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

POSSIBLE APPLICATION OF DATA MINING TECHNIQUES TO TARGET POTENTIAL VISA CARD USERS IN DIRECT MARKETING (THE CASE OF DASHEN BANK S.C.)

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

BY

TILAHUN MULUNEH

JANUARY, 2009

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

POSSIBLE APPLICATION OF DATA MINING TECHNIQUES TO TARGET POTENTIAL VISA CARD USERS IN DIRECT MARKETING (THE CASE OF DASHEN BANK S.C.)

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

BY

TILAHUN MULUNEH

JANUARY, 2009

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

POSSIBLE APPLICATION OF DATA MINING TECHNIQUES TO TARGET POTENTIAL VISA CARD USERS IN DIRECT MARKETING (THE CASE OF DASHEN BANK S.C.)

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

BY

TILAHUN MULUNEH

JANUARY, 2009

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

POSSIBLE APPLICATION OF DATA MINING TECHNIQUES TO TARGET POTENTIAL VISA CARD USERS IN DIRECT MARKETING. (THE CASE OF DASHEN BANK S.C.)

BY

TILAHUN MULUNEH

Name and Signature of Members of the Examining Board

----------	----------	----------
Chair person, Examining Board	Signature	Date
Dr. MANOJ VNV	----------	----------
Advisor	Signature	Date
----------	----------	----------
Chair person, Faculty	Signature	Date
----------	----------	----------
Chair person, Graduate Council	Signature	Date

Dedication

I would like to dedicate the whole thesis to my dear father (**Muluneh Arage Tessema**) and mother (**Aguagu G/Egziabher G/Hiwot**), to express my deepest appreciation towards them and for their never-ending support that they have extended me in every step of my life.

God bless them!

Acknowledgment

First and for most thanks to the almighty God!

This thesis is written as part of Master's program in Information Science at the faculty of Informatics, in the department of Information Science, Addis Ababa University.

I would like to thank sincerely all those who helped me with their valuable support during the entire process of this thesis. I would especially like to express my deepest gratitude to my advisor Dr. Manoj VNV, Prof from Addis Ababa university, faculty of technology, department of electrical and computer engineering, for his invaluable guidance, strong support, encouragement and his helpful comments throughout the progress of this thesis.

I would also like to thank all respondents who gave me their valuable time during the data collection phase. I really appreciate their cooperative attitude towards this research. I pay my special gratitude to Meseret, Library staff, for providing excellent facilities and full cooperation to all of the students.

Finally I would like to thank my loved ones, my parents, brothers and sisters for their love, affection, prayers and endless support. Above all, I would like to appreciate my father Muluneh Arage, for his incredible support, he has always been behind every achievement that I have made in my life, and I am really grateful and proud of him.

Tilahun Muluneh Arage
Addis Ababa University
January, 2009

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI: - Artificial Intelligence

AD: -Advertising

ANN: - Artificial Neural Networks

ATM: Automated Teller Machine

CART: - Classification and Regression Trees

CHAID: - chi-squared Automatic Interaction Detection

CIC: - Customer Interaction Centers

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

CRM: - Customer Relation Management

DSF: - Decision-Support Functions

EM: - Expectation Maximization

ESL: - Ethiopian Shipping Line

KDD: - Knowledge Discovery in Database

K-NN: - K-Nearest Neighbors

LCV: - Lifetime Customer Value

MIS: - Management Information Systems

OLAP: - Online Analytic Process

POS: - Point Of Sale

SOM: - Self Organizing Map

SQL: -Standard Query Language

WEKA: -Waikato Environment for Knowledge Analysis

Abstract

Identifying customers who are more likely to respond to a product offering is an important issue in direct marketing. In direct marketing, data mining has been used extensively to identify potential customers for a new product (target selection).

The final goal of this thesis is to build a model that helps to classify the book customers of DASHEN BANK S.C, according to their expected response to the direct marketing campaign. Since there are no predefined classes, that describe the customers of the bank according to their expected response to visa card offer, the researcher uses a clustering techniques that resulted in the appropriate number of clusters. Then, a predictive response model was developed to predict the degree of likelihood (as high, medium and low) that a customer is going to respond to a visa card offer. This predictive model achieved accuracy of 96.14%.

For modeling purpose customers' data was gathered from DASHEN BANK S.C. Since irrelevant or redundant features result in bad model performance, data preparation like attribute selection was performed in order to determine the inputs to the model.

Thus various data mining techniques and algorithms were used to implement each step of the modeling process and alleviate related difficulties. K-Means was used as a clustering algorithm to segment customers' record into clusters with similar characters. Different parameters were used to run the clustering algorithm before reaching at segments that made business sense. J48 decision tree algorithm was used for classification purpose. In addition to those attributes that are believed by the experts to have high impact on customers probability to be the potential customer of the visa card service, this research found the attributes "Occupation" and "Accommodation" to have a big influence. Moreover, with respect to the attribute "Age" a new pattern was found.

Generally the result of the study was encouraging, which reinforce the possible application of data mining solution to the banking industry, particularly in direct marketing campaign at the DASHEN BANK S.C.

Chapter One

Introduction

1.1Background

Analyzing, interpreting and making maximum use of data has been difficult and resource demanding due to the exponential growth of many business, governmental and scientific databases. It is estimated that the amount of data stored in the world's database grows every twenty months at a rate of 100% [49].

This fact shows that we are getting more and more exploded by data or information and yet ravenous for knowledge. Data mining therefore appears as a useful tool to address the need for shifting useful information such as hidden patterns from databases [44].

Data mining can help banking firms make crucial business decisions and turn the newly discovered knowledge into actionable results in business practices such as new service development, marketing, claim distribution analysis, solvency analysis and customer identification process [44]. As can be clearly seen, data mining appears as a useful tool to address the need for effective customer relation management (CRM).

Since the banking industry in Ethiopia is expanding, the researcher believes that data mining technology have a lot to do with it. Currently many Ethiopian organizations are making important business decisions merely based on experts' judgment, which could miss hidden but very important knowledge that is crucial for the decision being made [2].

Modern banking in Ethiopia has been operating for about hundred years, In spite of long-term existence and operation, the banking industry is still in its infant stage in terms of expansion, service delivery and application of modern information Technology [33].

DASHEN BANK S.C. is a private bank which was established in 1995 by 11 investors and now has 200 shareholders and also it has more than 500,000 book customers. It is the first bank in the history of Ethiopia to introduce the visa technology and also the first to be appointed as a principal member of Visa association to issue and acquire visa card [25].

The bank originally imported the visa card technology from American companies at a cost of 3.5 million dollars. In addition to this the bank has made huge payment for experts, who install and maintain the system [25]. After installation process was completed, the bank has incurred many costs in terms of replacing damaged machines, paying of telecommunication cost and many more running costs. Currently the major available e-banking (electronic banking) service at the bank is visa card service, which is provided widely in two ways, one is ATM service and the other is Point Of Sale (POS). The bank has an immediate target to offer world-class card payment services to over 260,000 customer accounts in Ethiopia. Currently it has nearly 25,000 visa card customers but the number of customers is not growing as it was expected [41].

Formal banking has been unable to provide access to poor rural and urban Ethiopians that comprise 45% of the overall populations [3]. In addition to this before DASHEN BANK S.C. launched e-banking service, the commercial Bank of Ethiopia has asked the Visa Association to start the e-banking service but could not be acceptable by the association on the grounds that the service would not find enough customers [41].

The first step toward successful business is making sure that there are enough users of the service and drives its venture to reach them through appropriate mechanism. Bank of America could be taken as an example in studying its customer's data and able to device a way that resulted in increasing of its customers in a short period of time.

CRM is used for the overall process of exploiting customer related information and using it to enhance the revenue flow from existing customers. Customer segmentation is the process of dividing customers into homogenous groups on the basis of shared attributes, and is at the heart of CRM [16].

Before data mining caught on several years ago, a direct mail campaign was thought to be successful if it achieved a response rate of 6 to 7 percent. In 1998, the Canadian Imperial Bank of Commerce utilized CRM and data mining to achieve a phenomenal response rate of 47 percent [49]. Much of the success was attributed to targeting the right customers and being able to predict their responses. Fleet Bank also used data mining and CRM to identify the best prospects for marketing its mutual funds based on customer demographics and account data.

As clearly described, the ability to forecast the hidden behavior of customers is one of the major criteria towards a successful business. Unless the customer base is critically identified and addressed, crisis will be the most probable destiny of the business.

Banks database is full of complicated data, where critical data about the customers is stored. But it is surprising to see the fact that little effort is being made to use this valuable data as a base for decision making. Identifying and knowing a given segment of customers in the banks database will help the decision makers in deciding on the specific future actions regarding to their customers and business.

Thus in Ethiopia lack of capacity to reach the population together with lack of emphasis on historical data manipulation techniques, makes the effort towards modern e-banking service difficult and full of obstacles [3].

In this contemporary world, where the accumulation of data is increasing in an alarming rate, understanding pattern in each and every segment of customers is an important issue to be considered to adjust strategies, to make maximum use of it, and find new opportunities. Organizations keeping data on their domain area takes every record as an opportunity in learning facts. But the simple gathering of data is not enough to get maximum knowledge out of it.

Thus for an effective learning, data from many sources must first be gathered together and organized in a consistent and useful way, data warehousing. Data warehousing allows the enterprise to recognize what it has noticed about its domain area. The data must also be analyzed, understood, and turned into actionable information. This is the point where the application of data mining is needed.

Before the introduction of sophisticated tools, the only analysis made on data to get meaning out of it is simple statistical manipulations that have very small power to show all the necessary hidden information content of a given data. But data mining technology, on the other hand has the greatest potential in identifying various interesting patterns for enabling organizations to control data resources for strategic planning and decision making in their domain area.

Until recently many researches have been done to be evidence for the possible applications of data mining techniques in different parts of the world including Ethiopia. Researchers tried to prove its applicability in many domain areas and organizations. It was the researcher belief that

data mining techniques are also applicable to facilitate the direct marketing campaign process particularly in identifying potential visa card users of DASHEN BANK S.C.

It is obvious that the output of different researches may vary from time to time or place to place depending on the socio-economic and other specific situation.

Thus throughout this research an attempt was made to apply data mining tools and techniques in analyzing and determining the appropriate number of segments of customers according to their response to visa card offer. In addition to this, a classification model was built.

1.2 Statement of the Problem

DASHEN BANK S.C. seems to forget the importance of keeping and analyzing of customers' data with the appropriate tools and techniques, which help to come up with a better view of its customers and design the appropriate business strategy. Currently, there is a very poor traditional means of knowing what is unique (distinguishing) features and need of their visa card customers and how to use this knowledge to make better future decisions by facilitating the ordinary process of direct marketing, that involves identification and reaching of new customers.

Thus the underlying problem that necessitates this research is the inability of the bank to identify potential customers of visa card service to conduct direct marketing campaign. This problem interns has its long lasting penalties such as unable to have a clearer and bigger picture of their customers, lost revenue, lessening of customers or a market not being fully realized.

Although data on customers detail is always gathered and stored in the bank's database, due to lack of proper attention and appropriate data analysis tools, the data is not practically used to lighten the difficulty faced on the DASHEN BANK S.C, that is the identification of the book customer that are potential candidates for the visa card service and doing strategies accordingly.

As the researcher mentioned, the bank has made a huge investment to bring the e-banking system into function. But having all those expenses, the technology could not be acceptable by the population as being expected.

The reasons for the above case could be many, but as being explained by the marketing experts in the bank, the main reason for the low rate growth of visa card customers is weak promotional strategy, which is a direct result of not knowing what are the distinguishing hidden features of its visa card customers and hence inability to devise a mechanism that possibly takes the bank to the customers than the other way round. They added that, as DASHEN BANK S.C. is the only bank that provides both the ATM and POS service, not only its present book customers but also it could attract other banks' customers for its unique service. But this could not be real, let alone attracting others it can't even devise a good enough way to identify and reach its own book customers to make them the user of the visa card service.

Currently the bank does not have a clear classification scheme of its book and visa card customers or there are no clearly defined classes. But the employees try to promote the visa card service to book customers in a mass marketing way. In rare cases there is an attempt of direct marketing, but it is based on little and simple criteria. This is a clear indication of the fact that the company is not using its past dataset efficiently and effectively. Thus in order to plan and implement effective marketing strategies there is a need for actionable information which is obviously the result of a research works.

Accordingly, in the effort of improving the current situation by identifying factors, which have a strong link with being a visa card user, effective bank prediction model that can provide with unseen and hidden knowledge is very fundamental. For this purpose, timely and reliable data of

customers is crucial. This data helps to get the knowledge needed in identification of customers according to their likelihood to be visa card customer and hence to come up with the appropriate business strategies.

As a matter of fact, most analysis made by using traditional methods focus on problems with manageable number of variables and cases than may be encountered in real world, which has limited capacity to discover new and unanticipated patterns and relationships that are hidden in conventional databases [47]. Thus, in this research an attempt is made to come up with the appropriate model, which helps to predict the likelihood customer to be the visa card user. Specifically the model can successfully categorize each instance to the appropriate potential customers' group.

1.3 Justification of the Study

DASHEN BANK S.C. has a mission of providing efficient, customized, focused and international banking service to its customers. To accomplish its mission the bank should have well organized mechanisms to identify and serve its customers by studying their special character. Unfortunately the bank has tried to reach and convince people to use its new service in a more backward and customary ways. But the process of identifying and reaching the potential customers needs more than this simple traditional technique because of the huge amount of customers data and the critical nature of the process. By using data mining techniques the bank could simply identify the key characteristics and behaviors of each book customer and then predict the likelihood of that customer to be the visa card user. This intern helps the bank to acquire new customers by tailoring its promotional strategies and campaign in the way that targeted the particular potential customer.

Recent projects have indicated more than a 20-fold decrease in costs for targeted mailing campaigns over conventional approaches [33]. By predicting customers' behavior in advance, businesses can then market the right products to the right segments at the right time through the right delivery channels [41]. Consequently the bank will be in a better position to assign appropriate budget and manpower for the future expansion of its service. Generally it will help the bank to use more advanced techniques that helps to reach at measurable and actionable recommendations that support decision makers in strategic planning and decision-making.

On the other side this research will help other banking industries, that are trying to start the visa card service, as a lesson to what extent they should keep and analyze their customers' data to get the required hidden facts about their customers and what is the benefit of doing so. In addition to this, whenever the banks ability to expand their service increase, the integration of the society with global community through e-banking will also increase and consequently contributes to the globalization and the modernization process.

Because no attempt has been made so far to this level in identifying the appropriate numbers of clusters of the visa card customers and also knowing major determinants for being potential customer of visa card service in DASHEN BANK S.C, this research will be ground work for the effort of future works in the same area.

1.4 Objectives

1.4.1 General Objective

The main objective of this research is to explore the possible applications of data mining methods and techniques for potential visa card customer selection.

1.4.2 Specific Objectives

To achieve the general objective indicated above, the researcher has accomplished the following specific objectives:

1. Develop a very high level understand of the application domain.
2. Collect data on which the mining process is conducted.
3. Prepare the data for model building by selecting, cleaning and integrating it.
4. To explore the relationship between different variables (like monthly income, age, educational level and type of job) and being a potential visa card customer.
5. To come up with the appropriate number of clusters, this is based on the possibility of customers to be potential visa card user.
6. Build and test a predictive model that will help in the classification of book customers into one of the identified clusters.
7. Make conclusion and recommendations.

1.5 Research Methodology

1. 5.1 Literature Review

A literature review was conducted on the role of data mining technologies in transforming raw data into valuable information and applications of data mining for direct marketing purpose. In addition to this the researcher conducted a general literature review of related works. The literatures reviewed include journals, articles, internet resources, books, banking related thesis.

1.5.2 Data Collection and Preprocessing

Primary and secondary data were collected through discussion with the domain experts. Document analysis was made for the purpose of getting the visa card customers data, and for

extracting essential information regarding to different information need that are of value for the research.

The collected data (from the company's database) was stored in Microsoft excel. Then the researcher conducted the preprocessing steps on the data in order to improve the accuracy, efficiency and scalability of the clustering and classification process. The preprocessing step deals with missing values, exclusion of some attributes that are believed to have no use, like id and name of the customers, inclusion of attributes that are not explicitly stated in the table, which are important in the decision making process.

1. 5.3 Model Building and Experiments

In building a model the algorithm is expected to learn the different patterns in the dataset and this learned knowledge by the algorithm could be applied on the new datasets. The required data was collected from DASHEN BANK S.C. database. The researcher has followed the CRISP-DM.

The researcher used decision tree for the classification purpose and a clustering algorithm for segmentation. Specifically for the clustering purpose the researcher used K-Means clustering algorithm while for the classification purpose a decision tree algorithm called J48 was used.

In order to analyze the datasets and build the above type of model there were different available tools. From these tools the researcher used WEKA by considering the following factors:

- The algorithm supported (K-Means and J48)
- The data mining tasks that the tool intended to do(Clustering and Classification)
- Architecture and operating systems:- The hardware requirement on which the software run and MS window operating system

- The maximum number of records the software can comfortable handle
- The availability of the software
- The familiarity of the researcher with the software

Microsoft excel was used for storing and doing some preprocessing tasks.

1.6 Scope and Limitations

The scope of this research is strictly limited to assessing the possible application of data mining techniques for direct marketing purpose at the DASHEN BANK S.C. The exploration was totally done on the data collected from the visa card customers' database.

This study, like all others, is not without its limitations. First, the bank gave a very poor coordination and support to the researcher. Secondly, finding a good database which stores needed information for building customer response model was very hard. Also convincing managers to give the customers' information was another problem and hence this makes the data collection process hard. All the above reasons forced the researcher to confine himself only on the available attributes at that particular moment.

1.7 Significant of the Study

This research will help to show how data mining could be effectively used to come up with a good segmentation model that could serve the purpose of direct marketing campaign. Specifically, the DASHEN BANK S.C. can use the output of this research to be in a better position to answer questions like, what are the characteristics of customers that have high, medium or low likelihood to be the visa card user?

On the other hand, this research will provide various types of information to make vital decisions on developing and implementing marketing strategies to retaining profitable customers and attracting new ones. These all contributions make the bank to stay competent and hence profitable. Finally it could serve as a base for further researches that are mainly concerned with direct marketing.

1.8 Thesis Organization

This research consists of five chapters. The first chapter deals with a general background, motivation, statement of the problem, justification of the study, objectives, research methods, limitation of the study, and the possible applications of the research work. The second chapter deals with giving a general overview of data mining techniques. The third chapter deals with direct marketing and customer relation management and also the current status of direct marketing at the DASHEN BANK S.C. Chapter four, the most relevant chapter, discuss the different stages of experimentation toward building the data mining model and interpretation of results. The final chapter, chapter five, deals with giving of conclusion and recommendation based on the investigation of the study.

Chapter Two

Data Mining

2.1 Overview

To understand the term 'Data Mining' it is useful to look at the literal translation of the word: to mine in English means to extract. The verb usually refers to mining operations that extract from the Earth hidden, precious resources. The association of this word with data suggests an in-depth search to find additional information which is previously went unnoticed in the mass of data available. From the viewpoint of scientific research, data mining is a relatively new discipline that has developed mainly from studies carried out in other disciplines such as computing, marketing, and statistics [30].

Although data mining is relatively a new term, the technology is not new. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

Many of the methodologies used in data mining come from two branches of research, one developed in the machine learning community and the other developed in the statistical community, particularly in multivariate and computational statistics.

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses [5].

Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Most companies already collect and refine massive quantities of data.

Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought online.

2.2 Data Mining and Knowledge Discovery in Database

Historically the notion of finding useful patterns in data has been given a variety of names including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. The term data mining has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities. It has also gained popularity in the database field.

As defined by many scholars, data mining is the process of exploration and analysis of large quantities of data in order to discover meaningful patterns and rules from large amounts of data. Data Mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data [16].The data mining process is sometimes referred to as knowledge discovery or KDD (Knowledge Discovery

in Databases). The phrase knowledge discovery in databases was coined at the first KDD workshop in 1989 to emphasize that knowledge is the end product of a data-driven discovery [31]. It has been popularized in the AI and machine-learning fields. Knowledge Discovery in Databases refers to the overall process of discovering useful knowledge from data. There is a difference in understanding the terms "knowledge discovery" and "data mining" between people from different areas contributing to this new field.

In general KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data. Blind application of data-mining methods (rightly criticized as data dredging in the statistical literature) can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns.

Knowledge discovery in databases is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns/models in data. Whereas, data mining is a step in the knowledge discovery process consisting of particular data mining algorithms that, under some acceptable computational efficiency limitations, finds patterns or models in data.

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of

digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases [5].

Today the digital revolution has made digitized information easy to capture, process, store, distribute, and transmit [5]. With significant progress in computing and related technologies and their ever-expanding usage in different walks of life, huge amount of data of diverse characteristics continue to be collected and stored in databases. The rate at which such data are stored is growing phenomenally. We can draw an analogy between the popular Moore's law and the way data are increasing with the growth of information in this world of data processing applications. The advancement of data processing and the emergence of newer applications were possible, partially because of the growth of the semiconductor and subsequently the computer industry. According to Moore's law, the number of transistors in a single microchip is doubled every 18 months, and the growth of the semiconductor industry has so far followed the prediction. We can correlate this with a similar observation from the data and information domain. If the amount of information in the world doubles every 20 months, the size and number of databases probably increases at a similar pace. Discovery of knowledge from this huge volume of data is a challenge indeed. Data mining is an attempt to make sense of the information explosion embedded in this huge volume of data [16].

2.3 Data Mining and Related Fields

2.3.1 Data Warehouses

One of the global definitions of data warehouse is, it is a collection of integrated, subject-oriented databases designed to support the decision-support functions (DSF), where each unit of data is relevant to some moment in time [16].

Although the existence of a data warehouse is not a prerequisite for data mining, in practice, the task of data mining, especially for some large companies, is made a lot easier by having access to a data warehouse. A primary goal of a data warehouse is to increase the "intelligence" of a decision process and the knowledge of the people involved in this process. A data warehouse means different things to different people. Some definitions are limited to data; others refer to people, processes, software, tools, and data. One of the global definitions is that the data warehouse is a collection of integrated, subject-oriented databases designed to support the decision-support functions (DSF), where each unit of data is relevant to some moment in time. A data mart is a data warehouse that has been designed to meet the needs of a specific group of users. It may be large or small, depending on the subject area.

Data warehouse includes the following categories of data, where the classification is accommodated to the time-dependent data sources [16]:

1. Old detail data
2. Current (new) detail data
3. Lightly summarized data
4. Highly summarized data
5. Metadata (the data directory or guide).

To prepare these five types of elementary or derived data in a data warehouse, the fundamental types of data transformation are standardized. There are four main types of transformations, and each has its own characteristics:

1. Simple transformations
2. Cleansing and scrubbing

3. Integration

4. Aggregation and summarization

These transformations are the main reason why we prefer a warehouse as a source of data for a data-mining process. If the data warehouse is available, the preprocessing phase in data mining is significantly reduced, sometimes even eliminated. So that, we can avoid the most time consuming phase.

2.3.2 Data Mining and Online Analytical Processing

A popular approach for analysis of data warehouses is called online analytical processing (OLAP), named for a set of principles proposed by Codd in the year 1993 [42]. OLAP tools focus on providing multidimensional data analysis, which is superior to SQL in computing summaries and breakdowns along many dimensions. OLAP tools are targeted toward simplifying and supporting interactive data analysis, but the goal of KDD tools is to automate as much of the process as possible. Thus, KDD is a step beyond what is currently supported by most standard database systems.

One of the most common questions from data processing professionals is about the difference between data mining and OLAP [9]. As could be understood from this literature, they are very different tools that can complement each other. OLAP is part of the spectrum of decision support tools. Traditional query and report tools describe what is in a database. OLAP goes further; it's used to answer why certain things are true.

The user forms a hypothesis about a relationship and verifies it with a series of queries against the data. For example, an analyst might want to determine the factors that lead a person to commit a crime. He or she might initially hypothesize that people who watch films that promote

bad character are highly exposed to do crime and analyze the database with OLAP to verify (or disprove) this assumption. If that hypothesis were not borne out by the data, the analyst might then look at other factors like their childhood environment as the determinant factor to commit a crime. If the data did not support this guess either, he or she might then try both bad films and unfavorable childhood environment are the determinant factor.

In other words, the OLAP analyst generates a series of hypothetical patterns and relationships and uses queries against the database to verify them or disprove them. OLAP analysis is essentially a deductive process. But what happens when the number of variables being analyzed is in the dozens or even hundreds. It becomes much more difficult and time-consuming to find a good hypothesis (let alone be confident, there is no better explanation than the one found), and analyze the database with OLAP to verify or disprove it.

Data mining is different from OLAP because rather than verify hypothetical patterns, it uses the data itself to uncover such patterns. It is essentially an inductive process. For example, suppose the analyst who wanted to identify the factors that lead a person to commit a crime were to use a data mining tool. The data mining tool might discover that people who watch bad films and live in unfavorable childhood environment are highly suspected to commit a crime. But it might go further and also discover a pattern the analyst did not think to try, such as that age is also a determinant of risk.

On the other way, OLAP is complementary in the early stages of the knowledge discovery process because it can help to explore the data, for instance by focusing attention on important variables, identifying exceptions, or finding interactions. This is important, because the better you understand your data, the more effective the knowledge discovery process will be [5].

2.3.3 Data Mining, Machine Learning and Statistics

Most data-mining problems and corresponding solutions have roots in classical data analysis. Data mining has its origins in various disciplines, of which the two most important are statistics and machine learning.

Data mining takes advantage of advances in the fields of artificial intelligence (AI) and statistics. Both disciplines have been working on problems of pattern recognition and classification. Both Communities have made great contributions to the understanding and application of neural nets and decision trees. Data mining does not replace traditional statistical techniques. Rather, it is an extension of statistical methods that is in part the result of a major change in the statistics community. The development of most statistical techniques was, until recently, based on elegant theory and analytical methods that worked quite well on the modest amounts of data being analyzed.

The key point is that data mining is the application of these and other AI and statistical techniques to common business problems in a fashion that makes these techniques available to the skilled knowledge worker as well as the trained statistics professional. Data mining is a tool for increasing the productivity of people trying to build predictive models.

2.4 Data Mining Tasks and Functionalities

Several data mining problem types or analysis tasks are typically encountered during a data mining project. Depending on the desired outcome, several data analysis techniques with different goals may be applied successively to achieve a desired result.

Data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database while Predictive mining tasks Perform inference on the current data in order to make predictions [16].

Based on the different mining tasks, we can categorize data mining functionalities (methods) as classification, clustering, regression, association rules, sequence discovery, prediction, and so on [9]. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks [16]. Figure 2.1 shows the data mining methods.

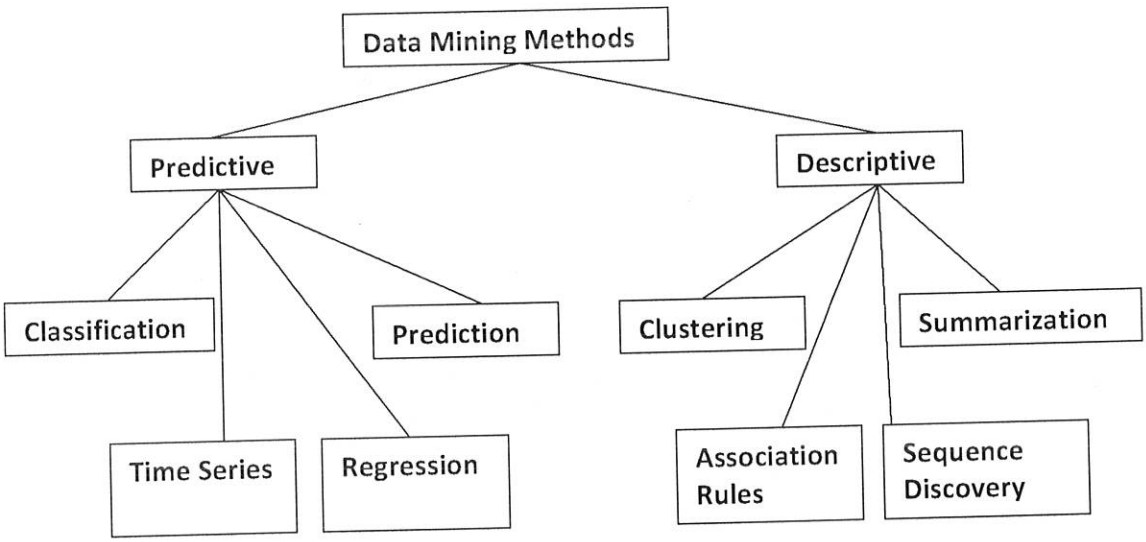

Figure 2. 1: Data Mining Functionalities [9]

Basic data mining functionalities are: Classification, Estimation, Prediction, Affinity grouping or associating rules, Clustering, Description and visualization [25]. The first three are all examples of directed data mining, where the goal is to find the value of a particular target variable. Affinity grouping and clustering are undirected tasks where the goal is to uncover structure in data

without respect to a particular target variable. Profiling is a descriptive task that may be either directed or undirected.

The most known data mining methods are [16]:

1. Clustering (descriptive)
2. Classification (predictive) and Regression (predictive)
3. Association rule discovery (descriptive) and Sequential Pattern Discovery (predictive)

2.4.1 Clustering

Clustering is a technique that puts similar entities into the same groups based on similar data characteristics and those with dissimilar entities are put in different groups. Similarity is measured according to a distance measure function. The meaning of the clusters is therefore dependent on the distance function used. Thus, clustering always requires significant involvement from a business or domain expert who needs to both propose an appropriate distance measure and to judge whether the clusters are useful.

Unlike classification, we don't know what the clusters will be when we start, or by which attributes the data will be clustered. That is why we need some one who is knowledgeable in the business must interpret the clusters. After we have found clusters that reasonably segment our database, these clusters may then be used to classify new data. Some of the common clustering techniques include EM (Expectation Maximization), K-Nearest Neighbors (K-NN), a special type of neural network called Kohonen net or self-organizing maps (SOM).

2.4.1.1 K-Nearest Neighbors

K-nearest neighbor is a predictive technique suitable for classification models. K represents a number of similar cases or the number of items in a group. With the K-NN technique, the training data is the model. When a new case or instance is presented to the model, the algorithm looks at all the data to find a subset of cases that are most similar to it and use them to predict the outcome.

There are two principal parameters in the K-NN algorithm:

1. The number of nearest cases to be used (K)
2. A metric to measure the similarity

Each use of the K-NN algorithm requires that a positive integer value for K is specified. These determine how many existing cases are looked at when predicting a new case. For example, 4-NN indicates that the algorithm will use the four nearest cases to predict the outcome of a new case.

2.4.1.1.1 The K-Nearest Neighbors Algorithm

K-NN decides into which class to place a new case by examining some number (the K) of the most similar cases or neighbors. Most of the time K ranges from 2 to 20. The algorithm computes the distance from the new case to each case in the training data. The new case is predicted to have the same outcome as the predominant outcome in the K closest cases in the training data. So, the new case is as signed to the same class to which most of the similar cases belong.

K-NN is based on a concept of distance, and this requires a metric to determine distances. For continuous attributes Euclidean distance can be used, for categorical variables, one has to find a

suitable way to calculate the distance between attributes in the data. Choosing a suitable metric is a very delicate task because, different metrics, used on the same training data, can result in completely different predictions. This means that a business expert is needed to help determine a good metric.

2.4.1.2 Self Organizing Map

When the set of inputs is multi-dimensional, traditional clustering algorithms do not offer an easy way to visualize the "closeness" of other clusters. A self-organizing map(SOM) or Kohonen feature map is a special kind of neural network architecture that provides a mapping from the multi-dimensional input space to a lower-order regular lattice of cells (typically 2 dimensional grid). Such a mapping is used to identify clusters of elements that are similar (in a Euclidean sense) in the original space.

In SOM, the clusters are usually organized into a lattice of cells, usually a two- dimensional grid but also one-dimensional or multi-dimensional. The grid exists in a space that is separate from the input space; any number of input features may be used as long as their number is greater than the dimensionality of the grid space. SOM tries to find clusters such that any two clusters that are close to each other in the grid space have cluster close to each other in the input space. But the converse does not hold: cluster centroids that are close to each other in the input space do not necessarily correspond to clusters that are close to each other in the grid.

2.4.2 Classification and Regression

Represent the largest part of problems to which data mining is applied today, creating models to predict class membership (classification) or a value (regression). Classification is used to predict what group a case belongs to. Regression is used to predict a value of a given continuous valued

variable based on the values of other variables, assuming a linear or non-linear model of dependency. Logistic regression is used for predicting a binary variable. It is a generalization of linear regression, because the binary dependent variable cannot be modeled directly by linear regression. Logistic regression is a classification tool when used to predict categorical variables such as whether an individual is likely to purchase or not, and a regression tool when used to predict continuous variables such as the probability that an individual will make a purchase. There are several classification and regression techniques including decision trees, neural networks.

2.4.2.1 Decision Tree

Decision trees are powerful and popular tools for classification and prediction. They are attractive due to the fact that in contrast to other machine learning techniques such as neural networks, they represent rules that human beings can understand. Decision tree is a classifier in the form of a tree structure, where each node is either a leaf node, indicating the value of the target attribute or class of the examples, or a decision node, specifying SOM test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test. A decision tree can be used to classify by starting at the root of the tree and moving through it until a leaf node is reached, which provides the classification of the instance. Figure 2.2 shows the possible decisions made during tree construction step.

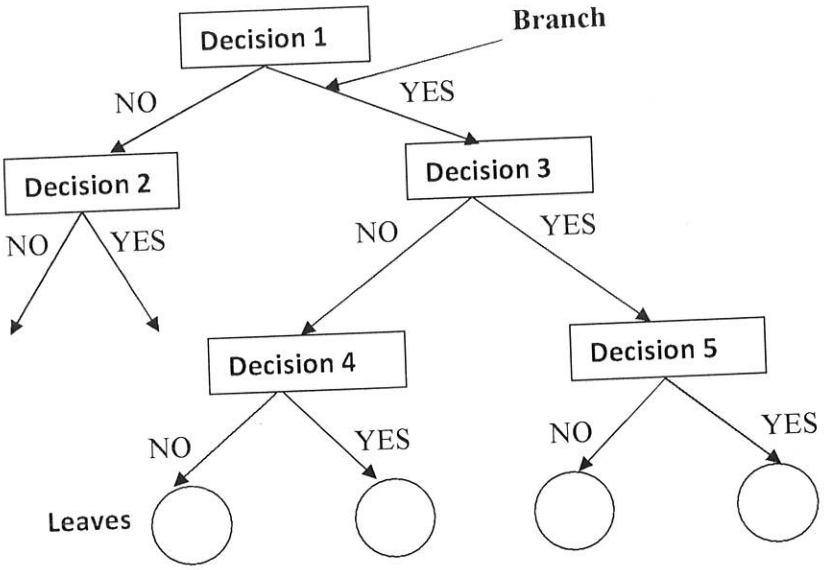

Figure 2. 2: Decision Tree

Decision trees represent a set of decisions. These decisions generate rules for classification of a dataset using the statistical criterion: entropy, information gain, Gini index, chi-square test, measurement error, classification rate, etc. There are two stages, tree construction and post-pruning, and five tree algorithms are in common use CART, CHAID, ID3, C4.5 and C5.0. Most algorithms that have been developed for learning decision trees are variations on a core algorithm that employs a top-down, greedy search through the space of possible decision trees [48].

A decision tree can be grown until every node is pure, i.e., the leaf nodes can be divided no further and the members within each leaf node belong to only one class. A maximal classification tree gives 100% accuracy on training data, but it is a result of over fitting and would give poor prediction on test data. Tree complexity is a function of the number of leaves, the number of splits and the depth of the tree.

A well-fitted tree has low bias and low variance. To avoid over fitting a tree needs to be right sized by either forward-stopping or stunting the growth or growing the tree to its full length and then pruning it back [30].

2.4.2.1.1 Tree Induction

The training process that creates the decision tree is called induction and requires a small number of passes through the training set. As described above most decision tree algorithms go through two phases: a tree growing (splitting) phase followed by a pruning phase.

- **Splitting: -** The first iteration considers the root node that contains all the data. Subsequent iterations work on derivative nodes that will contain subsets of the data. At each split, the variables are analyzed and the best split is chosen: The tree growing phase is an iterative process which involves splitting the data into progressively smaller subsets. One important characteristic of splitting is that it is greedy, which means that the algorithm does not look forward in the tree to see if another decision would produce a better overall result.
- **Stopping criteria**: - Tree-building algorithms usually have several stopping rules. These rules are usually based on several factors including maximum tree depth, minimum number of elements in a node considered for splitting, or its near equivalent, the minimum number of elements that must be in a new node. In most implementations the user can alter the parameters associated with these rules. Some algorithms, in fact, begin by building trees to their maximum depth. While such a tree can precisely predict all the instances in the training set (except conflicting records), the problem with such a tree is that, more than likely, it has over fit the data.
- **Pruning:** - After a tree is grown, one can explore the model to find out nodes or sub trees that are undesirable because of over fitting or rules that are judged inappropriate. Pruning

removes splits and the sub trees created by them. Pruning is a common technique used to make a tree more general.

2.4.2.1.2 Understanding the Output

Once trained, a tree can predict a new data instance by starting at the top of the tree and following a path down the branches until encountering a leaf node. The path is determined by imposing the split rules on the values of the independent variables in the new instance. A decision tree can help a decision maker identify which factors to consider and how each factor has historically been associated with different outcomes of the decision.

Decision trees have obvious value as both predictive and descriptive models. Prediction can be done on a case-by-case basis by navigating the tree. More often, prediction is accomplished by processing multiple new cases through the tree or rule set automatically and generating an output file with the predicted value or class appended to the record for each case. Many implementations offer the option of exporting the rules to be used externally or embedded in other applications.

2.4.2.1.3 Different Decision Tree Algorithms

Decision tree algorithms commonly implemented includes, chi-squared Automatic Interaction Detection (CHAID), Classification and Regression Trees (CART), C4.5 and C5.0 [35]. All are well suited for classification; some are also adaptable for regression. The distinguishing features between tree algorithms include the following:

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

2.4.2.2 Artificial Neural Networks

Artificial neural networks (ANN) are among the most complicated of the classification and regression algorithms. They are often considered as a black box. Neural networks require a lot of data for training, thus consuming time, but once trained, it can make predictions for new cases very quickly, even in real time. Moreover, neural networks can provide multiple outputs representing multiple simultaneous predictions. A key feature of neural nets is that they only operate directly on numbers. As a result, any nonnumeric data in either the independent or dependent (output) columns must be converted to numbers, e.g. variables with " yes/no " values must be replaced by " 0/ 1 ".

2.4.3 Association and Sequential Pattern Discovery

Tools analyze data to discover rules that identify patterns of behavior, e. g. what products or services customers tend to purchase at the same time, or later on as follow-up purchases. While these approaches had their origins in the retail industry, they can be applied equally well to services that develop targeted marketing campaigns or determine common (or uncommon) practices. In the financial sector, association approaches can be used to analyze customers' account portfolios and identify sets of financial services that people often purchases together. They may be used, for example, to create a service "bundle" as part of a promotional sales campaign. Market basket analysis is a good example of application of association rule mining.

Chapter Three

Customer Relation Management and Direct Marketing

3.1 Customer Relation Management

CRM is a combination of policies, processes, and strategies implemented by a company that unify its customer interaction and provides a mechanism for tracking customer information. CRM includes many aspects which relate directly to one another:

- Front office operations: — Direct interaction with customers, e.g. face to face meetings, phone calls, e-mail, online services etc.
- Back office operations: — Operations that ultimately affect the activities of the front office (e.g., billing, maintenance, planning, marketing, advertising, finance, manufacturing, etc.)
- Business relationships: — Interaction with other companies and partners, such as suppliers/vendors and retail outlets/distributors, industry networks (lobbying groups, trade associations). This external network supports front and back office activities.
- Analysis: — Key CRM data can be analyzed in order to plan target-marketing campaigns, conceive business strategies, and judge the success of CRM activities (e.g., market share, number and types of customers, revenue, and profitability).

Customer Relationship Management emerged in the last decade to reflect the central role of the customer for the strategic positioning of a company. CRM takes a holistic view over customers. It encompasses all measures for understanding the customers and for exploiting this knowledge to design and implement marketing activities, align production and coordinate the supply-chain.

CRM puts emphasis on the coordination of such measures, also implying the integration of customer-related data, meta-data and knowledge and the centralized planning and evaluation of measures to increase customer lifetime value. CRM gains importance for companies that serve multiple groups of customers and exploit different interaction channels for them. This is due to the fact that information about the customers, which can be acquired for each group and across any channel, should be integrated with existing knowledge and exploited in a coordinated fashion [19].

However according to many literatures, CRM is a broadly used term and covers a wide variety of functions not all of which require data mining. These functions include marketing automation (e.g., campaign management, cross- and up-sell, customer segmentation, customer retention), sales force automation (e.g., contact management, lead generation, sales analytics, generation of quotes, product configuration), and contact center management (e.g., call management, integration of multiple contact channels, problem escalation and resolution, metrics and monitoring, logging interactions and auditing), among others.

CRM has a strong relationship with direct marketing. It is an innovative element that moves and broadens the knowledge of direct marketing and thus management's knowledge, which generates value for companies and proposes tools that the administrators can employ to increase the organizations' competitiveness. Moreover, the correct implementation of CRM processes as a strategy of massive personalization optimizes the efforts of direct Marketing in companies basing themselves in the development of customer loyalty and knowledge, creating a change in the culture of customer management which generates sustained profitability in the organization [29]. A problem with CRM is that CRM means different things to different people. For some, CRM

means direct e-mails. For others, it is mass customization or developing products that fit individual customers' needs. For IT consultants, CRM translates into complicated technical jargon related to terms like OLAP (on-line analytical processing) and CICs (customer interaction centers).

3.1.1 Principles of Customer Relation Management

The overall processes and applications of CRM are based on the following basic principles [29]:

- **Treat Customer Individually**

Remember customers and treat them individually. CRM is based on philosophy of personalization. Personalization means the content and services to customer should be designed based on customer preferences and behavior.

- **Acquire and Retain Customer Loyalty through Personal Relationship**

Once personalization takes place, a company needs to sustain relationships with the customer. Continuous contacts with the customer especially when designed to meet customer preferences – can create customer loyalty.

- **Select "Good" Customer instead of "Bad" Customer based on Lifetime Value**

Find and keep the right customers who generate the most profits. Through differentiation, a company can allocate its limited resources to obtain better returns. The best customers deserve the most customer care; the worst customers should be dropped. In summary, personalization, loyalty, and lifetime value are the main principles of CRM implementation.

3.1.2 Component of Customer Relation Management

The basic model contains a set of 7 basic components [37]:

A. A database of customer activity

B. Analyses of the database

C. Given the analyses, decisions about which customers to target

D. Tools for targeting the customers

E. How to build relationships with the targeted customers

F. Privacy issues

G. Metrics for measuring the success of the CRM program

A. Creating Customers Database

A necessary first step to a complete CRM solution is the construction of customers database or information file. This is the foundation for any customer relationship management activity. For existing companies that have not previously collected much customer information, the task will involve seeking historical customer contact data from internal sources such as accounting and customer service. The database should contain information about the following:

1. Transactions
2. Customer contacts
3. Descriptive information
4. The data should also be over time

B. Analyzing the Data

Traditionally, customer databases have been analyzed with the intent to define customer segments. A variety of multivariate statistical methods ranging such as cluster and discriminant

analysis have been used to group together customers with similar behavioral patterns and descriptive data which are then used to develop different product offerings or direct marketing campaigns. Direct marketers have used such techniques for many years. Their goals are to target the most profitable prospects for catalogue mailings and to tailor the catalogues to different groups. More recently, such segmentation approaches have been heavily criticized. As a result, a new term, lifetime customer value or LCV, has been introduced into the lexicon of marketers.

C. Customer Selection

Given the construction and analysis of the customer information contained in the database, the next step is to consider which customers to target with the firm's marketing programs. The results from the analysis could be of various types. If segmentation-type analyses are performed on purchasing or related behavior, the customers in the most desired segments (e.g., highest purchasing rates, greatest brand loyalty) would normally be selected first. Other segments could also be chosen depending upon additional factors. For example, if the customers in the heaviest purchasing segment already purchase at a rate that implies further purchasing is unlikely, a second tier with more potential would also be attractive. The descriptor variables for these segments (e.g., age, industry type) provide information for deploying the marketing tools. In addition, these variables could be matched with commercially-available databases of names to find additional customers matching the profiles of those chosen from the database.

D. Targeting the Customers

Mass marketing approaches such as television, radio, or print advertising are useful for generating awareness and achieving other communications objectives, but they are poorly-suited for CRM due to their impersonal nature. More conventional approaches for targeting selected

customers include a portfolio of direct marketing methods such as telemarketing, direct mail, and, when the nature of the product is suitable, direct sales

E. Relationship Programs

While customer contact through direct e-mail offerings is a useful component of CRM, it is more of a technique for implementing CRM than a program itself. Relationships are not built and sustained with direct e-mails themselves but rather through the types of programs that are available for which e-mail may be a delivery mechanism.

The overall goal of relationship programs is to deliver a higher level of customer satisfaction than competing firms deliver. There has been a large volume of research in this area. From this research, managers today realize that customers match realizations and expectations of product performance, and that it is critical for them to deliver such performance at higher and higher levels as expectations increase due to competition, marketing communications, and changing customer needs. Thus, managers must constantly measure satisfaction levels and develop programs that help to deliver performance beyond targeted customer expectations.

A comprehensive set of relationship programs comprise of the followings:

- Customer service
- Frequency/loyalty programs
- Customization
- Rewards programs
- Community building

F. Privacy Issues

The CRM system depends upon a database of customer information and analysis of that data for more effective targeting of marketing communications and relationship-building activities. There

is an obvious tradeoff between the ability of companies to better deliver customized products and services and the amount of information necessary to enable this delivery. Particularly with the popularity of the Internet, many consumers and advocacy groups are concerned about the amount of personal information that is contained in databases and how it is being used.

G. Metrics

The increased attention paid to CRM means that the traditional metrics used by managers to measure the success of their products and services in the marketplace have to be updated. Financial and market-based indicators like profitability, market share, and profit margins have been and will continue to be important. However, in a CRM world, increased emphasis is being placed on developing measures that are customer-centric and give the manager a better idea of how CRM policies and programs are working. Some of these CRM-based measures, both web and non-web based are:

- Customer acquisition costs
- Conversion rates (from lookers to buyers)
- Retention/churn rates
- Same customer sales rates
- Loyalty measures.
- Customer share or share of requirements

All of these measures imply doing a better job acquiring and processing internal data to focus on how the company is performing at the customer level.

3.1.3 Customer Relation Management Issues

3.1.3.1 Customer Privacy

Customer privacy is an important issue in CRM. It deals with large amounts of customer data through various touch points and communication channels. The personalization process in CRM requires identification of each individual customer and collections of demographic and behavioral data. Yet, it is the very information that most customers consider personal and private [29].

Firms want to collect as much information as possible about each customer to further its sales, yet in doing so it treads at and beyond the bounds of personal privacy. Privacy issues are not simple. There are overwhelming customer concerns, legal regulations, and public policies around the world. Still it is unclear and undetermined what extent of customer privacy should be protected and shouldn't be used, but the following four basic rules might be considered [46].

- The customer should be notified their personal information is collected and will be used for specific purposes
- The customer should be able to decline to be tracked
- The customer should be allowed to access their information and correct it
- Customer data should be protected from unauthorized usage

Some companies try to ask the customer to agree to information collection and usage. Providing personalized service to customer is a way to satisfy customers who provided their personal information. All of these efforts are designed to build trust between the company and its customers.

3.1.3.2 Technical Immaturity

The concept, technologies, and understanding of CRM are still in its early adapter stage. Most of the CRM technologies are immature and the typical implementation costs and time are long enough to frustrate potential users. Many software and hardware vendors sell themselves as complete CRM solution providers but there are little standardized technologies and protocols for CRM implementation in the market [29].

Moreover the scope and extent of 'what CRM includes' differ from vendor to vendor; each has different implementation requirements to achieve the customer's expectations. CRM is one of the busiest industries which occurs frequent merger and acquisition. Many small companies merge together to compete with large vendor. Most of the time, the technical immaturity together with unclear customer requirement leads the project to failure. These technical immaturities may be overcome overtime, but the process may be long and painful.

3.1.4 Customer Segmentation

Segmentation is a way to have more targeted communication with customers. The process of segmentation describes the characteristics of the customer groups (called segments or clusters) within the data. Segmenting means, putting the population into segments according to their affinity or similar characteristics. Customer segmentation is a preparation step for classifying each customer according to the customer groups that have been defined.

Segmentation is essential to cope with today's dynamically fragmenting consumer marketplace. By using segmentation, marketers are more effective in channeling difficulties in making good segmentation [49]. Mainly this segmentation takes into consideration the followings:

Relevance and quality of data: - These are essential to develop meaningful segments. If the company has insufficient customer data or too much data, that can lead to complex and time-consuming analysis. If the data is poorly organized (different formats, different source systems) then it is also difficult to extract interesting information. Furthermore, the resulting segmentation can be too complicated for the organization to implement effectively. In particular, the use of too many segmentation variables can be confusing, resulting in segments which are unfit for management decision- making. Alternatively, apparently effective variables may not be identifiable. Many of these problems are due to an inadequate customer database.

Intuition: - Although data can be highly informative, marketers need to continuously develop segmentation hypotheses in order to identify the ` right ' data for analysis.

Continuous Process: - Segmentation demands continuous development and updating as new customer data is acquired. In addition, effective segmentation strategies will influence the behavior of the customers affected by them; thereby necessitating revision and reclassification of customers. Moreover, in an e-commerce environment where feedback is almost immediate, segmentation would require almost a daily update.

Over-Segmentation: A segment can become too small or insufficiently distinct to justify treatment as separate segments. The data mining methods used for customer segmentation belong to the category of clustering or K-nearest-neighbors algorithms.

3.2 Direct Marketing

3.2.1 What is Direct Marketing?

Direct marketing is any unsolicited contact a business unit makes with existing or potential customers in order to generate sales or raise awareness. For many businesses, it's by far the most cost-effective form of marketing. From direct mail and leaflet drops to telemarketing and email marketing, it allows to target customers with greater accuracy than any other method available.

Sometimes direct marketing is a controversial sales method by which businesses approach potential customers directly with products or services with out any intermediary. Some of the most common forms of direct marketing are telephone sales, solicited or unsolicited emails, catalogs, leaflets, brochures and coupons. Successful direct marketing also involves compiling and maintaining a large database of personal information about potential customers and clients. These databases are often sold or shared with other direct marketing companies.

For many companies or service providers with a specific market, the traditional forms of advertising (radio, newspapers, television, etc.) may not be the best use of their promotional budgets. For example, a company which sells a hair loss prevention product would have to find a radio station whose format appealed to older male listeners who might be experiencing this problem. There would be no guarantee that this group would be listening to that particular station at the exact time the company's ads were broadcast. Money spent on a radio spot (or television commercial or newspaper ad) may or may not reach the type of consumer who would be interested in a hair restoring product [46].

Direct marketing works best when the recipients accept the fact that their personal information might be used only for the agreed upon purpose. Some customers prefer to receive targeted

catalogs which offer more variety than a general mailing.

3.2.2 Major Uses of Direct Marketing

Direct marketing allows businesses to generate a specific response from targeted groups of customers. It's a particularly useful tool for small businesses because it allows to focus limited resources where they are most likely to produce results so that it is easy to target a representative sample of the target audience and see what delivers the best response rates before any attempt to develop a full campaign. A direct marketing campaign can help you to achieve the following key objective [47]:

- Building customer loyalty
- Increasing sales to existing customers
- Re-establishing lapsed customer relationships generating new business

Direct marketing can be used in both business-to-business and consumer markets. Of course, the strategy will need to be modified depending on the target receptive to mail shots or telemarketing calls. Writers such as Peppers and Rogers urged companies to begin to dialogue with their customers through these targeted approaches rather than the usual one way communication channel. In particular, the new mantra, "1-to-1" marketing, has come to mean using the Internet to facilitate individual relationship building with customers. An extremely popular form of Internet-based direct marketing is the use of personalized e-mails [37].

However, it showed be noted that the results of direct marketing aren't guaranteed. A poorly planned and executed targeted campaign will be a waste of money. A badly designed mail shot, for example, could simply end up in the bin. And worse still, it may irritate recipients and damage the business' reputation as a result.

3.2.3 Target Selection in Direct Marketing

Direct marketing has become an important application field for data mining. As explained above in direct marketing, companies or organizations try to establish and maintain a direct relationship with their customers in order to target them individually for specific product offers or fund raising. Apart from commercial firms and companies, charity organizations also apply direct marketing for fund raising. In fact, nowadays more and more companies are using the information about their customers' preference and behavior, which is provided by their databases, to do this kind of marketing. Moreover, many companies are using this type of relationship as their main strategy for interacting with their customers [15].

Kaymak [22], potential customers could be selected by analyzing data from previous campaigns or by organizing test mail campaigns from which models can be generated to select the customers who will be targeted. Usually data from previous campaigns is used. What group of customers should then be targeted? Based on what facts will the company address one group of customers instead of another?

Large databases of customer and market data are maintained for this purpose. In a specific campaign, the customers or supporters to be targeted are selected from the database, given different types of information such as demographic information or personal characteristics like profession, age and purchase history.

Target selection is an important data mining problem from the world of direct marketing. In fact, the answer to how could the company prefer one group of customer than the other is the goal of target selection: determine the potential customers for a new product by identifying profiles of customers that are known to have shown interest in a product in the past; that is, the generation

of customer models (profiles) for a given product by analyzing customers' data obtained from similar previous marketing campaigns [22].

The purpose of target selection in general is the selection of those customers who will be most interested in a particular product offer. So that as large percentages as possible of the targeted customers respond to the product offer. The key to target selection is to maximize the profits of selling the product and minimize the cost of the marketing campaign.

Many techniques have been applied to select the targets in commercial applications, such as decision tree methods like CHAID or CART (Haughton and Oulabi ,1993), statistical regression (Wansbeek, 1993), neural computing (Zahavi and Levin ,1997) and fuzzy clustering (Setnes and Kaymak, 2001) [40].

3.2.3.1 Target Selection Methods

The methods for target selection can be divided into two main groups [22]:

1. Segmentation methods: - The segmentation approach divides the customers' database into segments with similar properties. Segmentation based target selection models thus divide the customers into several groups depending on similarity of relevant features. An estimate of the response percentage for each group can be computed given the training data available. The customers within the groups that have a higher response percentage are then selected for targeted offers, i.e. they are sent a product offer by mail or otherwise.

2. Scoring methods: - The scoring approach assigns a separate score to each individual customer, and is interesting for tailoring the marketing campaign to individual customers. The score is indicative of the likelihood of response of the customer. The customers are then ordered

As part of relationship marketing programs, marketing executives are taking advantages of vast quantities of customer data. Models commonly used in the direct marketing arena to predict response to mailings and other forms of direct marketing promotions are increasingly being used to up-sell or cross-sell customers who contact companies through call centers. For example, the models can be used to decide which of various possible products or services to offer the customer based on a predicted probability of accepting an offer that is estimated on the fly from data already available on the customer or obtained with a couple questions. A class of such models is called response models, in which the dependent variable is a simple response or not [31].

Response modeling for database marketing is concerned with the task of modeling the customers purchasing behavior. The information at the level of the individual consumer is typically used to construct a response score.

Response modeling is a well known technique commonly used by direct marketing analysts [8]. It has proven to be a profitable tool in fine-tuning direct marketing strategies, since even small improvements attributed to modeling can create great financial gains [47]. The substantive relevance of response modeling comes from the fact that an increase in response of only one percentage point can result in substantial profit increases [45].

Given a tendency of rising mailing costs and increasing competition, the importance of response modeling increased. Improving the targeting of the offers may indeed counter these challenges by lowering non response.

customer. Finally the card directly goes back to the particular branch and the branch gives the card together with a pin number, which is only known by the customer, to the appropriate customer. The bank has a help desk that solves customers' problems which are concerned with the visa card usage or malfunction of machines. In the case where customers forget their pin number they will be given another pin number.

With the expansion of the banking industry in Ethiopia, the DASHEN BAK S.C. needs to improve its old and backward way of handling relationship with its customers. Particularly in the case of direct marketing the bank has to do a lot of works which help it in the following ways:

- Can maximize customers' response to a product offering
- Minimize the overall marketing cost
- Improve customer relationship management.

All these benefits can be achieved by using the direct marketing techniques properly to their customers' data.

3.3.2 Direct Marketing Programs

In the developed countries, banks like the Imperial Bank of Canada are making use of target selection to identify potential customers for a new product. As being repeatedly indicted the main purpose of this research is to build a model that helps to classify book customers according to their probability of being a visa card user. To accomplish this purpose a clustering and predictive models using customer demographic and financial data were built.

One of the main benefits of direct marketing is to select a target in a cost effective way. The key idea is to avoid investing in a scattershot means of advertising. Companies with a specific type of

marketing campaigns and select the appropriate marketing channel and advertising for the campaign. It is then possible to target those customers most likely to exhibit the desired behavior by creating predictive models.

But most of the time in contrary to this idea of direct marketing the bank uses mass marketing strategies, which is believed to be obsolete. Thus one can easily observe the absence of automated system in the bank makes the journey to the world of direct marketing very much difficult. If the bank starts to use the data mining techniques for the purpose of direct marketing campaign, it can build strong marketing strategy that helps the bank to meet its objectives in a cost effective means.

3.4 Review of Related Works

In recent year different researches made on the customer relation management are increasing in alarming rate. Currently CRM is applied in enhancing different business issue like customer identification, segmentation, customer differentiation, customization and many more. In this part the researcher has tried to review some of the researches conducted in the department of information science at Addis Ababa University and elsewhere in areas of CRM that employed decision tree and clustering as a technique.

The work of Henock [17], entitled "Application of Data Mining Techniques to Support Customer Relation Management at Ethiopian Air Lines" is one of the works in the area of CRM. He attempted to study the possible application of data mining techniques to enhance business productivity. He used decision tree and clustering techniques, to support CRM at ETHIOPIAN AIR LINES Corporation. The research was conducted in five major phases, namely business understanding, data understanding, data preparation, model building and evaluation. He applied a

K-Means clustering algorithm to segment the customer data into meaningful groups. The study seems to validate the business norm that customer value is based on the 'Total Revenue' contribution. The decision tree model that was generated from the cluster results correctly assigned 92.18% of new records to five clusters with 'Total Revenue' as the splitting variable. This result was found to be lower than the result obtained with 'Total Trip' and 'Tenure in Months' as splitting variable, but the he said that "since the difference was quite negligible, the decision tree model with 'Total Revenue' making the initial split was chosen as a working model" but the researcher believes that, there is no enough reason to ignore other splitting variables as far as they can bring better accuracy.

Generally the results from his study were encouraging, because it shows that it is possible to segment customer into different clusters that made business sense by using the K-Means clustering algorithm. Finally, the researcher do believe that, the research showed that knowledge of data mining techniques, marketing strategies and business companies should be integrated to successfully implement CRM. As a recommendation, he suggested the importance of further data mining projects by including more demographic data.

The other research was conducted by Kumneger on customer relation management for Ethiopian shipping line [23]. In her study the entire population was used to train the clustering model whereas for the decision tree model 60% was used for training the model and 30% served as a test data and the remaining data 10% was employed as validation set. As she pointed initially, her major objective was to segment customer into similar groups based on their revenue generating behavior. At the end, the result reveals that it is possible to segment customers based on their profitability and hence long term potential to generate revenue. The Cross Industry

Standard Process for Data Mining (CRISP-DM) model was followed to complete the data mining task.

The researcher believes that, the result of her research seems to be acceptable because the expert at the ESL (Ethiopian Shipping Line) observe and appreciate the result. This shows how the model accurately handles the business cases. In the experimentation part she used different values for K, which were 3, 4 and 5. She got accuracy of 98.37 %, 98.62% and 97.88% for 3, 4, 5 values of K respectively. On the other hand, for the validation set the accuracy was 98.45%, 98.55% and 97.72% accordingly for 3, 4 and 5 value of K. As could have been understood from the result, the decision tree and the K-Means clustering were good enough to segment and predict instances accurately. Finally, she emphasized the importance of taking further data mining researches in different area with different tools and techniques.

Still other research was done by Hiwot on applying data mining tools and techniques for effective CRM at Ethiopia hotel [18]. In her research different CRM concepts were revised and one can understand that CRM is the best marketing strategy for acquiring, retaining and partnering with selected customers to create a high value for the hotel industry. She applied data mining process and principles on customers data obtained from the Ethiopian Hotel. For the purpose of data understanding, data preprocessing, and modeling data mining software called Knowledge STUDIO was used. She used the K-Means clustering algorithm and J48 decision tree algorithm. For the value of K different numbers were attempted i.e. K=6, K=5, K=4. The best result was found when K = 5, in that case all the five clusters were different and meaningful. At last, the decision tree was used to classify the customers into one of the five clusters and the splitting attribute selected is the 'Room Revenue' because it scores a test accuracy of 96.72%,

which is greater than any other attribute.

Other important works, particularly those involves in application of data mining in the Banks are the work of Meretework [27] and Askale [2]. Both showed that the application of data mining in the Bank industry is invaluable and further researches are recommended by both of them.

On the other hand, there are many researches made in different part of the world concerning the application of data mining techniques for effective CRM. "Response Modeling in Direct Marketing- Data Mining Based Approach for Target Selection" was done by Sadaf Hussein [38]. In his research, a response model for target selection in direct marketing with data mining techniques was constructed for Persian Bank. As he explained, the bank is faced with challenges of increasing competition and decreasing of response rate. To solve the problem the bank need to select the "Customers" that should be contacted in the next marketing campaigns. Hence he tried to predict whether an existing customer will purchase on the next marketing campaign or not, based on information provided by the purchase behavior variables. For this purpose, he developed a predictive response model with data mining techniques to select the customers that should be targeted in order to obtain a percentage as high as possible of positive responses. The customers were divided into two classes, respondents and non-respondents.

Various classification methods (classifiers) were used for response modeling such as statistical and machine learning methods. Neural networks, decision trees and support vector machines. In the response modeling procedure different steps were taken, such as: data collection, data preprocessing, feature construction, feature selection, class balancing, classification and model evaluation. Finally he pointed out that the result obtained is satisfactory and also recommend on future works that need attention. He suggested future work in the area of response modeling

using other techniques (other than support vector machine).

Either directly or indirectly all the above researchers impress the importance of identifying the best group of customers in the market.

Chapter Four

Experimentation

4.1 Overview

To successfully accomplish the objectives of the research the contribution of this section is central and most important. In order to enable successful CRM, the initial task is to identify market segments containing the highest potential customers [46].

This research work includes all the fundamental stages that are incorporated in the data mining process and specifically in the Cross Industry Standard Process for Data Mining (CRISP-DM), which is shown in Figure 4.1.

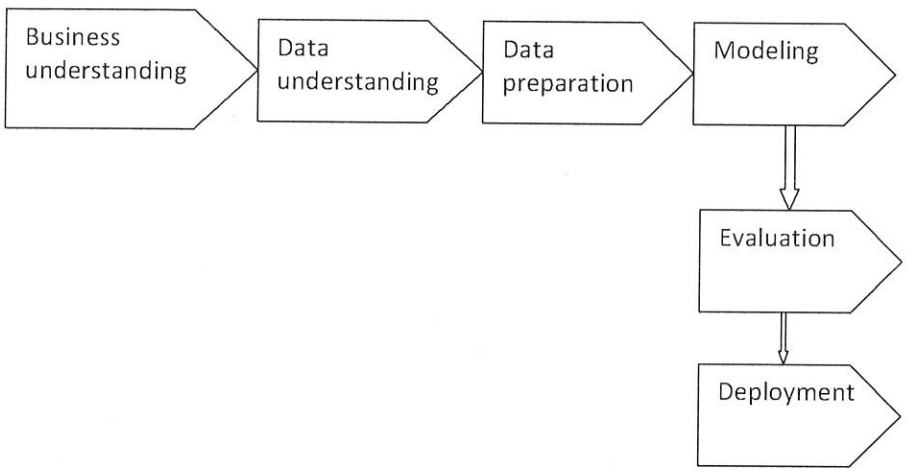

Figure 4. 1: Phases of the CRISP-DM life Cycle [36]

4.2 Business Understanding

Based on the discussion in chapter three (section 3.3.2), which focus on the current direct marketing program in the bank, it is clear that the bank needs a quick solution to its current challenges in conducting a successful potential visa card customer selection process.

4.2.1 Data Mining Goals

One of the data mining goals is to identify the most important variables from the data collected that can be used for cluster model building. Hence the next step, that is segmenting customers into different groups, is achieved by using these selected variables. This will greatly helps to have a clearer and bigger picture of the customers. As a matter of fact, better understanding of customers plays a great role in taking actions that are very much efficient and effective.

Since the understandability of the clusters obtained is highly dependent on the data preparation and analysis phases, these steps are given a higher attention in the process of this research. The data preparation and analysis help the identification of important attributes that could serve as an input for the model building. The last data mining goal is to build a classification model that can automatically assign a new record to the already identified cluster indexes by using the decision tree learning method. In this classification process the cluster index served as the dependent variable whereas the other attributes as the independent.

The successful accomplishment of this research is evaluated against the ability of the clustering model to come up with the appropriate number of clusters and at the same time the ability of the decision tree to classify a new record as accurate as possible to the correct cluster index.

4.3 Data Understanding

Having defined the data mining goals, the next step is the understanding of the customers' data (demographic and financial data). In this phase an investigation on the availability of data that are useful for achieving the research goal were dealt. Therefore, in order to fulfill the data requirement, data was initially collected regarding customers behavior from the DASHEN BANK S.C. database. As careful analysis of data and its structure is invaluable, the researcher has gone through many processes that ensured the availability of a well organized data. As part of the researcher endeavor to come up with a best sort of data, the evaluation of the relationship of the data with the problem at hand and the particular data mining tasks were dealt with the experts at the bank. Figure 4.2 shows the different tasks in the data understanding phase.

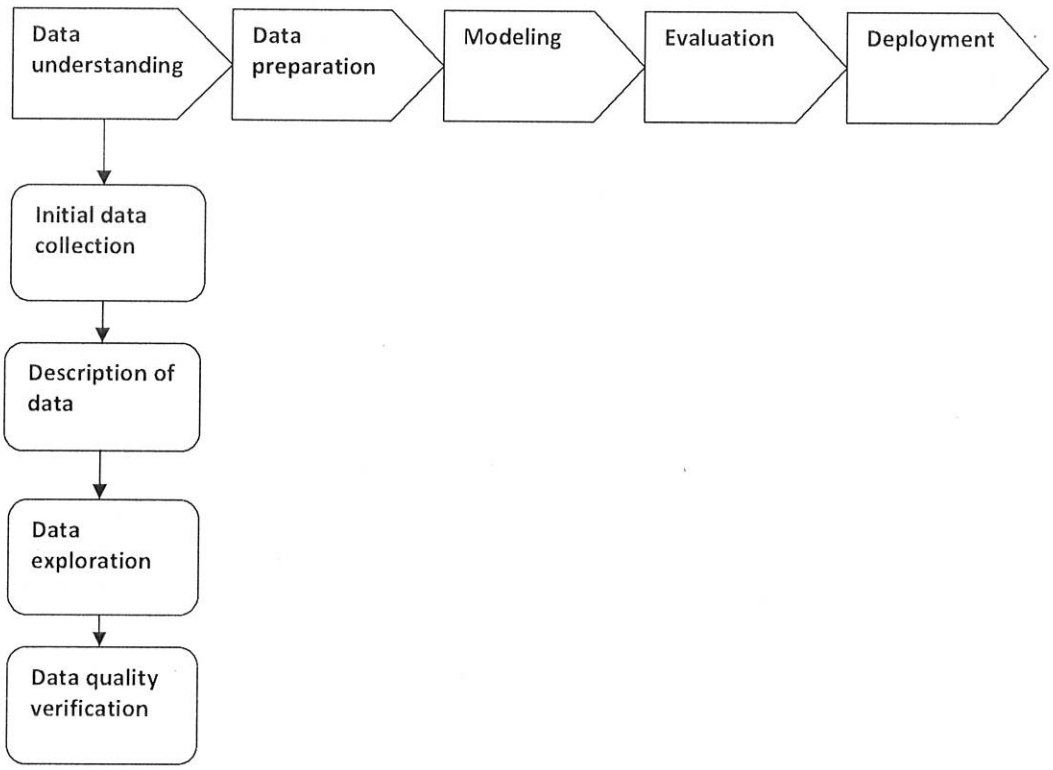

Figure 4. 2: The Data Understanding Phase [36]

4.3.1 Initial Data Collection

Identify the sources of the data that will serve the mining process is one of the major steps of data understanding. The primary data source for this research is the DASHEN BANK S.C. database. The server which stores the customers' data is not connected to the rest of the banks servers instead it directly connected to the international visa and master card servers. The responsible body for storing and maintaining this database is the IT section of card payment department. The database stores demographic and financial data. The visa card customers' data has been electronically stored since 2006.

4.3.2 Description of the Data Collected

This step helps to describe the contents of the database. The collection has nearly 25000 records and more than 40 attributes. Some of the attributes that were obtained from the initial data collection using MS-Excel is shown in Table 4.1.

Attribute Name	Data Type	Description
Name	Text	The name of the customer
Id .no	Number	Identification number of the customer
Id issued by	Text	The authority that issue the id
Date of birth	Date	Date of birth of the customer
Place of birth	Text	Place where the customer born
Gender	Text	sex of the customer
Name of employer	Text	Name of the company where the customer works
Employer address	Text	Address of the company where the customer works
Position	Text	The position of the customer in his company
Occupation	Text	The occupation of the customer
Monthly income	Number	The amount of monthly income of the customer
City/town	Text	The city where the customer lives
Sub city/woreda	Text	The sub city where the customer lives
Kebele	Number	The Kebele where the customer lives
House no	Number	House number of the customer
Home tel	Number	The home phone number of the customer
Office tel	Number	The office phone number of the customer
Mobile	Number	Mobile number of the customer

Table 4. 1: Some of the Attributes of Customer Registration Sheet

4.3.3 Data Quality Verification

The reliability of the data and completeness of records are relatively good as the data is produced electronically. Since demographic and financial data play a great role in determining a customer response to the direct marketing campaign, the researcher believes that the available data is enough for the intended purpose. Despite the above fact, some of the collected data have missing, incomplete and irrelevant values. The researcher has tried to solve these problems in the data preparation phase.

4.4 Data Preparation

This phase involves a number of steps to provide the final dataset for modeling. As shown in Figure 4.3, it includes data selection, cleaning, construction, integration and formatting.

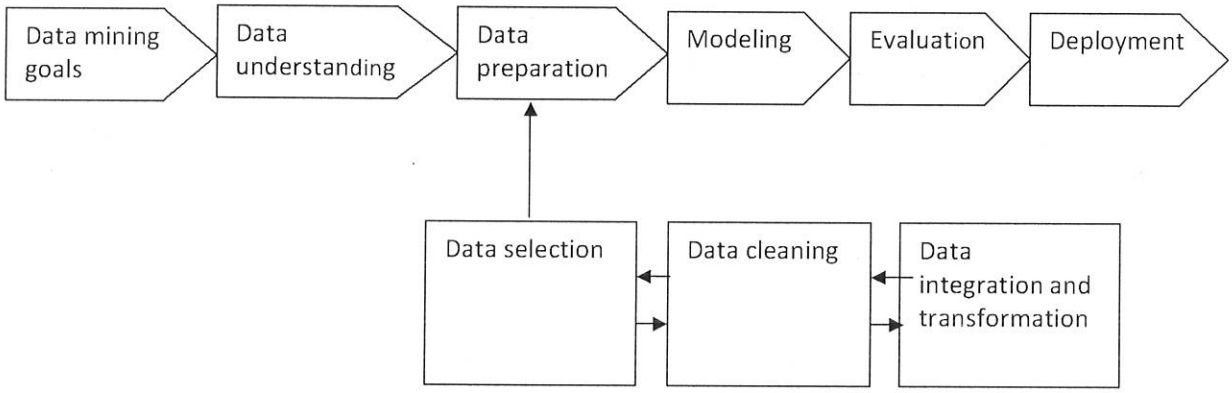

Figure 4. 3: Data Preparation Phases [36]

4.4.1 Data Selection

The collected data for this researcher was selected from huge collection of customers' data. Since it was very much difficult to personally access the database, the researcher used those data that were made available by one of the experts at the bank. As noted in Appendix C, there are

nearly 40 attributes. By discussing with the experts at the DASHEN BANK S.C. some attributes that were considered crucial for the analysis purpose were selected. The criteria used for the selection include relevance to the data mining task as well as quality constraints. The initial attributes collection includes some irrelevant ones like address of the customer, name of employer, identification number, phone number, and employer address of both primary and secondary applicants and others. Since these attributes are believed to have less significance for this researcher, all of them were excluded from the dataset. The "Age" attribute was obtained by subtracting the "Date of Birth" attribute from the current date. Therefore the "Date of Birth" attribute was excluded. Since customer information is recorded only once, there is no any work done with respect to excluding redundant records in the collection. After the data preparation process was done, the dataset has left with 5110 records and 8 attributes. Since the data mining task to be performed need relatively higher number of records, all of the 5110 records were used for the experimentation purpose.

4.4.2 Data Cleaning

Data cleaning helps to clean the data by filling in the missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Different methods were used to handle the missing values such as ignoring the tuples, filling the missing values by using the modal value (for nominal and ordinal variables), and the mean (for continuous variables). The researcher mainly used four methods namely: using mean value, modal value and filling the missing value manually and also ignoring tuple.

From the available attributes "Monthly_Income", "Educational_Level" and "Occupation" were found relatively to have the highest missing value. In the case of " Monthly-Income" there were

180 missing values, so that to fill in the missing values the researcher used the mean value of monthly incomes of the customers who are in the same group with respect to the type of occupation(hired or private). For the attribute "Educational_Level" there were 163 missing values, to handle this situation the technique of filling the most frequent value (College) for the variable "Educational-Level" was used.

In the case of "Occupation" there were 60 missing values, to handle this situation the researcher has gone through the dataset manually and filled the value "hired" for those customers who earn a monthly income less than 5000 and "private" for those who have a monthly income greater than 5000. There were also 28 missing values for the attribute "Marital_Status" and filled by considering the age and gender of the customer, if the age is greater than 26 and the gender is female or if age is greater than 30 and the gender is male then the missing value was filled with "Married" unless "Single" was filled. In addition to this some records which contain large number of missing values were discarded. For the purpose of data cleaning Microsoft excel was used.

4.4.3 Data Transformation and Integration

Since the attributes were derived from customers' profile, this data construction step was important. This is because the customer file contains raw data, which is not in the form appropriate for the business goal to be addressed, and the corresponding data mining tasks to be performed. In the case of data formatting the data was converted into .arff format, which is suitable for the data mining tool. Since all the data available for this research purpose was available in a single file there is no need for integrating different data sources. Creating the new attribute from the existing attributes can improve the result of data mining models [39]. As

discussed earlier in this section there was a derived attribute. This attribute is "Age" which was derived from "Date-of-Birth". Age = current date (2008) – Date-of-Birth. Finally, the attributes that are ready for the analysis purpose are shown in Table 4.2.

Attribute/Field Name	Data type	Description
Age	Number	Age of the Customer
Gender	Text	Sex of the Customer
Occupation	Text	The Occupation of the Customer
Monthly income	Number	The Amount of Monthly Income of the Customer
Educational level	Text	The Educational Status of the Customer
Accommodation	Text	Accommodation of the Customer
Marital status	Text	Marital Status of the Customer
Money deposited	Number	The amount of money deposited.

Table 4. 2: Final Attributes of the Final Dataset

4.5 Model Building

This is the major phase of the CRISP-DM process and hence there are a number of tasks the researcher has gone through. These include selection of the modeling technique, laying out a test design, building a model, and finally the assessment of the model built.

The K-Means clustering algorithm examines all the records and assigns each record to the nearby cluster center. The main thing to be remembered here is that to come up with a good segmentation model, a great deal of care should be taken in selecting of decisive attributes, scales and configuration information for the K-Means algorithm. In this research a thorough analysis and interpretation of each and every cluster is made by the researcher and the domain experts. For the understanding and interpretation purposes the following approaches were used:

- Visually understand how the output differs by changing in the input variables.
- Examine the difference in the distribution of variables from cluster to cluster, one variable at a time.
- Lastly, automatically grew a decision tree with the cluster indices as the dependent variable, and use it to derive rules explaining how to assign new records to the right cluster.

4.5.2 Test Design

During this step data should be partitioned into training and test sets. It is important to create both a training dataset, which is used to build the model, and a test or hold- back dataset, which is used to test the model.

Most of the time researchers randomly split data into training and test sets. Training is typically done on a large proportion of the total data available, whereas testing is done on some small percentage of the data that has been held out exclusively for this purpose. 70% of the data was used for training purpose while the remaining was used for test set.

For the clustering purpose all the records of dataset was used for the training purpose, because clustering is un supervised learning, where the algorithm is provided with the data points with out labels, the task is to find suitable representation of the underlining distribution of the data.

4.5.3 Clustering Modeling

After the completion of all the above steps, the next step is to build the clustering model followed by decision model using the selected tools. The basic configuration parameters available in WEKA for the K-Means algorithm include:

- Display standard deviation: -Display standard deviations of numeric attributes and counts of nominal attributes. The available choices are true and false.
- Do not replace missing values: - To replace or not missing values globally with mean/mode. The available choices are true and false.
- Number of clusters: - The number of clusters (K in K-Means) that need to be created. This value has to be manually input into the system. Most of the time the value of K ranges from 2-20, but it has to be determined by the number of segments that the business can successfully handle or manage.
- Seed: -The random number seed to be used.

In this research four clustering experiments (experiment 1,2,3,4) were conducted by changing the values of K.

Experiment 1

Before starting this experimentation part, the researcher believes that it is important to mention the fact that there was a discussion with the experts at the DASHEN BANK S.C. This discussion focused on assessing the influential factors for being a visa card customer. Generally the experts

were discussing some of the most important variables that are used to select a potential customer of the visa card service. Thus the researcher would like to point out the important points raised by the experts in the following paragraphs.

There is no actual quantitative definition of a good segmentation output, assessing the clusters based on certain decisive attribute is sensible [34]. Thus the attributes "Monthly_Income", "Money_Deposited", "Educational_level" and "Age" were given a very high weight by the experts. Consequently, in the experimentation part the analysis and interpretation of each and every cluster was highly dependent on these attributes. But this doesn't mean that the rest of the attributes have no importance, rather it is to note the weight given to these variables in the real world by the experts. As the experts explained, if a customer has the following characteristics he/she considered as having a higher probability to be the potential customer of the visa card service:

- very high score in the variables Monthly_Income and Money_Deposited
- The age category is in the young and middle age group , like thirties and twenties
- The Educational status is high like college

On the other hand the customer will most probably consider as having a low probability to be a potential customer of the visa card service if the following conditions are met:

- A low score in the variables Monthly_Income and Money_Deposited
- The age category fall around late Forties, Fifties, above these
- The educational level is low like elementary

Finally, all the other possible combination of the above value together with the rest of the attributes values evaluated as being fall between these two extreme cases and hence evaluated

accordingly. Thus, it is important to note that the evaluation and interpretation of the clusters was fully dependent on the above points and additional advices of the experts.

As shown in Table 4.3, the final attributes that were taken for the experimentation purpose are "Money_Deposited", "Age", "Gender", "Occupation", "Monthly_Income", "Educational_Level", "Accommodation" and "Marital_Status". The rest of the attributes which are shown in Appendix C were excluded. The decision was based on the consensus reached by the researcher and the experts on the degree of useful information provided by the variables. The discarded variables were believed to provide very little useful information. Rather they may reduce the accuracy of the algorithm. Since clustering is unsupervised data mining technique, all the selected variables were set as independent variable.

In order to easily see the pattern discovered, the researcher used the dataset mean, maximum and minimum values together with the suggestion of the domain experts to determine threshold values. This threshold values are found at Appendix D.

To represent the values of the variables the researcher used short forms, which is shown in Table 4.3.

65

Values of the Variables	Short Form
College	CO
Elementary	EL
High School	HS
Forties	FR
Fifties	FT
Thirties	TH
Twenties	TW
High	H
Low	L
Medium	MD
Very High	VH
Married	MR
Single	SG
Female	F
Male	M
Private	PT
Hired	HD
Rent	RT

Table 4. 3: Short Forms for the Values of the Attribute Used

Figure 4.4 shows the training result of the clustering model including the name, the number of attributes used, the number of records used, the test mode, and other information.

```
== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 5 -S 10Relation: dataset both numeric and
nominal-weka.filters.unsupervised.attribute.Remove-R3, 6, 9

Instances:    5110

Attributes:   8   Gender, Occupation, Monthly_Income, Marital Status, Age, Educational
Level, Money_Deposited, and Accommodation.

Number of iterations: 3

Within cluster sum of squared errors: 11051.0
```

Figure 4. 4: Training Result of the First Cluster Run

Cluster index	Freq. records	Gender	Occupation	Monthly Income	Marital status	Age	Educational Level	Money Deposited	Accommodation
1	1661	M	HD	L	SG	TH	CO	L	PT
2	764	M	HD	L	SG	TW	CO	L	RT
3	2079	M	PTT	VH	MR	FR	HS	VH	PT
4	150	F	HD	MD	SG	FT	EL	H	PT
5	456	M	PT	L	SG	FR	HS	L	RT

Table 4. 4: Summarized Result of the First Experiment.

As shown in Table 4.4, on the summary report, all the variables are taken as independent variable. It also shows the size of the dataset used in the clustering analysis i.e. 5110 records (all the dataset).The algorithm assigns appropriate cluster index for each of the records in the dataset. This type of visual output provides a descriptive classification model of the clusters, which plays invaluable role in exploring and identifying the characteristics of the clusters.

The result shows that Cluster_4 has very small number of instances (3%). This condition prompts the researcher to suspect the presence of some outlier values, which contain values that are beyond the normal trend. Most of the clustering algorithms consider these types of exceptional values as customers who have a unique character, so that put as a single cluster. Thus the researcher go through the clustering result and found that there were figures beyond the normal value, which were assumed to be encoding error on the attributes "Marital_Status" and "Age". Then the dataset was assessed visually for the particular values of the two attributes and found that there are unusual or strange values for these two attributes. Consequently nine records were manually removed from the dataset. In addition to this, Cluster_1 and Cluster_2 contain customers who show more or less a similar pattern. Thus the researcher believed that the produced clusters are not good enough to represent different group of customers.

As indicated at the beginning of this experimentation, the researcher together with the domain area experts' selects attributes that are considered important for classifying customers according to their likelihood of being a visa card customer. In addition to this, the researcher believes that it is also possible to improve the performance of the model by selecting the most statistically significant attributes by using decision tree. For this purpose the researcher used decision tree to classify the records to the appropriate cluster index. Finally the attributes that are found at the top

most of the decision tree and that are believed to increase the performance of the model were selected as statistically significant variables. Thus by including the above adjustments the next experiment was conducted.

Experiment 2

The following attributes were selected by the decision tree as statistically significant attributes: Gender, Occupation, Monthly_Income, Age, Educational_Level, Money_Deposited, and Accommodation.

After the removal of the records that have outliers and also by selecting the above attributes, this second experiment is conducted. The specific values and percentage of composition for each attribute in each of the clusters are shown in Figure 4.5.

```
=== Run information ===

Scheme:   weka.clusterers.SimpleKMeans -V -N 5 -S 6
Relation:  dataset-weka.filters.unsupervised.attribute.Remove-R3, 5, 8
Instances:  5101
Attributes: 7   Gender, Occupation, Monthly_Income, Age, Educational_Level,
                Money_Deposited, Accommodation
Test mode:  evaluate on training data
=== Model and evaluation on training set ===
KMeans
Number of iterations: 3
Within cluster sum of squared errors: 10708.0
```

Figure 4. 5: Training Result of the Second Cluster Run

Cluster index	Freq. records	Gender	Occupation	Monthly Income	Age	Education Level	Money Deposited	Accommodation
1	1087	81%F	72%PT	65%L 19%MD 10%H 4%VH	53%TH 24%TW 31%FR 13%FT	60% CO 25% HS 14% EL	54% L 17% MD 16% H 12% VH	71% PT 28% RT
2	1339	87%M	78%PT	7%L 27%MD 12%H 52%VH	8%TH 7%FR 28%FT	59% CO 21% HS 18% EL	12% MD 17% H 70% VH	94% PT 5% RT
3	955	80%M	86%HD	95%L 4%H	68%TH 32%FR 5%FT	66% CO 32% HS	86% L 4% MD 7% VH	12% PT 87% RT
4	692	61% M	76% HD	94% L 5% MD	100% TW	81% CO 13% HS 5% EL	83% L 16% MD	5% PT 94% RT
5	1028	67% M	95% PT	30% L 18% MD 21% H 28% VH	16% TH 83% FR	80% HS 19% EL	14% MD 10% H 74% VH	86% PT 13% RT

Table 4. 5: Summarized Result of the Second Experiment

From the result of the second experiment shown in Table4.5, the third cluster and the fourth cluster have a very similar pattern except for the variable "Age". Even the dissimilarity of the "Age" attribute is the smallest possible disparity that is Cluster_3 is dominated by customers around the age of thirties while in Cluster_4 the majority of customers are at the age of twenties. In simple word both of the clusters represent customers who have a low probability to be the potential customers of the visa card service. The experts believe that it is not a wise decision to

Cluster index	Freq. records	Gender	Occupation	Monthly Income	Age	Educational Level	Money Deposited	Accommodation
1	1598	94%M	81%PT	17%L 35% MD 22%H 24%VH	22%TH 2%TW 45%FR 29%FT	48%CO 26%HS 24%EL	1% L 18%MD 22%H 57%VH	92%PT 7%RT
2	1259	64%F	96%PT	8%L 14%MD 11%H 65% VH	19%TH 52%FR 28%FT	12%CO 73%HS 24%EL	5%L 8%MD 12%H 72%VH	83%PT 16%RT
3	1450	65%F	78%HD	92%L 5%MD 2%H	63%TH 28%TW 7%FT	77%CO 17%HS 5%EL	86%L 10%MD 2%VH	27%PT 72%RT
4	794	100%M	65%HD	100%L	63%TW 29%FR 6%FT	74%CO 25%HS	80%L 14%MD 4% H	22%PT 77%RT

Table 4. 6: Summarized Result of the Third Experiment.

As clearly shown in Table4.6, Cluster_3 and Cluster_4 have customers with a low probability of being a potential visa card customer and at the same time these two clusters have similar pattern. The characteristic of these clusters with respect to the different attributes is summarized below.

the same as that of Cluster_1 and classified it as containing high probability customers. On the other hand, others said that it contains the mixture of those customers with very high, high and medium probability of being a potential customer of the visa card service.

Cluster_3

This cluster contains the second largest number of customers (1450). Most of the customers (92%) earn a low Monthly income, 86% and 10% of the customers have low and medium amount of money deposited respectively. In terms of educational level 77% and 17% of the customers are college and high school graduate respectively. As being evaluated by the experts and the researcher this cluster was considered as containing those customers with low probability to be the visa card customers.

Cluster_4

This cluster contains the least number of customers 794(16%). All of the customers in this cluster earn a low Monthly_Income, 80% and 14% of the customers have a low and medium amount of deposited money respectively. In the case of Educational_level 74% and 25% of the customers are college and high school graduates' respectively. This cluster also considered as containing those customers that have a low probability of being a potential visa card customer as that of Cluster_3.

Accordingly, the two clusters, Cluster_3 and Cluster_4 have almost similar patterns in terms of the variables "Monthly_Income", "Money_Deposited" and "Educational_ Level". In addition to this, the two clusters show a very similar pattern for the variables "Accommodation" and "Occupation". The only difference between these two clusters is in the variable "Age" and "Gender" even in this case it is not a big one. Thus it became difficult to differentiate between

the two clusters; rather the researcher found himself in dilemma to say Cluster_3 contains customers with better probability to be the potential customers than Cluster_4 or vice versa.

Consequently, the researcher believes that all the above problems happened because of the fact that the clustering algorithm did not perform well. A good clustering algorithm tries to maximize the similarity within the class while decreasing similarity between classes. But in this experiment not only the third and fourth clusters but also the first and the second clusters show a close similarity. If this model were a good model, it would be simple to rank clusters according to their probability of being a potential visa card customer. Because of the above mentioned points, the next experiment was conducted by changing the value of K to 3.

Experiment 4

In this experiment K was set to three but everything including the number of records are same as the previous experiment (experiment two)

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -V -N 3 -S 6
Relation:     dataset-weka.filters.unsupervised.attribute.Remove-R3, 5, 8
Instances:    5101
Attributes: 7   Gender, Occupation, Monthly_Income, Age, Educational_Level,
                Money_Deposited, Accommodation
Test mode:    evaluate on training data
=== Model and evaluation on training set ===
KMeans
======
Number of iterations: 3
Within cluster sum of squared errors: 8958.0
Missing values globally replaced with mean/mode
```

Figure 4. 7: Training Result of the Fourth Cluster Run

Cluster index	Freq. records	Gender	Occupation	Monthly Income	Age	Educational Level	Money Deposited	Accommodation
1	1184	34% M 65% F	83% PT	49% L 28% MD 12% H 9% VH	21%TH 10%Tw 43%FR 22%FT	10% CO 68% HS 21% EL	17% L 36% MD 22% H 23% VH	75% PT 24% RT
2	2006	85% M 14% F	86% PT	7% L 19% MD 17% H 55% VH	19%TH 3%TW 45%FR 31%FT	48% CO 33% HS 17% EL	3% L 6% MD 12% H 77% VH	92% PT 7% RT
3	1911	58% M 41%F	80% HD	93% L 4% MD 2% H	44% TH 40% TW 10% FR 4% FT	80% CO 17% HS 2% EL	89% L 6% MD 3% VH	18% PT 81% RT

Table 4. 7: Summarized Result of the Fourth Experiment.

In this experiment it became possible to clearly distinguish between the clusters and also to classify each of the customers to a different level according to their expected response to the direct marketing campaign (as high, medium and low). From the above experiment, the following behaviors are detected from each cluster.

Cluster_1

- This is the smallest cluster that constitutes 23% of the total customers. Most of them are female (65%).

- The age group that dominate this cluster is forties (43%) followed by thirties (23%), fifties (22%) and twenties (10%).From the whole dataset 13%, 18%, 31% and 27% of the twenties, thirties, forties and fifties are found in this cluster respectively. This indicates that, the fifties and forties have a better chance to be the medium probability customer. While the twenties and thirties have less chance to join these medium probability customers.
- A great number of customers (83%) run their private business while the remaining customers are hired by others. From the whole dataset 31% of the private and 9% of the hired customers are found in this cluster.
- Almost half of the customers (49%) earn a low monthly income while (28%) medium, (12%) high and (9%) very high. In terms of monthly income, from the whole dataset 42% of the medium, 23% of the low, 29% of the high and 8% of the very high are found in this cluster.
- 68% of them are high school graduates, (21%) elementary and (10%) college. From the whole dataset 45% of the high school graduates, 40% of elementary and 3% of the college graduates are found in this cluster. This shows that the college graduates have a negligible chance to have a medium probability.
- With respect to the amounts of money deposited, the medium depositors (36%) take the higher proportion followed by very high (23%), high (22%) and low (17%). From the whole dataset 64% of the medium, 10% of the low, 52 of the high and 14 of the very high depositors are found in this cluster.
- The majority of the customers (75%) have there own residence. From the whole dataset 28% of the private and 14% of the rent customers are found in this cluster.

The researcher supported by the suggestion of the domain experts analyzed this cluster and consider it as containing a "medium" probability customer. The number of customers with

medium monthly income in this cluster is greater than any other cluster. While in terms of the variable money deposited this cluster consists of the second largest number of medium depositors. Generally this cluster consists of customers who show an intermediate behavior between customers in Cluster_2 and Cluster_3.

Cluster_2

- This cluster consists of 39% of the dataset and 85% of them are male.
- The majority of them (86%) run their own business. From the whole dataset 56% of those customers who have their private job are found in this cluster.
- 55% of the customers gain a very high monthly income, while the remaining 19 %, 17% and 7 % are medium, high and low in that order. From the whole dataset 91% of the very high and 70% of the high monthly income customers are found in this cluster. Whereas, only 5% of the low income customers are found in this cluster.
- In term of age group 45% are forties, 31% fifties, 19% thirties and 3% twenties. From the whole dataset 6%, 26%, 56%, and 65% of the twenties, thirties, forties and fifties are found in this cluster respectively. This clearly shows that, most of the customers with the age of fifties and forties are found to be high probability customers, while very few numbers of customers with the age of twenties and thirties have this same chance.
- Most of the customers are college graduates (48%) followed by high school (33%). From the whole dataset 54% of the elementary, 37% high school and 37% college graduates are found in this cluster.
- 77% of them store a very high amount of money. From the whole customers in the dataset who store a very high and high amount of money 82% and 48% of them are found in this cluster respectively. On the other hand only 3% of the low depositors are found in this cluster.

- Almost all of them (92%) have a private residence. From the whole dataset 60% of the customers who have their private residence and 7% of the customers who use a rent house are found in this cluster.

As being explained by the experts, this cluster contains those customers who are most probably merchants with good educational level and also who deposit and withdraw their money frequently. Moreover, the experts said that the highest amount of profit has been gained from this type of customers. These customers are also considered by the experts as containing most of the POS service user. So that, this cluster is categorized as having customers who are highly expected to respond positively to the direct marketing campaign and classified as "high" probability customers.

Cluster_3

- This cluster consists of the second largest number customers (37%).

- The number of male (58%) and female (41%) are considerably proportional.
- 80% of the customers are hired, while there are small numbers of customers who run their own business. From the whole dataset 76 % of the hired and 12% of the private customers are found in this cluster.
- Majority of the customers (93%) are characterized by low monthly income. In this cluster none of the customers earned a high monthly income. Those customers with medium (4%) and high (2%) took very small percentage of the total. From the whole dataset 71% of the low income customers and only 7% of the high income customers are found in this cluster. On the contrary there is no any costumer with a very high monthly income.

- The age composition of this cluster is highly dominated by thirties (40%) and twenties (44%). The remaining 10% and 4% are shared by forties and fifties respectively. From the whole dataset 81%, 56%, 12%, and 8% of the twenties, thirties, forties and fifties are found in this cluster respectively. This clearly shows that, most of the twenties and thirties are found to be low probability customers while very few numbers of the forties and fifties are found to be low probability customers.
- 80% and 17% of the customers are college and high school graduates in that order. In contrary to this the numbers of customers whose level of education fall under the category elementary are negligible, only 2%. From the whole dataset 58% of the college graduates, 18% the high school graduate and 6% of the elementary are found in this cluster.
- The majority of customers (89%) accumulated a low amount of money while medium (6%) and very high (3%) depositors comprise the remaining. In this cluster there is no any customer with the label high. From the whole dataset 87% of the customers with a low deposit are found here at the same time only 3% of the very high are found in this cluster.
- Most of the customers (81%) have not their private home rather they rely on rent. From the whole dataset 78% of the rent and 11% of the private customers are found in this cluster.

As being interpreted by the experts, this cluster typically represents the young and middle age of both male and female customers of the bank. Most of them are college graduates who are hired in different governmental or private organization and their monthly income is the primary source of their livelihood. This cluster is considered as having customers with a "low" chance to positively respond to the direct marketing campaign. The experts explained that, in the past three or four years these customers show a positive attitude toward the visa card service and the

bank has been trying to reach these customers in different ways. On the course of reaching these customers, the bank has been investing much amount of money for convincing this group of customers (young and middle age customers). But the result of this research shows that the bank would have been a beneficiary if it has tried to invest much on customers who are found in Cluster_1 and Cluster_2. The experts said that, the trend in the bank is to contact more of the young and middle age customers with out giving much attention to their financial status and other variables and this cost the bank a lot.

From the output of this research the researcher realize that the variable "Occupation" and "Accommodation" play a great rule in determining customers' response to the visa card offer. In this respect if a customer runs a private business, then the probability to respond positively to the visa card offer is relatively high. Similarly customers with a private residence are found to have a high probability to respond positively. A large proportion of customers with educational level of elementary and high school are found in the high and medium probability customers than the college graduates. The output also shows that from current visa card customers with elementary educational level most of them are grouped under the high probability customers while a very few of them are categorized as a low probability customer. With respect to the variable age, this research reveals that customers at the age of fifties have high probability to respond positively to the visa card offer. And the probability decrease as the age decrease.

After seeing the result, experts understood that the bank should not confine itself only to the four attributes (monthly income, educational level, age and money deposited) rather it should take into consideration all the attributes that were included in this research. It is important to note that only one or two variables can not determine the probability of being a visa card customer, rather

all the above explanation are meant to impress the importance of the attributes in influencing the response of the customers to the visa card offer.

The overall result of this experiment (the fourth experiment) looks satisfactory because of the fact that it satisfies the criteria of a good segmentation model, it is the clarity of the segments to be explained by the domain experts. The result shows different group of customer segments and most of the drawbacks indicated in the previous experiments are solved. As clearly indicated, some of the clusters in the previous experiments are suffering from having patterns which are difficult to interpret. In addition to this, the clustering algorithm put customers with similar pattern in different clusters. More than any thing else, this research found attributes that play a critical role in determining the customers ability to be a potential customer.

4.5.3.1 Choosing the Best Clustering Model

Four experiments were conducted to come up with the appropriate segmentation model. Finally the segmentation model that satisfies the criteria of good clustering model more than any one of the others was selected. The best set of clusters may be simply the one that shows some expected pattern in the data [4].

There is no actual quantitative definition of a good segmentation output, assessing the clusters based on certain decisive attribute is sensible [34]. Thus it is more of a subjective judgment and hence in this evaluation part, the following major and common criteria are mainly used:

- Good clustering model maximize the intra class similarity while minimize the inter class similarity.
- A good clustering model is the one that could be easily understandable and interpretable by the domain experts.

The first experiment, which used five (5) for the value of K, indicated that very few records are assigned to Cluster_4. The researcher suspects the presence of outliers and by going through the dataset some numbers of outliers were found. Cluster should contain enough customers to develop a separate marketing strategy [16]. In addition to this the patterns generated in different clusters show similarity, as an example the pattern between Cluster_1 and Cluster_2. All the above reasons are highly supported by the domain experts.

Thus, this situation brought the need for the second experiment with same value of K (5) but with removed outliers and also without the attribute "Marital_Status". As the result shown in Table4.5, there are customers showing similar patterns categorized in different groups. In addition to this it is difficult to give a clear interpretation to some of the clusters because of the great degree of heterogeneity of customers within the clusters. Particularly Cluster_5 shows this heterogeneity character.

The third experiment was conducted by reducing K to 4 and the result is summarized in Table 4.6. Still some of the problems that are clearly seen in the previous experiment (experiment two) could not be solved. Consequently, the researcher was forced to conduct other experiment.

The fourth experiment was conducted by setting the value of K to 3, and the result is shown in Table4.7. As can be seen from the table there are three clusters behaving differently. There exists a better defined separation or differentiation among the three clusters and also the homogeneity within the clusters is better than the previous experiments. In addition to this, all the clusters are distinct and meaningful to the domain experts and hence it is easy to develop a separate marketing strategy for each of the three clusters.

Finally, the researcher together with the suggestion of domain experts decided that the appropriate numbers of clusters are three and hence the fourth experiment was selected as good model showing a good segmentation of the visa card customers of the DASHEN BANK S.C. The output of this clustering model (cluster index) is used further as an input for decision tree building.

4.5.4 Classification Modeling

During this phase, the output of the clustering model is used as an input to the classification purpose. For this classification purpose the decision tree algorithm called J48 is used to classify an instance to the already identified cluster index. The cluster index serves as the dependent variable and all the attributes as independent.

The whole dataset is used for classification purpose and the dataset is divided into training and test set as shown in Table 4.8.

Total Dataset	Training Set (70%)	Test or Evaluation Set (30%)
5101	3570	1531

Table 4. 8: Partition of the Total Dataset

3570
1531
5101
10,202

Decision tree performs best when all the attributes contain non-continuous values. Since all the continuous data were converted to the appropriate form there is no need to do the descritization task.

In this classification sub phase two experiments are done. The first one is by using the default value for the parameter "number of minimum objects" that the leaf node should contain, it is two. For this default value the accuracy is found to be 96.14%. The second experiment is conducted by using different values for the parameter, number of objects at each leaf node. The researcher found that the accuracy of the classification model decrease for those values other than the default value. As an example the researcher takes the value 25 for the number of minimum objects and found an accuracy of 95.25%. Finally, based on their accuracy level the default (96.14%) is found to be better than the other values (95.25%).

Thus the decision tree with default value of parameter is selected as the best classifier.

Actual	Predicted			Total	Score (Actual Rate)
	Cluster_1	Cluster_2	Cluster_3		
Cluster_1	495	8	14	522	94.83%
Cluster_2	11	386	9	406	95.07%
Cluster_3	5	4	598	607	98.52%
Total	511	398	621	1535	96.14%

Table 4.9: Output from the J48 Decision Tree learner by Using the Default Value of the Parameter Number of Objects

Actual	Predicted			Total	Score (Actual Rate)
	Cluster_1	Cluster_2	Cluster_3		
Cluster_1	493	8	16	517	95.36%
Cluster_2	10	373	23	406	91.87%
Cluster_3	5	4	598	607	98.52%
Total	511	398	621	1535	95.25%

Table 4.10: Output from the J48 Decision Tree learner by Adjusting Value of the Parameter Number of Objects

4.6 Evaluation

During this phase the degree to which the model meets the business objectives is assessed. As directly or indirectly indicated in different part of this research, the business goal is to come up with a model that could find the appropriate number of clusters of customers according to their likelihoods of response to the direct marketing campaign and also to assign new customers to the appropriate cluster index. Consequently, the business can have appropriate response modeling techniques.

The basic criterion to evaluate the segmentation output is based on the probability of customers to respond positively to the direct marketing campaign. This customer response model was defined based on the different demographic and financial information of the visa card customers. Thus clustering and classification models were developed to fulfill the basic business objective of the marketing department of the DASHEN BANK S.C.

The analysis, which was closely undertaken with domain experts, revealed that the final segmentation experiment indeed discover patterns that are really interesting. The best set of clusters may be simply the one that shows some expected pattern in the data [4]. As clearly indicated, the clustering model brought customers into different clusters according to their expected response to the direct marketing campaign. Obviously this is the underling criterion of a good clustering model. In addition to this, the decision tree model(with the default value of the parameter) provides a very good description of the segments and it clearly shows a number of rules that have invaluable help to assign a new customer record to one of the clusters.

Generally, the final result of this research is encouraging and at least it shows the possible application of data mining techniques for the current marketing problems of the DASHEN BANK S.C, particularly direct marketing problem. The researcher believes that, if the result is further analyzed by different marketing and IT experts, it could give them much insight to their customers' behavior and helps the bank to improve its current direct marketing campaign in a cost effective and well studied approach.

4.7 Deployment of the Result

Since the segmentation result is encouraging, it could be used even for more sophisticated marketing purpose. First and foremost to implement this output all the generated rules should be converted to more specific rules, which are simple to understand and apply. In addition to this, the bank should invest on all the appropriate preconditions that are necessary for proper deployment and execution of the process. It demands the right integration and availability of qualified personnel, technology and resources. In addition to this, the bank should keep every important information about its book customers to make use of this model. If all the above

conditions can be met, the output of this study can help the DASHEN BANK S.C. to conduct effective and efficient direct marketing campaign.

Chapter Five

Conclusion and Recommendation

5.1 Conclusion

The major focus of this research is the application of data mining techniques in the area of CRM and more specifically for direct marketing purpose at the DASHEN BANK S.C. To this effect related literatures on data mining, CRM and direct marketing were reviewed. The investigation was conducted based on the CRISP_DM process model. For the experimentation purpose clustering and classification models were built.

The objective of the research is to come up with appropriate number of clusters of customers according to their response to the direct marketing campaign. This model helps to identify potential visa card customers who should be contacted in the direct marketing campaigns by using the information available in the demographic and financial variables.

Consequently, the researcher has tried to build a model that segment customers in different groups according to the likelihood of their response to the direct marketing campaign and achieved encouraging result. The basic criteria for evaluating the segmentation output were measured with respect to four main attributes. These four attributes were "Monthly_Income", "Money_Deposited", "Educational_Level" and "Age". In addition to this, the research reveals that the attributes "Occupation" and "Accommodation" have a very high impact on determining the response of customers to the direct marketing campaign. With respect to the attribute "Age" and "Educational_Level" the experts have gained a new insight that was out of the experts previous thinking. This research shows that a better proportion of the visa card customers with

5.2 Recommendation

Even though the investigation is done for academic purpose, it revealed the possible application of data mining techniques for modeling the response of customers in a direct marketing campaign.

With the development of data mining techniques and databases technology, some areas which are not covered in this study are interesting and need to be explored. In addition, the limitations and shortcoming of this study also provide suggestion for future research.

Thus based on the finding the researcher makes the following recommendations:

Further data mining researches

This study was mainly focus on building a clustering and predictive response model with K-Means clustering algorithm and decision tree classifier. It was not the researcher objective to compare the performances of different classification and clustering algorithms when applying to response model. Other algorithms can be used for response model. Thus further research is suggested that to compare the performances of different clustering and classification algorithms when apply to response model.

Customer segmentation and classification procedure consists of various steps. Different data mining techniques can be applied for implementing each step of modeling. Considering available tools, time and literature the researcher tried to select the better possible techniques and algorithms for each step. Since data mining is an iterative and interactive process it needs refinement and update thus the researcher suggested that for future research, apply other data

mining techniques for classification(like neural network) and clustering(EM). Therefore, the predictive accuracy of model might even increase when other techniques apply for modeling.

In this research most of the attributes are demographic customer information. Only the "Monthly_Income" and "Money_Deposited" contain financial information. Therefore, further research is needed to use more financial information for building a model. Predictive ability of model might even increase when more financial and demographic variables are included in the model. In addition to this, the researcher believes that increasing the number of records could bring a better performance. Thus further research is recommended by including more attributes (financial and demographic) and increasing the number of records.

Develop a customer warehouse

The researcher realized that the data collection and preprocessing phase was by far the most difficult job. Thus, the researcher strongly recommends that DASHEN BANK S.C. should strive to build a data warehouse that contains all the important information that could successfully serve the data mining process. In addition to this, the bank should practice the culture of collecting as much information as possible about the book customers so that different analysis could be made.

Providing the necessary support for researchers

The researcher observes that, the interest of the bank toward the different researches that are being done in different sections was very much low. Let alone providing help and support, they do not want to give information which are simple and nothing to do with privacy. Thus the researcher would like to recommend that the bank should come up with a special research center, which could facilitate and support the academic researchers endeavor in terms of giving the

required information and also experts' advice. This will boost the fruitfulness of the researches which are conducted in the bank. In addition to this since the banking industry operate on a huge and complex data, this research centers could explore the potential application of data mining in different operational areas of the bank.

REFERENCES

1. Abraham (1998). Market segmentation: can you really divide and conquer? Hawks.plc, UK. Available at URL: http://www.Abramhawkes. plc.uk/ub/mktseg.htm. Visited on August 15, 2008.
2. Askale, w. (2001). Data Mining Application in Support of Loans Disbursement Activity at Dashen Bank sc, Unpublished Master's Thesis, Addis Ababa University, Addis Ababa.
3. A.Devamohan (2008). E–Banking Problems and Prospects in Ethiopia, Ethiopia. Available at URL: http://wA.Devamohan%20-%20E-banking.htm. Visited on November 19, 2008.
4. Berry, M. J. and Linoff, G. (2004). *Mastering Data Mining: The Art and Science of* Customer Relationship Management, John Wiley & Sons Inc, New York.
5. B.K.Hansen (2000). Weather Prediction using case-Based Reasoning and Fuzzy set Theory, Published Master Thesis, Technical university of Nova Scotia, Halifax, Nova scotia, Canada. Available at URL: http://www.cs.dal.ca/~bjarne/thesis/htm. Visited on September 23, 2008.
6. Bult, J. and Wansbeek, T. (1995). "Optimal selection for direct mail." *Marketing Science,* No. 4, November 4 1995 pp378-394. Available at URL: http: *//www2.computer.org/portal/web/csdl/doi/10.1109/TKDE.* Visited on September 2, 2008.
7. Catherine, B. and Esa, R. (2001). Information Technology Research report and overview of Data Mining for customer Behavior Modeling version1, 29 June 2001. Available at URL: http:// www.vtt.fi/inf/julkaisut/muut/2001/customerprofiling.pdf. Visited on November 6, 2008.
8. Desarbo, W.S., Ramaswamy, V. (1994). "CRISP: Customer Response Based Iterative Segmentation Procedures for Response Modeling in Direct Marketing" Journal *of Direct Marketing,* vol. 8, No. 3, pp 7-20.
9. Dunham, M. (2002). *Data Mining:* Introductory and Advanced Topics, Prentice Hall, USA. *Available at URL:* http://search.barnesandnoble.com/Data-Mining/Margaret-Dunham/e/9780130888921. Visited on September 5, 2008.
10. Elams, J. (2004). Regional ICT Developments: The AISI Perspective African Development Forum. Available at URL: http://www.uneca.org/adf99/ adf99 ecommerce. Html. Visited on September 9,

2008.

11. Elsner, R.; Krafft, M.; Huchzermeier, A. (2004). Optimizing Rhenania's Direct Marketing Business through Dynamic Multilevel Modeling (DMLM) in a Multicatalog-Brand Environment, *marketing Science*. Vol. 2, No. 23. Available at URL: *www.whu.edu/cms/index.php?id=4052.* Visited on July 23, 2008.

12. Fadlalla, A. and Lin, C.H. (2001). An analysis of the applications neural networks in finance. Vol. 31, No. 4, July 2001 PP 112-122. Available at URL: http://*www.eurojournals.com/IRJFE4%2017%20piotr.pdf.* Visited on November 11, 2008.

13. Fayyad, U.; *Piatetsky-Shapiro, G.* & Smith, P. (1996). *From Data Mining to Knowledge Discovery in Database, Jossey-Bass, San Francisco*. Available at URL: http:// www.citeseer.nj. nec.com/fayyad96 from. html. Visited on November 15, 2008.

14. Felix, S. and Andrew, C. (2008). Electronic Money: Preparing the Stage, Zurich, Switzerland. Available at URL: http://www.felix.openflows.org/html/excash.html. Visited on October 10,2008.

15. Ha, K.; Cho, S. and MacLachlan, D. (2005). "Response models based on bagging neural networks" *Journal of Interactive Marketingl*. Vol. 19 No. 1, 2005 pp 17-30. Available at URL: http://www.dmlab.snu.ac.kr/ResearchPapers/%5BHaK_ChoS_DMac%5D. Visited on November 13, 2008.

16. Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques,* San *Francisco, U.S.A, Morgan Kaufman Publishers. Available at* URL: http://www.crito.uci.edu/publications/pdf/crm.pdf . Visited on September 15, 2008.

17. Henock, w. (2002). Application of data mining techniques for effective customer relation management at EAL, Unpublished Master's Thesis, Addis Ababa University, Addis Ababa.

18. Hiwot, A. (2005). Appling data mining tools and techniques for effective CRM at Ethiopia hotel, Unpublished Master's Thesis, Addis Ababa University, Addis Ababa.

19. Jaideep, S. (2008). Data Mining for Customer Relationship Management CRM, the Dblp Computer Science Bibliography Publisher. Available at URL: http://www srivasta@cs.umn.edu. Visited on October 13, 2008.

20. John W. (2003). *The Art and Science of Customer Relationship Management*, New York, John Wiley & Sons Inc. Available at URL: http://www.ieeexplore.ieee.org/iel5/5992/28097/01255819.pdf?arnumber=1255819. Visited on December 12, 2008.

21. Kaymak, U. and Setnes, M. (2001). Extended fuzzy clustering algorithms, ERIM report series Research in Management, Erasmus Research Institute of Management, Erasmus University Rotterdam, Netherlands. Available at URL: http://www.narcis.info/dare/RecordID/oaiepeurnl176557/Language/en/repository_id/eurdare/;jsessionid=6a9g8h5o8on. Visited on June 4, 2008.

22. Kaymak, U. (2001). Fuzzy target selection using RFM variables, Proc. of Joint 9th IFSA World Congress and 20th NAFIPS Int. Conference, Vancouver, Canada, pp. 1038. Available at URL: http://www.google.co.uk/search?hl=en&q=Kaymak%2C+U.+%282001%29.+Fuzzy+target+selection+using+RFM+variables%2C+&meta. Visited on November 13, 2008.

23. Kumneger, K. (2006). Application of data mining techniques to support customer relation management for Ethiopian shipping lines (ESL), Unpublished Master's Thesis, Addis Ababa University, Addis Ababa.

24. Lahiri, R. (2006). Comparison of Data Mining and Statistical Techniques for Classification Model, Published Master's Thesis, Louisiana State University.

25. Issayas, M. (2007). Dashen to Resume Issuing visa cards, AllAfrica Global Media publisher. Available at URL: http://www.allafrica.com/stories/200708271429.html. visited on June 13,2008. Visited on December 9, 2008.

26. Mao, R. (2001). An Efficient and Effective Method for Multi-Level Multi-Dimensional Frequent Pattern Mining, Unpublished Master's Thesis, Simon Fraser University, Canada. Available at URL: http://www.sal.cs.uiuc.edu/~hanj/pubs/theses.html. Visited on September 11, 2008.

27. Meretework, S. (2004). Data mining application in support of credit risk assessment, Unpublished Master's Thesis, Addis Ababa University, Addis Ababa.

28. Michael, P. (2003). What is direct marketing? conjecture corporation, Sparks, NV 89432

U.S.A. Available at URL: http://www.wisegeek.com/what-is-direct-marketing.htm. Visited on November 15, 2008.

29. Paul, G. and Jongbok, B. (2001). Customer Relationship management. Vol. 10 No. 23, November 2001 PP 140-150 Available at URL: http:// www.cit.ac.ug/events/srec/PapersSubmitted/paper%20tonny.pdf- . Visited on September 15, 2008.

30. Peggy, W. (1998). Knowledge Discovery in Databases: Tools and Techniques, Vicksburg, USA. Available at URL: http://www.acm.org/crossroads/xrds5- 2/kdd.html title. Visited on October 2, 2008.

31. Piatetsky, S. (1991). Tools for Data Mining and Knowledge Discovery, G. Software, USA. Available at URL: http://www.kdnuggets.com/siftware.html. Visited on September 22, 2008.

32. Plate, T. (1997). A comparison between neural networks and other statistical techniques for modeling the relationship between tobacco and alcohol and cancer, Jordan. Available at URL: http://citseer.nj.nec.com/plate96comparison.html visited on October 7,2008. Visited on October 25, 2008.

33. P.T, JOSEPH (2002). E-COMMERCE: a Managerial Perspective, Prentice-Hall of India Private Limited Inc, pp 179.

34. Pritscher, L. and Hans, F. (2008). Data mining and strategic marketing In the Airline Industry, Zurich-Airport, Switzerland. Available at URL: http://www.luc.ac.be/iteo/article/pritscher1.pdf. Visited on July 8, 2008.

35. Raghavan, V.V.; Deogun, J. S.; Sever, H. (1998). "Data Mining: Trends and issues, Vol. 49, No. 4, April 1998 pp. 397-402. Available at URL: http://cuadra.nwrc.gov/pubs/rds97.pdf. Visited on November 2, 2008.

36. Rüdiger, W.; Colin, S.; Thomas, R.; Thomas, K.; Randy, K.; Julian, C. and Pete, C. (2000). CRISP-DM 1.0: A step by step data mining guide. Available at URL: http://www.crisp-dm.org/CRISPWP-0800.pdf. Visited on September 2, 2008.

37. Russell, S. W. (2001). Customer Relationship Management: A Framework, Research Directions,

and the Future, Haas School of Business, USA. Available at URL, http://groups.haas.berkeley.edu/fcsuit/PDF-papers/CRM%20paper.pdf. Visited on November 20, 2008.

38. Sadaf, H.J. (2007). Response Modeling in Direct Marketing Data Mining Based Approach for Target Selection, Published Master's Thesis, Lulea University of Technology, Sweden. Available at URL: http://epubl.ltu.se/1653-0187/2008/014/LTU-PB-EX-08014-SE.pdf. Visited on September 20, 2008.

39. Saarenvirta, G. (1988). Mining customer data, IGI Hershey, PA, USA. Available at URL: http://www.db2mag.com/db_area/archives/1998/q3/98fsaar.html. Visited on August 15, 2008.

40. Sara, A. (2002). Comparison of Target election Methods in Direct Marketing, Published Master thesis, Universidade Técnica De Lisboa, Italy. Available at URL: http:// www.di. ubi.pt/smadeira/ MSC_thesis_2002.pdf. Visited on September 15, 2008.

41. Semeneh, T. (2007). Payment System Reform First Steps in Birritue, NBE Bulletin, Addis Ababa, Nov 2004-Jan 2007.

42. Simon (1993). The Panacea for the Ills of Management Information Systems. Available at URL: http://www.books.google.co.uk/books?id=n2nIM0l1TQ0C&pg=PA350&lpg=PA350&dq=Codd+(1993).++olap&source=web&ots=sRL1BUORgN&sig=6eo8dnSElI8JTxwnQu5sMn58zhk&hl=en&sa=X&oi=book_result&resnum=6&ct=resultOLAP. Visited on December 1, 2008.

43. Steinbock (2000). Data mining and privacy, Journal of Science & Technology, Vol. 11, pp. 105-113, USA. Available at URL: http://www.law.utoledo.edu/students/faculty/Steinbock/steinboc.htm. Visited on September 15, 2008.

44. Sterne and Jim. (2000). Customer service on the internet, john wiley and sons, New York. Available at URL: http://www.epubl.luth.se/1653-0187/2005/03/LTU-PB-EX-0503-SE.pdf. Visited on December 15, 2008.

45. S.Viaene; baesens, B.; T. Van Gestel; M. Stepanova; J. Suykens; J. Vanthienen. (2003). "Benchmarking state-of-the-art classification algorithms for credit scoring" *Journal of the Operational Research Society.* Vol. 54 No. 6 pp 627-635. Available at URL:

http://www.doi.*ieeecomputersociety.org/10.1109/TKDE.2008.131. Visited on November 12, 2008.*

46. Thearling, K. (2003). An introduction to data mining, Praeger publisher, New York. Available at URL: http://www.thearling.com/text/dmwhite/dmwhite.htm. Visited on October 9, 2008.

47. The Direct Marketing Association. (2008).How to Stop Junk Mail, London. Available at URL: http://www. dma@dma.org.uk. Visited on December22, 2008.

48. Two Crows Corporation (1999). *Introduction to Data Mining and Knowledge Discovery,* Chicago, IL, USA. Available at URL: http://www.twocrows.com/intro-dm.pdf. Visited on November 1, 2008.

49. Witten, I.H and Frank, E. (2000). Data Mining: practical machine learning tools and techniques with java implementations, Morgan Kaufmann publishers, San-Francisco.

Glossary of Terms

- **ATM card**: - ATM Card can be used to withdraw money from ATM machines, at any time during 24 hour of the day.
- **Book customer**: - The customer who has a savings or current account in the bank.
- **Direct Marketing campaign: -** It is a marketing strategy the bank uses to persuade book customers to be the visa card customers. The strategy communicates each person individually.
- **E-Banking**: - It is used as a synonymous for Internet banking, though in reality banking activities carried out through the internet just constitute a part of the whole gamut of e-banking.
- **E-payment: -** It is a transfer of fund from one person (payer) to other person (Payee). In E-payments, the funds are transferred through electronic mode.
- **POS**: - It is an acronym for Point of Sale and it is online payment used for the purchase of goods and service.
- **Positive Response to Direct Marketing Campaign: -** It is equivalent to positive response to the visa card offer. It means the customer is going to accept the visa card offer. The degree of acceptance could vary as high, medium or low.
- **Visa Card: -** A Visa Card provides for online electronic payment (POS) like Credit Card but from savings or current accounts of the cardholder for purchases and for withdrawal of money from ATM machines. This card is a deposit access product where cardholder uses his/her own money in his bank account through the visa card on the principle of "Pay First and Use Later". Visa card can be used to make purchase at retail shops and merchant

establishments in the same way as the credit card is used. But in order to use the visa card, the cardholder must have sufficient balance in the account.

- **Visa Card Customer**: - A customer who is registered to use either ATM card (to withdraw money) or POS service (for online purchase) provided that he/she has saving or current bank account.

Appendices

Appendix A: Some of the Rules Generated from Decision Tree

1. If monthly income is equal to low and educational level is equal to high school and money Deposited is equal to medium or high or very high and gender is equal to male and occupation is equal to private then cluster index is 1.

2. If occupation is equal to private and accommodation is equal to private and gender is equal to male and educational level is equal to high school and monthly income is equal to low and money deposited is equal to low then cluster index will be 3.

3. If occupation is equal to hired and monthly income is equal to low and money deposited is equal to low then cluster index is 3.

4. If occupation is equal to hired and monthly income is equal to low and money deposited is equal to low and age is equal to thirties and educational level is equal elementary then cluster index is equal to 3.

5. If money deposited is equal to high and monthly income is equal to low and accommodation is equal to rent and occupation is equal to hired then cluster index 1.

6. If monthly income is equal to medium and accommodation is equal to rent and occupation is equal to private and educational level is equal to high school then cluster index 2.

7. If monthly income is equal to very high and accommodation is equal to rent and occupation is equal to private and educational level is equal to college and gender is equal to male then cluster index 1.

8. If occupation is hired and monthly income is equal to low and money deposited is equal to medium and age is equal to forties and accommodation is equal to private then cluster index 3.

9. If money deposited is equal to medium and age is equal to thirties and gender is equal to female and accommodation is equal to private and occupation is equal to private then cluster index is 1.

10. If occupation is hired and monthly income is equal to medium and gender is equal to female and educational level is equal to elementary or college then cluster index is equal to 3.

11. If money deposited is medium and monthly income is equal to low and occupation is private and age is forties and accommodation is equal to rent then cluster index 1.

12. If monthly income is very high and money deposited is very high and age is equal to thirties and gender is equal to female or female and accommodation is equal to private and occupation is equal to private then cluster index is 2.

13. If occupation is equal to hired and monthly income is equal to low and money deposited is equal to low and age is equal to thirties and educational level is equal high school then cluster index is equal to 2.

Appendix B: A Decision Tree Generated from the J48 Pruned Tree Learner

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

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


Occupation = private
|  Accomodation = private
|  |  Gender = male
|  |  |  Educational Level = college: Cluster_1 (566.0/17.0)
|  |  |  Educational Level = high school
|  |  |  |  Monthly_Income = low
|  |  |  |  |  Money_Deposited = low: Cluster_3 (6.0)
|  |  |  |  |  Money_Deposited = medium: Cluster_1 (35.0/1.0)
|  |  |  |  |  Money_Deposited = high: Cluster_1 (32.0)
|  |  |  |  |  Money_Deposited = veryhigh: Cluster_1 (2.0)
|  |  |  |  Monthly_Income = medium: Cluster_1 (116.0/3.0)
|  |  |  |  Monthly_Income = high: Cluster_1 (147.0/6.0)
|  |  |  |  Monthly_Income = veryhigh: Cluster_2 (397.0/9.0)
|  |  |  Educational Level = elementary: Cluster_1 (270.0/6.0)
|  |  Gender = female
|  |  |  Age = thirties
|  |  |  |  Money_Deposited = low
|  |  |  |  |  Educational Level = college: Cluster_3 (2.0)
|  |  |  |  |  Educational Level = high school: Cluster_2 (37.0)
|  |  |  |  |  Educational Level = elementary: Cluster_2 (0.0)
|  |  |  |  Money_Deposited = medium: Cluster_1 (1.0)
|  |  |  |  Money_Deposited = high: Cluster_1 (35.0/1.0)
|  |  |  |  Money_Deposited = veryhigh
|  |  |  |  |  Monthly_Income = low: Cluster_2 (0.0)
|  |  |  |  |  Monthly_Income = medium: Cluster_2 (0.0)
|  |  |  |  |  Monthly_Income = high
|  |  |  |  |  |  Educational Level = college: Cluster_1 (4.0)
|  |  |  |  |  |  Educational Level = high school: Cluster_2 (34.0/1.0)
|  |  |  |  |  |  Educational Level = elementary: Cluster_2 (0.0)
|  |  |  |  |  Monthly_Income = veryhigh: Cluster_2 (45.0)
|  |  |  Age = twenties: Cluster_1 (39.0/1.0)
|  |  |  Age = forties
|  |  |  |  Monthly_Income = low: Cluster_2 (2.0)
|  |  |  |  Monthly_Income = medium
|  |  |  |  |  Educational Level = college: Cluster_1 (3.0)
|  |  |  |  |  Educational Level = high school: Cluster_2 (71.0/1.0)
|  |  |  |  |  Educational Level = elementary: Cluster_1 (31.0/1.0)
|  |  |  |  Monthly_Income = high
```

```
| | | | | Educational Level = college: Cluster_1 (5.0/1.0)
| | | | | Educational Level = high school: Cluster_2 (4.0)
| | | | | Educational Level = elementary: Cluster_2 (63.0/1.0)
| | | | Monthly_Income = veryhigh
| | | | | Money_Deposited = low: Cluster_2 (0.0)
| | | | | Money_Deposited = medium: Cluster_2 (0.0)
| | | | | Money_Deposited = high: Cluster_1 (2.0)
| | | | | Money_Deposited = veryhigh: Cluster_2 (193.0/5.0)
| | | Age = fifties: Cluster_2 (162.0/1.0)
| Accomodation = rent
| | Monthly_Income = low
| | | Age = thirties: Cluster_3 (267.0/12.0)
| | | Age = twenties: Cluster_3 (162.0/9.0)
| | | Age = forties
| | | | Gender = male
| | | | | Money_Deposited = low: Cluster_3 (38.0/1.0)
| | | | | Money_Deposited = medium: Cluster_1 (2.0)
| | | | | Money_Deposited = high: Cluster_1 (4.0/1.0)
| | | | | Money_Deposited = veryhigh: Cluster_3 (0.0)
| | | | Gender = female: Cluster_2 (35.0)
| | | Age = fifties: Cluster_1 (3.0/1.0)
| | Monthly_Income = medium
| | | Educational Level = college: Cluster_1 (9.0/1.0)
| | | Educational Level = high school: Cluster_2 (78.0/6.0)
| | | Educational Level = elementary: Cluster_1 (8.0)
| | Monthly_Income = high
| | | Gender = male: Cluster_1 (79.0)
| | | Gender = female: Cluster_2 (40.0/3.0)
| | Monthly_Income = veryhigh
| | | Educational Level = college
| | | | Gender = male: Cluster_1 (10.0/1.0)
| | | | Gender = female: Cluster_2 (4.0)
| | | Educational Level = high school: Cluster_2 (48.0/2.0)
| | | Educational Level = elementary
| | | | Gender = male: Cluster_1 (6.0)
| | | | Gender = female: Cluster_2 (3.0)
Occupation = hired
| Monthly_Income = low
| | Money_Deposited = low: Cluster_3 (1321.0/34.0)
| | Money_Deposited = medium
| | | Age = thirties
| | | | Educational Level = college: Cluster_3 (33.0)
| | | | Educational Level = high school: Cluster_2 (4.0)
| | | | Educational Level = elementary: Cluster_3 (0.0)
| | | Age = twenties: Cluster_3 (75.0/2.0)
| | | Age = forties
```

Appendix C: List of All Attributes Taken from User Registration Sheet

Attribute Name	Data Type	Description
Name	Text	The name of the customer
Id .no	Number	Identification number of the customer
Id issued by	Text	The authority that issue the id
Date of birth	Date	Date of birth of the customer
Place of birth	Text	Place where the customer born
Gender	Text	Sex of the customer
Name of employer	Text	Name of the company where the customer works
Employer address	Text	Address of the company where the customer works
Position	Text	The position of the customer in his company
Occupation	Text	The occupation of the customer
Monthly income	Number	The amount of monthly income of the customer
City/town	Text	The city where the customer lives
Sub city/woreda	Text	The sub city where the customer lives
Kebele	Number	The kebele where the customer lives
House no	Number	House number of the

Secondary applicant woreda	Text	The sub city where the second applicant lives
Secondary applicant kebele	Text	The kebele of the secondary applicant
Secondary applicant house no	Number	The house number of the secondary applicant
Secondary applicant home tel no	Number	The home phone number of the secondary applicant
Secondary applicant office tel no	Number	The office phone number of the secondary applicant
Secondary applicant mobile	Number	The mobile number of the secondary applicant
Secondary applicant's employer name and address	Text	The address and name of the company where the secondary applicant works
Secondary applicant position	Text	The position of the secondary applicant
Secondary applicant position	Text	The position of secondary applicant
Secondary applicant income	Number	The monthly income of the secondary applicant
Date	Date	The date the form filled

Appendix D: Threshold Values for the Attributes Age, Monthly_Income and Money_Deposited

1. Monthly_Income (MI): - The amount of money the customer earns per month.

   If MI <=5,000 then MI is Categorized as "Low"

   If 5,000 < MI <= 10,000 then MI is Categorized as "Medium"

   If 10,000 < MI <= 20,000 then MI is Categorized as "High"

   If MI >20,000 then MI is Categorized as "Very High"

2. Money_Deposited (MD): - The amount of money deposited by the customer.

   If MD <=10,000 then MD is Categorized as "Low"

   If 10,000 < MD <= 30,000 then MD is Categorized as "Medium"

   If 30,000 < MD <= 50,000 then MD is Categorized as "High"

   If MD >50,000 then MD is categorized as "Very High"

3. Age (A): - The age of the customer.

   If A <30 then A is Categorized as "Twenties"

   If 30 <= A < 40 then A is Categorized as "Thirties"

   If 40 <= A < 50 then A is Categorized as "Forties"

   If A >=50 then A is Categorized as "Fifties"

Appendix E: Sample Dataset

1	Gender	Occupation	Monthly income	Monthly_income	Age1	Age	Educational Level	Money Deposited	Money_Deposited	Accomodation
2	male	private	33530	veryhigh		33 thirties	high school	500000	veryhigh	rent
3	female	hired	1500	low		28 twenties	college	20000	medium	rent
4	female	hired	6000	medium		30 thirties	college	10000	low	private
5	male	private	5000	low		26 twenties	high school	30000	medium	rent
6	female	private	12000	high		36 thirties	high school	33876	high	private
7	male	hired	5000	low		35 thirties	high school	5000	low	rent
8	male	hired	3000	low		56 fifties	college	7000	low	private
9	male	private	40000	veryhigh		45 forties	college	340000	veryhigh	private
10	female	hired	3000	low		30 thirties	high school	15000	medium	rent
11	male	hired	7000	medium		49 forties	college	35000	high	private
12	male	private	3000	low		25 twenties	high school	2000	low	rent
13	female	private	6000	medium		33 thirties	college	20000	medium	private
14	male	private	15000	high		39 thirties	high school	70000	veryhigh	rent
15	female	hired	3567	low		28 twenties	college	23000	medium	rent
16	female	private	12444	high		35 thirties	high school	56789	veryhigh	private
17	male	hired	6000	medium		30 thirties	high school	60000	veryhigh	private
18	female	private	12000	high		36 thirties	elementary	33876	high	rent
19	male	hired	12789	high		50 fifties	college	80000	veryhigh	rent
20	male	private	11600	high		38 thirties	elementary	123654	veryhigh	private
21	female	hired	3000	low		44 forties	high school	7896	low	rent
22	male	private	19500	high		34 thirties	high school	234879	veryhigh	private
23	female	private	4600	low		25 twenties	elementery	20000	medium	rent
24	male	hired	9000	medium		42 forties	college	45980	high	private
25	male	private	55000	veryhigh		40 forties	high school	500000	veryhigh	rent
26	female	hired	6500	medium		45 forties	college	66908	veryhigh	rent
27	male	private	20500	veryhigh		46 forties	college	76543	veryhigh	private

Sheet1 Sheet2 Sheet3

Declaration

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

Tilahun Muluneh Arage

January, 2009

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

---

Dr. Manoj VNV