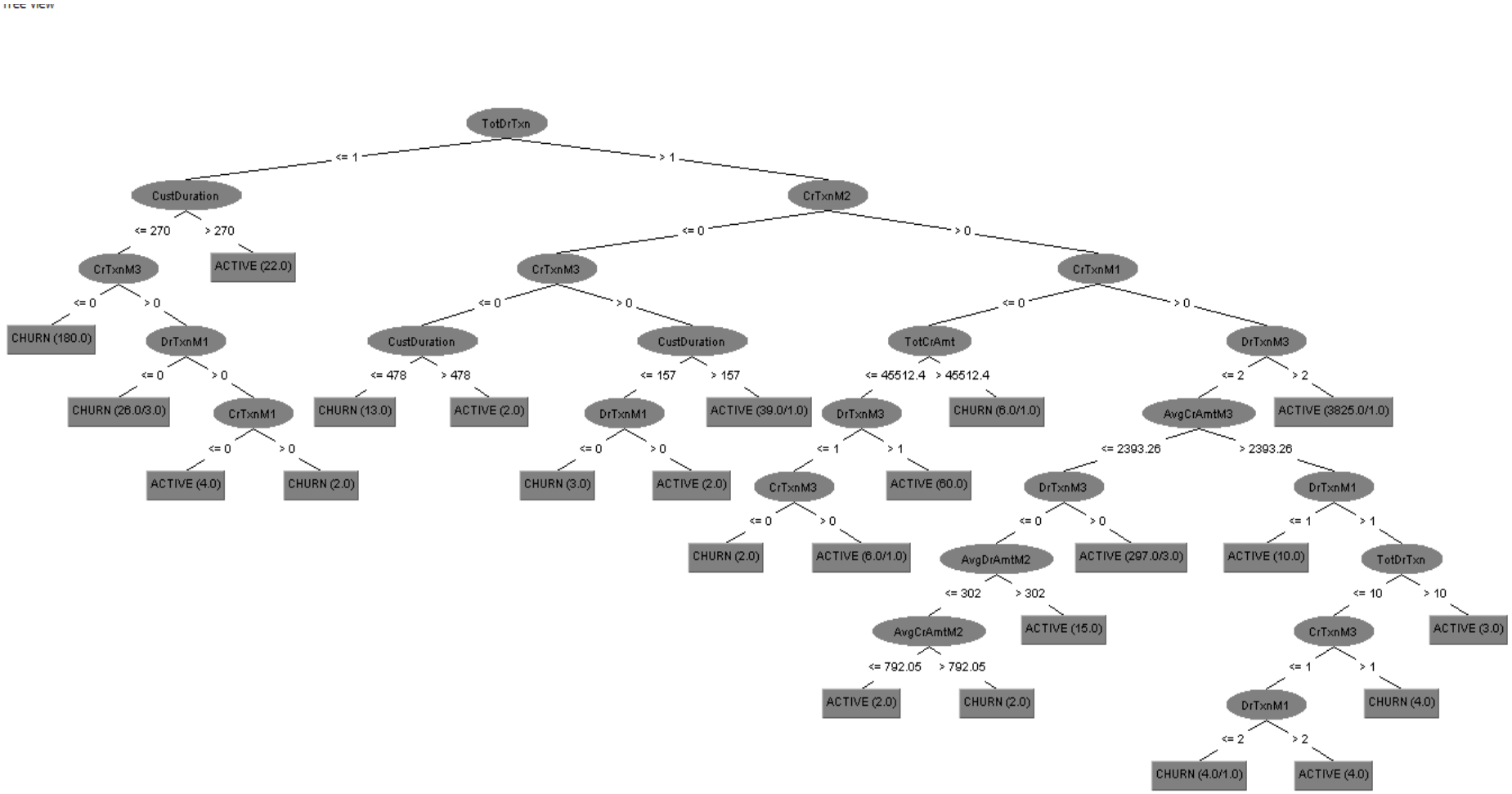
This is, therefore, to request you to provide him the required assistance and cooperation without compromising confidentiality.

Lubaba Damtie

ab /

Corporate HR Development
Tel: 0114-43-09-05/0114-43-06-01/0114-43-09-56/0114-43-08-48
Fax: 0114-43-08-43
P.O.Box 255 A.A

Appendix VI The Tree formed by the J48 Best Model (Before SMOTE is applied)



DECLARATION

This thesis is my original work, has not been presented for a partial fulfillment of the requirement of a degree in any university and that all sources of material used for the thesis have been duly acknowledged.

_____

KASSAHUN GEBREMESKEL

SEPTEMBER, 2013

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

_____

Dr. Dereje Teferi