ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

THE ROLE OF DATA MINING TECHNOLOGY IN ELECTRONIC TRANSACTION EXPANSION AT DASHEN BANK S.C.

LUEL BERHE

JULY, 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

THE ROLE OF DATA MINING TECHNOLOGY IN ELECTRONIC TRANSACTION EXPANSION AT DASHEN BANK S.C.

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

By
LUEL BERHE

JULY, 2011

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

THE ROLE OF DATA MINING TECHNOLOGY IN ELECTRONIC TRANSACTION EXPANSION AT DASHEN BANK S.C.

By

LUEL BERHE

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
_____	Chairperson	_____	_____
_____	Advisor(s),	_____	_____
_____	Advisor(s),	_____	_____
_____	Examiner,	_____	_____

DEDICATION

I would like to dedicate this paper to my brothers, Surafeal Berhe, Tesfay Berhe, and Simon Berhe who have been struggling and fighting for my education since I was a little boy. **My brothers' congratulations; your dream is now realized!!**

ACKNOWLEDGEMENTS

Above all, I would like to glorify the almighty GOD for giving me the ability to be where I am.

Next to this, I would like to express my gratitude and heartfelt thanks to my advisor Dr. Dereje Teferi for his constructive comments and overall guidance. Besides bringing the research area to my attention, his direction, guidance, and skilful pushes to get me explore had a huge impact both in this research and on my academic development. I also would like to thank all the Dashen Bank S.C. payment card system staffs for allowing me to carry out this research using the required data from the Bank.

My special thanks also goes to my sister, W/ro Kebedu Kassa, for her financial and moral support; I also thank Mihretab G/tsadik, Naod Surafeal, and Helen Surafeal whose advice has been worth considering starting from the beginning up to the completion of the program.

On top of this, I am fishing out my at most appreciation to my friends, Abreham Weldu (for making us daily firfir early in the morning), and Yemane Seged (for updating us latest infotainment from the FM's of Addis) for making my stay in Addis so memorable.

I would also like to extend my gratitude to all my instructors and classmates at college of Information Science for the lovely time and classes we have had together.

Finally, I would like to thank my friends for the constant assistance and encouragement they rendered to me since the time of my admission to the postgraduate program.

LIST OF ABBREVIATED WORDS

ANN: Artificial Neural Network

ATM: Automatic Teller Machine

CRISP: CRoss Industry Standard Process

CRISP-DM: CRoss Industry Standard Process of Data Mining

CRM: Customer Relationship Management

EFT: Electronic Fund Transfer

EPS: Electronic Payment Service

KDD: Knowledge Discovery in Databases

MLP: Multilayerperceptron

OLAP: Online Analytical Processing

POS: Payment of Sale

S.C.: Share Company

MAB: Monthly Available Balance

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In this study the application of decision tree J48, ANN classification algorithms, and K-means clustering algorithm of data mining on CRM the case of EFT of POS service of the Dashen Bank S.C. have been discovered within the framework of CRISP-DM model. The card holder customers data along with customer book information have been collected, cleansed, integrated and transformed for testing using the clustering and classification models. The final dataset consists of 11000 records in which different clustering models at k values of 6, 5, and 4 with different seed values have been traced and evaluated against their performances. The cluster model at k value of 6 with default seed value has shown a better performance. Hence, the output of the best clustering model (i.e. at k=6) has been used as an input for the decision tree and Artificial Neural Network (ANN) classification models.

Different classifications with the J48 decision tree algorithm are tested with 10-fold cross validation, and splitting the dataset into 70% for training and 30% for testing, techniques by setting the cluster index formed by the cluster model as dependent variable and the remaining variables as independent variables. Different decision tree classification models with **minNumObj** =default, 5, 10, 15, 20, and 25 have been experimented. From these decision tree parameters, a model with default parameter values showed the maximum overall classification accuracy (i.e. 99.55%).

Likewise, different classification models of Multilayerperceptron ANN have been tested by changing the hidden layer and learning rate parameter's value. As a result, a model with a classification accuracy of 99.97%, which is with default parameter value, was chosen. Lastly, a comparison of the decision tree and ANN models in terms of the overall classification accuracy , accuracy in classifying high level customers, and accuracy in classifying low level/value customers have been undertaken. Therefore, the ANN model has been the best in these evaluation parameters, and thus selected as a better classifier in EFT of POS service customers.

The result obtained in this study was encouraging as it has very high classification accuracy. This helps and strengthens the possible application of data mining to the

banking industry in general, and in the EFT of POS service expansion marketing strategy at the Dashen Bank S.C.

CHAPTER ONE

INTRODUCTION

1.1. Background

Information plays vital role in every of the human aspects. It is key resource, in parallel with land, human, and capital, to the success of any organization. It has become an invaluable input to any sound and acceptable strategic decision making processes in the twenty first century. Many private and public organisations depend on information to successfully accomplish their duties and responsibilities and to give better decision making. Information has been tightly integrated with the day –to-day life of almost all individuals. For example, human beings are uncertain of what they are doing without the right information. It is an integral part of every walks of human life. However not all individuals and organizations are using information in an equitable way. Based on the usage of information, the world nowadays is partitioned into agrarian, industrial, and information societies. Among them, information societies use information largely in their day-to-day activities (Deshpande and Thakare, 2010).

According to Deshpande and Thakare (2010), information societies are engaged in the collection, processing, storage, and dissemination of information. Thus, to efficiently inspire information, it is very important to generate information from massive collection of data. The data can range from simple numerical figures and text documents to more complex information such as multimedia data, spatial data, and hypertext documents.

However, the huge size of these data sources make it impossible for human experts to come up with interesting information or patterns that will help in the proactive decision making process (Rajanish, 2002). Therefore, processing information embraces application of sophisticated and expensive information technologies solutions. Due to these and other costs, the information is buried deeply under a massive amount data. Here it has to be noticed that data is the ultimate source of information (Deshpande and Thakare, 2010).

Today, in many business areas, detailed customer interaction data is abundant. Organizations maintain data about customer purchasing behaviour, returns, complaints, wishes, and more. Particularly, Banks gather data from various sources such as customer loan payments, saving history of customers, customer payment card system usages etc. Even though, these voluminous data are retained in business organizations legacy and current databases, it has been very difficult to analyze them manually by human experts; making it difficult for them to understand what information or knowledge is hidden. The reason for this is that the technology for generating, capturing, and storing data has far outpaced the human capacity to understand, analyze, and exploit it for maximum impact. To take complete advantage of these large amount of data; data retrieval is not enough rather it requires a tool for automatic summarization and classification of data, extraction of the essence of information stored, and the discovery of patterns in raw data and this tool is called data mining (Deshpande and Thakare, 2010).

The goal of data mining, when applied to a business data, is to allow a company to improve its marketing, sales, and customer support operations through better understanding of its customers. This application of data mining has, therefore, a potential impact in the banking business. The way in which banks interact with their customers has changed dramatically over the past few years. A customer's continuing business is no longer guaranteed. As a result, banks need to understand their customers better and quickly respond to the customers' wants. In addition, the time frame in which these responses need to be made has been shrinking. It is no longer possible to wait until the signs of customer dissatisfaction are obvious before action must be taken. In the highly competitive and information demanding environment of the banking industry, the speed of a decision could be as critical as the decision itself.

Data mining techniques can be applied to a wide variety of data repositories including databases, data warehouses, spatial data, multimedia data, Internet or web-based data and complex objects (Lori, 2006). Advances in computer hardware and data mining software have made data mining systems accessible and affordable to many businesses (SIM, 2002). It can be applied in many areas, including education, banking and insurance, medicine, security and communication (Deshpande and Thakare, 2010). Hence, it is not

surprising that data mining has gained widespread attention and increasing popularity among bankers in recent years.

Currently, huge electronic data repositories are being maintained by banks and other financial institutions across the globe. Valuable bits of information are embedded in these data repositories (Rajanish, 2002). The only problem is that this storehouse of data has to be mined for useful information. Normally, these terabytes of transaction data are collected, generated, printed, stored, only to be filled and discarded after they have served their short-lived purposes as audit and paper trails (Rene, 2010).

However, Global competitions, dynamic markets, and rapidly decreasing cycles of technological innovation provide important challenges for the banking and finance industry (Rajanish, 2002). As banking competition becomes more and more global and intense, banks have to fight more creatively and proactively to generate knowledge from large database. This is possible by introducing application of data mining system in banking sector (Rene, 2010).

Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume of data is too large or is generated too quickly to screen by experts (Rajanish, 2002). Specifically, data mining tools are used by bankers to identify business areas and acquire new customers. This is the first task in Customer Relationship Management. Businesses want to target their recruitment investment to choose the most profitable ones— achieved through the application of data mining models on customer profile database (David and Yasmin, 2006).

1.1.1. Banking history in Ethiopia

Emperor Menelik-II was the first to realize the importance of a banking service in the late 19^{th} century. But there was lack of educated human resources in this area. The poor economy of the country at that time and the absence of the necessary administrative framework didn't allow the formation and operation of local bank. Therefore,

introduction of foreign banks had to be appreciated to begin addressing the emerging need of a banking service.

Banking in Ethiopia started in 1905 which was called “Bank of Abyssinia” by Emperor Menelik-II. It was a private company controlled by the Bank of Egypt. In 1931 it was liquidated and replaced by the Bank of Ethiopia which was the bank of issue until the Italian attack of 1936. During the Italian living, Bank of Italy banknotes formed the legal tender. Under the subsequent British occupation, Ethiopia was briefly a part of the East Africa Currency Board (Bekezela, 2008).

According to Shiferaw (as cited in Askale, 2001), states bank of Ethiopia was set up and a new currency was designed in 1943. The increasing economy during the 1950’s led to the expansion of the money market and opening of new banks as well as beginning of branches of foreign banks. Once more in 1945 a new currency was issued and again in the same year, the agricultural bank was established and later in 1952 it was changed to development bank of Ethiopia.

In 1963, functions of the banks were formally separated and the National Bank of Ethiopia (the central and issuing bank) and the Commercial Bank of Ethiopia were formed. By the time the revolution broke out in 1975, there were four commercial banks operating in Ethiopia. These are commercial bank of Ethiopia, Addis Ababa bank (private bank), Banco di Roma and Banco di Napoli. On January 1st 1975 the provisional military administrative council (Derg) nationalized all banks. The national bank of Ethiopia was kept as a separate institution while the three commercial banks were merged with the commercial bank of Ethiopia (Bekezela, 2008).

Government monopolized the banking sector until 1994. By the end of 1994, establishment of private banks become a reality under the promulgation of proclamation No. 84 of 1994 for licensing and supervision of banking business in Ethiopia. These days there are many private and public banks operating in the country such as Commercial Bank of Ethiopia, Dashen Bank S.C., Awash International Bank S.C., Bank of Abyssinia S.C., Construction and Business Bank, Development Bank of Ethiopia, Zemen Bank, Wegagen Bank etc. There are different services provided by these banks. Some of the

services include loan, mobilization of deposits, electronic fund transfer via cards, and international money transfers.

1.1.2. About Dashen Bank S.C

Now a day, private banks are well established and some are opening currently at an increasing rate. Dashen Bank S.C. is one of the private banks established in Ethiopia since the promulgation of proclamation No. 84/1994 that allowed the formation and operation of private banking business in Ethiopia. It was established as per the aim of the new policy and the Ethiopian investment code. The bank came into existence on September 20, 1995 according to the Commercial Code of Ethiopia as a share company with an authorized and subscribed capital of Eth. Birr 50 Million. The first founding members were 11 businessmen and professional that agreed to combine their financial resources and expertise to form this new private bank with a mission of providing efficient and customer focused domestic and international banking services, overcoming the continuous challenges for excellence through the application of appropriate technology (Sofia and Seid, 2008). The bank has now 208 share holders, 650,000 book customers, 62 area banks, 28 ATM machines, 258 POS and it stood first on the market share among other banks in the country.

As Isayas (2007), described Dashen Bank S.C. is one of the private banks that has been introduced the modern banking system through the visa cards in the history of Ethiopia. He has also stated that, this bank has introduced the technology (visa card) firstly from an American company at a cost of 3.5 million dollars, and it also invests a lot for professionals/experts who install and control the system. Besides this the bank also pays a huge amount of money to replace and maintain the machines, telecommunication costs and much more after installation.

Currently the bank has different electronic banking services. These are visa card service which is provided with ATM (this is used to withdraw money) and Point of Sale (POS), which is used to transfer money electronically. These two services are now used in hotels and resorts, hospitals, tour and travel agencies, gallery and jewelry shops, tourisms, cafés and restaurants, fuel stations, supermarkets, and so on. Now the bank is working and

targeted in creating cashless society by encouraging and promoting the card holder and book account customers to use the EFT of POS service. Because, though the bank has currently 110000 card holder customers, they are not using the ATM and EFT of POS service as expected. The bank has introduced customer MasterCard and Visa card services so that customers can access and use their accounts internationally. The aim of the bank is to achieve its objective so as to provide world class modern banking system in Ethiopia.

The bank is interested in making more profit as much as possible by providing quality services to its customers. To achieve this goal, implementation of an information system and building of an information infrastructure is very crucial (Sofia and Seid, 2008).

While data mining technology is rapidly growing in the developed parts of the world and has been applied for a variety of tasks, it is still unknown in the Ethiopian banking industries. Taking experiences from the developed world in terms of the benefits acquired in applying data mining technology in the banking industry, it is only proper to discover the importance and application of the highest degree of development in the Ethiopian banking context. Detailed discussion about data mining is given in chapter-2.

1.1.2.1. Electronic payment

Payment represents both cash and non-cash financial transactions, which take place between two or more parties. In the history of mankind several payment mechanisms are observed ranging from traditional exchange system to the modern day electronic payments (Sumanjeet, 2009).

As Grover (2007) described, a real revolution in the meaning of electronic payment system came with the development of EFT (Electronic Fund Transfer) technology. EFT is one of the electronic commerce technologies that allow the transfer of funds from the bank account of one person or organization to another electronically. Consequently, the online payment of funds appeared to be the next logical step in a progressive move towards the electronic funds transfer and banking, a process that had begun long before the Internet itself.

Modern payment systems are almost nonexistent in our country. The country's economy is significantly cash based. Checks are also being used as an alternative payment mechanism but to a very limited extent. Today cards are becoming more popular in different parts of the world to effect payments safely and in a convenient manner. According to information obtained from Grover (2007), bank cards with a Visa or MasterCard logo are now used by over one and half billion people worldwide, accepted in more than 24 million retail locations as well as over one million ATMs and is the preferred method of payment for over 10% of consumer transactions.

The evolution of payment card came to existence in the late 1940's and early 1950's in the United States. It was developed in response to people's high instant demand for bank loan to purchase household items. Taking into consideration the inconvenience of the long process it takes to follow the traditional loan processing system, a credit officer in one of the bank's in New York, introduced a way of making approvals of loans in advance i.e. before the customer selects what to buy. The approved loan serves as a special currency which the merchant need to agree to accept. More clearly it works as follows.

- Bank approves loan to consumer and puts a special currency in his/her account
- Customer spends the special currency with the acceptable merchant
- The merchant deposits the special currency to the customer's bank
- Consumer pays loan to the bank in installments

In the early 1950's the special currency is changed into the first modern bank card. However, the usefulness of the card was still limited by the local nature of the agreement between merchants and banks. This problem forced banks to create association and try to manage usage of a bank's card at any merchant location. This leads to the emergence of Card Associations like VISA, MasterCard, American Express and Diners Club (Sumanjeet, 2009).

Sumanjeet (2009) states, in a typical complete card payment process, five parties are required to involve.

- The issuer- is a financial institution that issues the cards required for payment.
- The cardholder- is the customer of that financial institution who is provided with the card. The cardholder is expected to use the card for payment at Point of Sale (POS) terminal or for withdrawal of cash at an ATM that can accept the card.
- The merchant- is the business owner or trader who can accept cards as a payment tool in exchange for the products/services he/she sales to the cardholder.
- The acquirer- is the merchant's representative which will make a payment to the merchant when a transaction is effected using a card and later claim the amount from the issuer.
- The Card Association- Is someone like Visa or MasterCard who will facilitate the interchange for the whole payment process. They basically serve as central settlement and clearance body for the transaction made in between the issuer and the acquirer.

1.1.2.2. Benefits of using payment cards

Sumanjeet (2009) have explained the benefits of payment cards. Cards are becoming the preferred way of payment systems in the developed world due to the fact that they understand the benefits very well through long years of using them in their economy. The usage of payment cards can be seen from different angles as the following.

- To the cardholder- The cardholder can feel safe by having the card in his/her pocket and may develop peace of mind. Another is its convenience; he/she can put a single card in pocket and can buy whatever is needed at whichever convenient time and place.
- To the merchant- Company's can avoid or minimize management of cash flow. They can do business with cards to reduce such costs if collections are made electronically. It can also minimize risks as well as it can increase sales.
- To the Bank- the payment card is an additional product to its business which will serve as an additional means of income and it can increase customer if such innovative technology is introduced.
- To the economy- such technology can definitely attract tourists. Travelers are the major users of payment cards due to its convenience to carry and they safely

involved in it. There is also flow of foreign currency as it encourages tourists to spend more as well as minimizes printing of paper money.

1.2. Statement of the problem

Many researchers have studied how to handle customer-bank relationship by applying a data mining technology to come up with interesting information (or patterns) that will help in the decision making process. These researches have studied in different departments of the bank operation, such as Electronic payments in sub sahran Africa (Donald and Sparks, 2008), data mining in banks and finance (Rajanish, 2002), customer relationship management of credit card business (Chen et al., 2005), etc using different data mining techniques.

Chen et al. (2005), used classification to classify the selected customers into clusters using RFM model to identify high profit customers. Subsequently, they carried out data mining using association rules algorithm in the customer relationship management of credit card businesses.

Rajanish (2002) has also studied on data mining in banking and finance. As a huge size of data is impossible to analyze with human brain in order to get interesting knowledge, he decided to use data mining technology to discover the hidden knowledge from different electronic data repositories, both internal and external to support better decision making.

Banks in Ethiopia should apply such technology in order to explore interesting knowledge from the available databases in order to achieve their goals. Because data mining technology extracts hidden knowledge from the available datasets and this will give a direction to the banks on where to give more attention on the expansion of electronic payment cards. Banking industries in Ethiopia involves different activities related to doing bank business. Their primary task is to collect money from customers with justifiable interest rate, lend the collected money for different business owners with different criteria, performing foreign exchange, funding development, and supporting payment activities of other businesses.

In addition to these activities, the fundamental and the primary task of a bank is finding new customers and retaining the available customers of the bank. Prior to any activities and any investment, banks should have an identification mechanism of their potential and profitable customers.

As Ethiopia is a developing country, everyone is expected to work hard without wasting their time. Hence, banks should be concerned about the wastage of time in money transaction for their customers. In such situation where utilization of time is crucial for the country, people‘s should not waste their time for their business transaction in banks.

Dashen bank in Ethiopia is one of the prestigious banks that have many branches in most cities of the country. The bank is now working towards extending its electronic services throughout the country. Even though the bank is working on the expansion of the payment card system (E-transaction) to make the business transaction electronic, it lacks a documented data that can help the bank to classify its customers. Due to this, the bank does not have enough electronic transaction customers as expected. Hence, this can be solved by applying data mining technology and identifying potential electronic transaction users and expand the service accordingly.

Therefore, it is worth investigating and identifying potential Electronic Fund Transfer (EFT) users in order to extract hidden knowledge that will help banks for the expansion of payment card system (E-transaction) throughout the country by applying data mining technology.

Hence, the research questions this study attempt to answer are the following.

- What data mining technique will be used to identify potential customers?
- To what degree the electronic transaction service will be expanded by applying appropriate data mining technique?
- How the bank can create a cashless society?
- What benefits can be found and what possible way can be recommended?
- What the bank should accomplish to expand the EFT?

1.3. Justification of the study

CRM is at the heart of keeping existing customers of an organization satisfied while attracting and retaining new customers. In particular, organizations that provide different type of services should identify the market place for its various services though a good CRM in order to get a competitive advantage.

Hence Dashen Bank S.C. should do the same in order to achieve its objectives, as the bank has a mission of providing efficient, focused and international banking services to its customers. To achieve its mission, the bank is expected to identify the special behaviors of its customers in order to fulfill their interests. Though the bank is trying to change its customers from the traditional banking transaction to modern banking system, the change is not coming as expected.

Data mining can be used to study the behaviors of customers, using the huge data from book profiles of customers, in order to introduce modern electronic banking system. That is the data mining technology can easily identify the key characteristics and behaviors of each book customer and then predict the likelihood of that customer to the EFT of POS service customer. This helps the bank to acquire new customers by encouraging its promotional strategies in the way that targeted the particular potential EFT of POS service customer. According to Joseph (2002), recent projects have indicated a decrease in costs for targeted mailing campaigns over conventional approaches. By predicting customers' behavior in advance, business can then market the right products to the right segments at the right time through the right delivery channels.

Therefore, the bank could properly allocate the budget and man power for the future expansion of its modern services. This research will also help other banking industries that are trying to start the EFT of POS service as a lesson to what extent they should keep and analyze their customers' data to extract the required hidden facts about their customers. Besides this, whenever the bank's ability to expand their service increase, the integration of the society with global community through e-banking will also increase.

The bank has not made an attempt so far in identifying the appropriate numbers of customers of the EFT of POS service and also knowing major determinants for being a

potential customer of POS service in the bank. This study will be a direction for future researches in the same area.

1.4. Objectives of the study

1.4.1. General Objective

The general objective of the research is to expand Electronic Fund Transfer (EFT) of POS service activities at Dashen Bank S.C. by applying appropriate data mining techniques (Classification and clustering) on the payment card system customers' database to extract knowledge so as to specifically identify different segments of the customers. The aim is to test whether data mining techniques can be applied in customer segmentation in the context of Dashen Bank S.C.

1.4.2. Specific Objectives

In order to achieve the above stated general objective, the research has undertaken the following specific objectives.

- To review literature by identifying different papers which are related to this area of study ;
- To collect relevant data for the research from the card payment system of the bank;
- To select appropriate data mining algorithm that is best for the problem domain;
- To pre-process data and clean data, by fill in missing values, smooth noisy data, and resolve inconsistencies;
- To discover the relationship between different variables;
- To apply clustering algorithm to test behavioral segmentation based on the possibility of being potential EFT of POS service customers;
- To apply classification algorithms to build and train the classification models that classify new instances of customer into one of the class labels identified by the clustering algorithm;

- To evaluate the performance of the model in order to measure to what extent it enables marketing personnel to have dynamic access to marketing decisions concerning payment card system customers;
- To report the result and make appropriate recommendations for future research.

1.5. Research Methodology

Methodology is defined as the steps and procedures that one follows to achieve the objectives and research questions of a certain study. It is something that shows the direction in which a research follows to attain the end. In order to achieve the specified objectives of the current research, the researcher has used the CRISP-DM model. The CRISP-DM method is described in terms of a hierarchical process model, consisting of sets of tasks described in six levels: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This model is flexible and easy to work with (Pete, 2000). And it is used because it is popular and has been widely applied for data mining studies. In addition to this, this model is an open-source and industry standard data mining processes model. Accordingly, the research study has followed the following methodologies in order to develop customer segmentation and classification models.

1.5.1. Review of related literature

Thorough review of literatures in the area of data mining and KDD has been conducted particularly, clustering and classification tasks of data mining and their applications in the area of CRM are deeply analyzed and assessed. Various books, journal articles, and conference papers are consulted in order to get clear insight into the concepts and techniques of data mining, KDD and customer segmentation in the framework of CRM. These sources have given firm theoretical and practical background to the current research undertaking.

1.5.2. Fact Finding

To identify, understand, and analyze the business problems, the primary (observation and interview), and secondary data collection methods have been used. The potential source

of data used to undertake this research was mainly the payment card system section of the Dashen Bank S.C. database which contains about 110000 records of card holder members. The financial information of each customer and his/her status of ATM or POS usage are also available.

Moreover, the researcher has conducted interview and observation at different offices of Dashen Bank to get a thorough understanding of the business /domain.

1.5.3. Data identification, Collection, and preparation

The primary sources of data for this research were the payment card system section of Dashen Bank S.C. database. Primary and secondary data were collected through discussion with domain experts. Document analysis was made for the purpose of getting the EFT of POS service customers' data.

To this end, the payment card system (IT) section of the bank has provided the data in Excel format. Then the researcher conducted pre-processing on the data in order to improve the accuracy of clustering and classification process.

1.5.4. Training, Model Building and Testing

The model building phase in the data mining process of this investigation was carried out in two phases, clustering and then classification rule generation. Rules generated for classification purpose being the required final output of the model building phase, the clustering sub-phase, segmentation at a member level was done. To manage the clustering sub-phase and classification sub-phase, the WEKA 3-7-2 software was used. The k-means algorithm has been experimented to segment the customers' data. The clusters at k values of 6, 5, and 4 are examined. The models of the segments at these different values of K have been evaluated together with the domain experts and finally the best cluster model, which is at k value of 6, was selected as an input for the J48 decision tree and ANN algorithm.

Furthermore, different classification models of the J48 decision tree and ANNs have been experimented by changing some of its parameters. First, the algorithms were trained and tested with 10-fold cross validation learning technique, a technique that uses 90% of the

dataset for training and the remaining 10% for testing. In addition to this, other models have been tested with 70% of the dataset for training and 30% of the dataset for testing. The overall best model of J48 decision tree and ANN with default parameter was selected, and finally the ANN was achieved with a better overall classification accuracy.

1.6. Scope and Limitation of the Research

The scope of this research is limited to development of clustering and classification models for Dashen Bank S.C. EFT customers by applying data mining technology. Even though the findings of the research is important for other banks in Ethiopia at large, the scope of this research undertaking is limited in expanding Electronic Fund Transfer (EFT) by appraising data mining system at Dashen Bank S.C. This study has adopted and used some basic techniques from publication which are related to this area to apply it in the Ethiopian banking context. The basic target of this research is expanding Electronic Fund Transfer (EFT) of POS service by applying better data mining techniques and tools at the specified bank.

1.7. Significance of the Study

These days electronic transaction is becoming one of the profitable areas for bank industries and business owners. Some of the advantages of electronic banking are: it minimizes risk of theft, it is convenience to the card holder, it reduces cost of cash management, increases sales to the merchant, it is an additional steam of income and increases customer base due to such an innovative product to the bank, and it increases flow of tourists which can contribute to foreign currency flow to the country. The main beneficiaries of this study result are banks who wish to give quality modern banking services to their customers in particular and the country in general.

1.8. Application of the results

The result of this research can be used by Dashen Bank S.C. to identify trends and patterns of EFT of POS service customers. Dashen can also encourage its promotion and sales strategy based on the particular cluster of customers that will be identified in the course of the study. It also helps the division that manages customers that is solely

established to identify, maintain good relationships, and deliver all the necessary facilities and special services to both the existing and prospective company customers, to easily undertake research on similar problem domain.

1.9. Thesis organization

This research report is organized into six chapters. The first Chapter briefly discusses background to the problem area, states the problem, the general and specific objectives of the study, the research methodology, the scope and limitation, and application of the results of the research. Chapter two lays the foundation for other parts of this research. It thoroughly discusses important literatures in the area of data mining technology.

The concepts pertaining to the data mining technology and its application in the problem are reviewed in chapter two. Chapter three is dedicated for the discussion of three basic related works issues, electronic transaction, CRM, and segmentation concepts. Chapter four explains about data preparation and data pre-processing. It also defines the methods, the clustering techniques and decision trees, used in this research. Moreover, previous researchs, in which a researchable gap was left and become the concern of this research, is reviewed in this chapter. Chapter five presents the experimentation phase of the study at hand. Results of the clustering and classification experiments were also discussed here. Finally, chapter six provides conclusion of the research, and also presents recommendation for future work.

CHAPTER TWO

Literature Review

In this chapter, an attempt has been made to review literature on the concepts and techniques of data mining and its application in both private and public sectors in general and in the banking industry in particular with the aim to provide background about the models to be built.

2.1. Data mining concepts

Today, the capacity of production and collection of data is showing an immense increment. The growth and widespread usage of less expensive computers is enabling various organizations to gather and store huge volume of data. When organizational data bases keep growing in number and size, it is due to the availability of powerful and affordable database systems.

A need for new techniques and tools becomes essential as a result of the explosive growth in data and databases. These tools should help humans to automatically identify patterns; transform the processed data in order to draw meaningful conclusions; and then extract knowledge from the rapidly growing volumes of digital data. Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. Information is basically extracted from the huge size of data. The exponential increase in information primarily due to electronic capture of data and its storage in large data warehouses have created a demand for analysing the voluminous data produced by today's organizations, so that organizations can respond quickly to fast changing markets.

On the other hand, database sizes have significantly increased into larger size of data which have a complete advantage if the hidden information is extracted. To uncover this hidden information and knowledge, data mining is being used both to increase revenues and to reduce costs. The potential return that can be gained is massive. Innovative organizations worldwide are already using data mining to locate and appeal to higher value customers, to reconfigure their product offerings to increase sales, and to minimize losses due to error or fraud (Two Crows Corporation, 2005).

Jeffrey (2004) observes data mining as “a means for detecting fraud, assessing risk, and product retailing”. Data mining involves the use of data analysis, tools to discover previously unknown and valid patterns and relationships in large data sets. It is often seen as a potential means to identify terrorist activities (such as money transfers and communication), and to identify and track individual terrorists themselves in the context of security. Data mining is becoming common in both public and private organizations. Industries such as banking, insurance, medicine, and retailing use data mining for research enhancement and for increasing performance.

There is no question to the vital role that information plays in every sphere of the human life. However, to efficiently use information, it is important to generate information from massive collection of data. The data can range from simple numerical figures and text documents to more complex information such as multimedia data, spatial data, and hypertext documents. However, the huge size of these data sources make it impossible for human experts to come up with interesting information or patterns that will help in the proactive decision making process (Rajanish, 2002). To take complete advantage of data; simple data retrieval is not enough. It requires a tool called data mining, for automatic summarization and classification of data, extraction of the essence of information stored, and the discovery of patterns in raw data (Deshpande and Thakare, 2010).

The process of finding interesting patterns and knowledge in a large database that is not explicitly part of the data is referred as data mining. The extracted interesting patterns can be used to tell us something new and to make predictions and description (Lori, 2006).It is a young interdisciplinary field drawing from areas such as: database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, neural networks, pattern recognition, spatial data analysis, and many more (Han and Kamber, 2006).

As Lori (2006) states, the techniques of data mining can be applied to a wide variety of data repositories including databases, data warehouses, spatial data, multimedia data, Internet or web-based data and complex objects. Advances in computer hardware and data mining software have made data mining systems accessible and affordable to many

businesses (SIM, 2002). It can be applied in any area, including education, banking and insurance, medicine, security and communication (Deshpande and Thakare, 2010).

Among the diversified industries that benefit a lot from data mining are telecommunications and credit card companies. These organizations and companies apply data mining to detect fraudulent use of their services. Insurance companies and stock exchanges are also interested in applying this technology to reduce fraud. Medical applications are another fruitful area: data mining can be used to predict the effectiveness of surgical procedures, medical tests or medications. Companies active in the financial markets use data mining to determine market and industry characteristics as well as to predict individual company and stock performance. Retailers are making more use of data mining to decide which products to stock in particular stores (and even how to place them within a store), as well as to assess the effectiveness of promotions and coupons. Pharmaceutical firms are mining large databases of chemical compounds and of genetic material to discover substances that might be candidates for development as agents for the treatments of disease (Two Crows Corporation, 2005).

2.2. Data mining limitations

Data mining doesn't eliminate the need to know your business, to understand your data, or to understand analytical methods. It does not tell the value of the patterns to the organization. Rather, data mining assists business analysts in finding patterns and relationships in the data. Furthermore, the patterns uncovered by data mining must be verified in the real world. Remember that the predictive relationships found via data mining are not necessarily causes of an action or behaviour. For example, data mining might determine that males with incomes between $15,000 and $45,000 who subscribe to certain magazines are likely purchasers of a product you want to sell. While you can take advantage of this pattern, say by aiming your marketing at people who fit the pattern, you should not assume that any of these factors cause them to buy your product. To ensure meaningful results, it is essential that you understand your data. The quality of your output will often be sensitive to outliers (data values that are very different from the typical values in your database), irrelevant columns or columns that vary together (such as age and date of birth), the way you encode your data, and the data you leave in and the

data you exclude. Algorithms vary in their sensitivity to such data issues, but it is unwise to depend on a data mining product to make all the right decisions on its own. Data mining will not automatically discover solutions without guidance. Although a good data mining tool shelters you from the details of statistical techniques, it requires you to understand the workings of the tools you choose and the algorithms on which they are based. Data mining does not replace skilled business analysts or managers, but rather gives them a powerful new tool to improve the job they are doing. Any company that knows its business and its customers is already aware of many important, high-payoff patterns that its employees have observed over the years.

Technical experts and analytical specialists are required to change and interpret the output obtained from data mining in the way organizations can use it. After all, as Jeffery (2004) emphasizes, "the limitations of data mining are primarily data or personnel related, rather than technology-related. Although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. Such types of determinations must be made by the user. "

2.3. Data mining and Knowledge discovery in databases (KDD)

Different names were given for the extracted information from data. For instance, data mining, knowledge transaction, information discovery, information harvesting, and data pattern processing. As Fayyad et al. (1996) noted, the term data mining has mostly been used by statisticians, data analysts, database, and the management information system (MIS) , communities.

According to Piatetsky-Shapiro (as cited in Fayyad et al., 1996), 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. It has been popularized in the Artificial Intelligence (AI) and machine learning fields.

Many authors take data mining and KDD as one and the same, but there is a significant difference between the two. According to Fayyad *et al*. (1996) 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, is essential to ensure that useful knowledge is derived from the data.

2.3.1. The knowledge discovery process

The decisions made by users involve so many steps because KDD is iterative. This iterative process has been summarized by many researchers. Most of them agree that knowledge discovery process starts with a clear definition of the business problem or, equivalently, understanding of the application domain (Han and Kamber, 2006; Two Crows Corporation, 2005).

The method used in data mining process provides a hierarchical life cycle of a data mining project. It contains the phases of a project, their respective tasks and relationships between these tasks. At this description level, it is not possible to identify all relationships. Essentially, relationships could exist between any data mining tasks depending on the goals, the background and interest of the user and most importantly on the data (Two Crows Corporation, 2005).

The process model of a data mining project consists of six phases. **Figure 2.1** shows the phases of a data mining process. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase, for which phase or which particular task of a phase, has to be performed next. The arrows indicate the most important and frequent dependencies between phases (Pete et al., 2008).

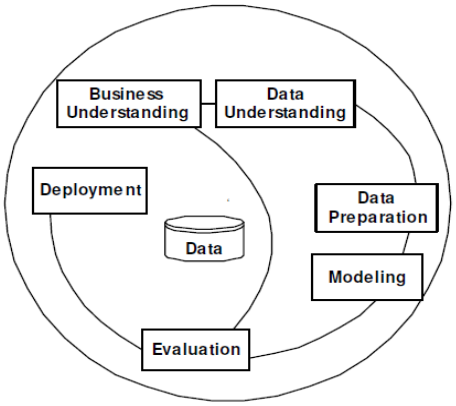


Figure 2. 1: Phases of the CRISP-DM process cycle

In the above **Figure 2.1,** the process model symbolizes the cyclical nature of data mining itself. Data mining is not over once a solution is deployed. The lessons learned during the process and the deployed solution itself can trigger new, often more focused business questions. Subsequent data mining processes will benefit from the experiences of previous ones. **Figure 2.1** shows the commonly agreed steps of knowledge discovery process. These steps are briefly summarized below.

2.3.1. 1. Business understanding

This helps to obtain brief understanding of the requirements and objectives of the project from a business perspective; after having this understanding, the knowledge could be converted into a data mining problem definition and a preliminary plan designed to achieve the objectives.

2.3.1. 2. Data understanding

This phase, on the other hand, starts with collecting data and proceeding with activities in order to be well acquainted with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

2.3.1. 3. Data preparation

The data preparation phase covers all activities performed to construct the final dataset that could be submitted to the tool from the initial raw data. The tasks are likely to be

performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

i. Data cleaning

Data cleaning is an attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data (Han and Kamber, 2006).

Missing values: The case here is tuples, may have no recorded values for some their attributes. There are different methods to handle missing values (Han and Kamber, 2006) such as:

- Ignore the tuple (usually done when the class label is missing).
- Fill in the missing value manually (time consuming and may not be feasible).
- Use global constant to fill in the missing values (Replace all missing values by the same constant).
- Use the most probable value to fill in the missing value.

Detecting noisy data and outliers: Han and Kamber (2006) describe noise as a random error or variance in a measured variable. Outliers are unusual data values that can not represent the data and that are not consistent with most observations. Commonly, outliers result from measurement errors, coding and recording errors, and sometimes are natural, abnormal values. Such non representative samples can seriously affect the model produced later.

There are two strategies for dealing with outliers and noisy values:

- Develop strong modelling methods that are insensitive to outliers.
- Detect and eventually remove outliers as a part of the pre-processing phase

ii. Data integration and transformation

Data mining often requires data integration – the merging of data from multiple data stores. The data may also need to be transformed into forms appropriate for mining. Data mining which involves many activities such as smoothing (to remove noise from data), aggregation (to apply summarization option), generalization (to replace low level data by

higher level concepts), normalization (to scale attributes such that they fall within a specified ranges), attribute construction (to construct new attributes from features) (Han and Kamber, 2006).

iii. Data reduction

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same analytic results (Han and Kamber, 2006).

2.3.1. 4. Modeling

A model is another CRISP-DM hierarchical step which is a high-level, global description of a data set that takes a large sample perspective. Modeling is describing or summarizing the data in a convenient and concise way. It is inferential, allowing one to make some statement about the population from which the data were drawn or about likely future data values.

In this phase, various modeling techniques are selected and applied and their parameters are adjusted to best values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary.

2.3.1. 5. Evaluation

Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

2.3.1. 5. Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be

organized and presented in a way that the customer can use it. It often involves applying live models within an organization's decision making processes. However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases it is the user, not the data analyst, who carries out the deployment steps. However, even if the analyst will not carry out the deployment effort, it is important for the user to understand up front what actions need to be carried out in order to actually make use of the created models.

2.4. Related fields of data mining

2.4.1. Database management and Data mining

There are different fields which play an essential role in data mining; and database management system is one of them. Some of the capabilities of DBMS include persistent storage, data model (e.g. relational or object oriented), and high level query language. These permits users to request 'what data' rather than 'how to access' it (example, SQL). It also provides efficient access to voluminous data. Data mining is realistic and applicable as long as there are large data collected and stored in a certain format. The information or data repository includes data warehouses, relational databases, transactional databases, flat files and World Wide Web. From these, the relational databases are one of the popularly available and rich in information repositories, and thus they are at the heart of data mining processes (Berry and Linoff, 1997).

Han and Kamber (2006) observed, data in a relational database is stored in a format suitable for mining. Due to the advancement of parallel databases, data warehousing and on-line analytical processing tools have greatly increased the efficiency with which databases can support large numbers of complex queries that are basic of data mining applications. Data mining must be integrated with databases to identify interesting pattern and to extract knowledge.

2.4.2. Data warehousing and Data mining

Since the beginning of using computers in data processing centres in the 1960's, almost every operational system in business has been automated. In companies, this automation has resulted in large number of data residing in dozens of dissimilar systems. In response to integrating the disparate systems, data warehousing is the process of bringing diverse data together from throughout an organization for decision-support purposes (Berry and Linoff, 1997).

As Two Crows Corporation (1999) describes data to be extracted is prepared in a company data warehouse first to be stored in data mining database. There is some actual benefit if the data mining database is already part of a data warehouse. A data mining endeavour includes the effort to identify, acquire, and cleanse data. This is due to the fact that the problems of cleansing data for a data warehouse and for a data mining are very similar. If the data has already been cleansed for a data warehouse, then it most likely will not need further cleaning in order to be mined.

Nevertheless, data warehouse is not a pre-requisite for data mining since building a data warehouse requires huge investment and long period of time, and the benefit of data mining projects should exceed the cost to be incurred. As useful as a data warehouse is, it is not the prerequisite for data mining and data analysis. Setting up a large data warehouse consolidates data from multiple sources, resolves data integrity problems, and loads the data into a query database. However, putting such a large database up can be an enormous task, sometimes taking years and costing much. One could, however, mine data from one or more operational or transactional databases by simply extracting it into a read-only database (Two Crows Corporation, 1999).

If the design of the data warehouse includes support for data mining applications, the warehouse facilitates and catalyzes the data mining efforts. In this case, the data mining database is a logical rather than a physical subset of the data warehouse. The two technologies work together to deliver value (Berry and Linoff, 1997).

2.4.3. On-Line Analytical Processing (OLAP) and Data mining

The basic and one of the common questions of the data processing experts is about the distinction between data mining and OLAP (On-Line Analytical Processing). As it is clearly stated in the following mode, data mining and On-Line Analytical Processing (OLAP) are very different tools that can complement each other.

On-Line Analytical Processing (OLAP) was a term coined by E. F. Codd was quoted in (Fayyad et al., 1996), and was defined by him as "The dynamic synthesis, analysis and consolidation of large volumes of multidimensional data." It is a popular approach for analysis of data warehouses. It mainly focuses on providing multidimensional data analysis, which is superior to Structured Query Language (SQL) in computing summaries and breakdowns along many dimensions.

According to Han and Kamber (2001), Traditional query and report tools describe what an OLAP in a database is. OLAP goes further; it is 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.

Even though, data mining involves more automated and deeper analysis than OLAP, it is expected to have broader applications. However, data mining does not replace but rather complements and interconnects with other decision support system capabilities such as OLAP. Before acting on a pattern, the analyst needs to know what the resulting implications would be of using the discovered pattern to govern the decision. OLAP can allow the analyst to answer those kinds of questions (Two Crows Corporation, 2005)

In other words, OLAP analyst generates different hypothetical patterns and relationships and uses queries against the database to verify them or disprove them. Whereas data mining, uses the data itself to uncover such patterns, instead of verifying hypothetical patterns. Additionally, OLAP is also complementary in the early stages of the knowledge discovery process because it can help explore data, for instance by focusing attention on important variables, identifying exceptions, or finding interactions. This is very important

because the better one understands the data, the more effective the knowledge discovery becomes (Two Crows Corporation, 2005).

2.4.4. Data mining and hardware/software trends

Some of the main reasons of using data mining technology are the increment of hardware performance, software advancement, and price reduction. The economics of collecting and storing voluminous amounts of data has radically changed due to the drop in the price of computer disk storage. The drop in the cost of computer processing has also been equally important. Each generation of chips greatly increases the power of the CPU. Virtually all servers today support multiple CPU using symmetric multi-processing and clusters of these servers can be created that allow hundreds of CPUs to work on finding patterns in the data. Technological advancement in database management systems to take advantage of the hardware parallelism also benefits data mining (Two Crows Corporation, 2005).

2.4.5. Data mining, machine learning and statistics

Since the beginning, people have been looking at the world and gathering data to explain natural phenomena such as the movements of the moon, the sun, and the stars. As a result they have created calendars to describe wonderful events. They managed to analyze data and look for patterns without the aid of computers even before recorded history began. It has started to change in the past few centuries. It was the codification of the mathematics and the creation of machines that facilitated the taking of measurements, their storage, and their analysis. Traditional statistics has developed over the past two centuries to help scientists, engineers, and later business analysts to make sense of the data they have collected (Two Crows Corporation, 2005).

Two Crows Corporation (2005) states as data mining takes advantage of advances in artificial intelligence (AI) and statistics fields. Both these disciplines have been working on problems of pattern recognition and classification, and have made great contributions to the understanding and application of neural nets and decision trees. Traditional statistical techniques can’t be replaced by data mining. Instead, it is an extension of statistical methods that is in part the result of a major change in the statistics community.

The basic point is that data mining is the application of AI and statistical techniques to common business problems.

2.5. Tasks of data mining

Data mining tasks are of different types depending on the use of data mining result, these are:

- Exploratory Data Analysis: This is exploring the data without any clear ideas of what we are looking for.
- Descriptive Modelling: describes all the data. It includes models for overall probability distribution of the data
- Predictive Modeling: This model permits the value of one variable to be predicted from the known values of other variables.
- Discovering Patterns and Rules: concerns with pattern detection, the aim is spotting falsified behaviour by identifying regions of the space defining the different types of transactions where the data points significantly different from the rest.
- Retrieval by Content: is finding pattern similar to the pattern of interest in the data set.

Osmar and Zaiane (1999) generally categorized data mining tasks into two:

- **Descriptive data mining**: Focuses on finding patterns that can be interpreted by humans describing the data. Descriptive data mining also describes the general properties of the existing data, and
- **Predictive data mining**: Involves using some variables or fields in the database to predict unknown or future values of other variables of interest. Or it is a task that attempts to do predictions based on inference on available data.

The goals of prediction and description can be achieved using a variety of data mining methods. In response to this, more often than not, a data mining project involves a combination of different activities, which together solve a business problem (Fayyad et al., 1996).

2.6. Successful data mining

The two basic things to be successful in data mining is to come up with a precise formulation of the problem you want to solve and using the right data. The more the model builder can play with the data, build models, evaluate results, and work with the data some more in the given time, the better the resulting model will become. Consequently, the degree to which a data mining tool supports this interactive data exploration is more important than the algorithms it uses (Two Crows Corporation, 2005).

2.7. Application areas of Data mining technology

Many business companies have been exploiting the advantage of data mining for many years. As a result it is becoming increasingly popular. The possible benefit of data mining can be to control costs as well as contribute to income increases (Two Crows Corporation, 2005).

These organizations were using this technology to control all phases of the customer life cycle, including obtaining new customers, increasing profit from existing customers, as well as keeping good customers. By determining characteristics of customers by their profile, an organization can focus prospects with similar characteristics. After profiling the characteristics of customers, the company can identify those who have bought a particular product. This helps to give more attention on similar customers who have not bought that product (Two Crows Corporation, 2005).

Even today, companies which are active in the financial markets use data mining to discover market and industry characteristics and to forecast individual company and stock performance (Two Crows Corporation, 2005). For instance, in the banking sector, data mining enables to detect patterns of fraudulent credit card use, identify faithful customers, predict customers likely to change their credit card affiliation, determine credit card spending by customer groups, find hidden correlations between different financial indicators, and identify stock trading rules from historical market data. With the objective of developing a model that can support the loan decision-making process at

Dashen Bank S.C., Askale (2001), for instance has explored the potential applicability of data mining technology in the banking sector.

Medical applications are also another productive area; data mining can be used to forecast the effectiveness of surgical procedures, medical tests or medications, to predict which medical procedures are claimed together, to characterize patient behaviour so as to predict office visits, and to identify successful medical therapies for different illnesses. In this regard, Shegaw (2002), having the objective of developing a model that can support in preventing and controlling child mortality at the district of Butajira, Ethiopia, has applied the data mining technology and reported an achievement in the development of the model.

According to Rajanish (2002) noted, pharmaceutical firms are also mining large databases of chemical compounds and of genetic material to discover substances that might be candidates for development as agents for the treatments of disease.

Moreover, the potential applications of data mining technology can be applied in other sectors such as airlines in order to support the implementation of an efficient customer relationship management (CRM), in which case the first task was to identify market segments containing customers with high profit. Since the research at hand is exploring the potential application of data mining in expanding Electronic Fund Transfer (EFT) of POS service in the banking industry, the next sub-section reviews the possible applications of data mining in this industry.

Voluminous amount of data is being collected and stored in databases of customers. As a result accessing and making deals in the web sites of the companies, traditional combinations of statistical techniques and data management tools are no longer adequate for analyzing the vast amount of data. Therefore, in order to find ways to intelligently help the enterprise in analyzing the huge amount of data to approach the potential resources, and to make fast and effective decisions for successful E-commerce, data mining technology is needed. As data collection and data storage rates are growing at a surprising rate in the business world in this information age, data mining is becoming a key component of electronic commerce (Xinying and Peizhi, 2008).

2.7.1. Application of Data Mining Technology in the Banking Industry

The advancement of information technology is widely used to power in an important way in the banking and financial industry. Since the 1980s, commercial banking has continuously innovated through technology-enhanced electronic transaction services. Over the past decade, the Internet has clearly played a critical role in providing online services and giving rise to a completely new channel. In the Internet age, the extension of commercial banking to the cyberspace is predictable (Tzyh and Michael, 2006).

Advancement of hardware and software performance also made the technology accessible and affordable to many businesses (SIM, 2002). It can be applied in any areas, including education, banking and insurance, medicine, security and communication (Deshpande and Thakare, 2010). Hence, it is not surprising that data mining has gained widespread attention and increasing popularity among bankers in recent years.

Currently, huge electronic data repositories are being maintained by banks and other financial institutions across the globe. Valuable bits of information are embedded in these data repositories (Rajanish, 2002). The only problem is that this storehouse of data has to be mined for useful information. Normally, these terabytes of transaction data are collected, generated, printed, stored, only to be filled and discarded after they have served their short-lived purposes as audit trails and paper trails (Rene, 2010).

Due to global competitions, dynamic markets, and rapidly decreasing cycles of technological innovation provide important challenges for the banking and finance industry (Rajanish, 2002). As banking competition becomes more and more global and intense, banks have to fight more creatively and proactively to generate knowledge from large database. This is possible by introducing application of data mining system in banking sector (Rene, 2010).

These companies maintain voluminous amount of data which is impossible to process with human brain in order to discover hidden and interesting knowledge from data repositories, as a result it was decided to use data mining technology to support better decision making (Rajanish, 2002).

CHAPTER THREE

ELECTRONIC TRANSACTION, CRM AND SEGEMENTATION CONCEPTS

3.1. Electronic Transaction

According to Microsoft Encarta Dictionary (2009), Transaction is defined as an instance of doing business of some kind. For example, a purchase made in a shop or withdrawal of funds from a bank account. It can also be explained as the act of communication or activity between two or more people that influences and affects all of them.

Electronic banking means a full access to cash through an automated teller machine (ATM) or direct deposit of pay checks into checking or saving accounts for different customers. Electronic banking sometimes called, electronic fund transfer (EFT), uses computer and electronic technology to replace for checks and other paper transactions. The Electronic Fund Transfers (EFTs) is becoming usable by devices like cards or codes that the system let you access your account. Many different financial organizations use ATM or debit cards and Personal Identification Numbers (PINs) for this purpose. EFT gives many services that customers may find practical. Some of the services of EFT are stated below (FTC, 2006).

- Automated Teller Machines (ATMs): are electronic terminals that let customers bank almost any time. Customers simply use an ATM card by inserting and entering personal identification number to withdraw money
- POS: is used to transfer funds between accounts.
- Direct Deposit: allows specific deposits such as pay checks and social security checks to customers account on a regular basis.
- Personal computer banking: enables different banking transactions via the customer's personal computer. For example customers may use a computer to view their account balance, request transfers between accounts, and pay bills electronically.
- Debit card purchase transaction: used to make purchases with a debit card, which also may be an ATM card. This could occur at a store or business, on the Internet

or online, or by phone. While the process is fast and easy, a debit card purchase transfers money fairly and quickly from customers bank account to the company's account and this plays an important role in creating an easy relationship between the purchaser and the company.

Electronic banking gives extra ordinary benefits that help them in the ways they organize financial product development, delivery, and marketing through the internet. Even though E-banking offers new opportunities; it has also created many problems. Some of the challenges are modernization of information technology application, the introduction of new competitors and the appearance of new business models. In general, the speed and scale of the challenges are rapidly growing from time to time due to the pervasiveness of the internet and expansion of the information economy. However to manage the challenge of the E-banking modernization, the present banks should understand the nature of the change and capability barriers that they face. If the banks try to transfer to E-banking without this understanding, it may lead them to failure (Tzyh and Michael, 2006).

Southard and Siau (as cited in Tzyh and Michael, 2006) states banks that are equipped with a good grasp of the electronic banking (E-banking) phenomenon will be able to make informed decisions on how to transform themselves into e-banks and to make use of e-banking in order to compete and survive in the new economy. Banks should continuously reconstruct, renew, or gain organizational capabilities and resources to satisfy the demands of the dynamic environment. Developing core capability can help the banks reorganize their resources and renew their competence to create a sustainable and competitive advantage.

Modern financial organizations have cashed on the electronic business opportunities of the internet by developing different payment systems to satisfy many payment service requirements. Due to the advancement of the computer systems and telecommunications, fast, convenient and secure financial transactions are conducted at service and security levels that are rarely achieved by traditional payment systems. Payment is the act of exchanging something of value for a product or service. There are different basic types of

payment used in a modern society. These include cash, check, credit card, electronic funds transfer, and so on. The essence of these mechanisms is the credit of participants in the payment process.

An innovation is the use of the advancement and use of new technological and business related knowledge to introduce new products and services in order to satisfy customers. It is important to organize customers systematically and to understand them fully to realize the scope and impact of innovation. Electronic banking is basically a financial innovation that is enabled by creative use of emerging IT and other business forces. Due to this the innovation includes a set of IT, customer, finance, marketing, and strategy aspects (Tzyh and Michael, 2006).

An Electronic Fund Transfer (EFT) system has the capability of displaying a menu including one or more user defined transactions associated with an identification card so the user can select and use his/her desired transaction. Electronic Fund Transfer (EFT) systems also include Point-of-Sale terminals and other many systems. Point-of-Sale (POS) are used in different business areas, such as grocery checkouts, Hotels, tour and travel agencies, gas station pumps and other retail locations to enable a user to pay for a purchase by using an ATM card, a credit card, a debit card or other similar methods. POS terminals also basically require a number of user inputs and transaction parameters to effect a transaction. The POS terminal plays an important role in creating a cashless society (James, 1996).

A typical system might consist of terminals at the point of sale; that is on the retailer's premises. The task of these terminals is to communicate with different financial organizations. Anyone who wants to use the Electronic Fund Transfer Point-Of-Sale (EFTPOS) system will be given a plastic card bearing a magnetic stripe. Besides this the customers will also be issued with certain personal identification data. When these customers are wishing to purchase from a retailer, they are expected to bring their card in which it can be read by the terminal, thus giving the terminal card holder related data such as account number, expiry date and so on. The card holder also separately enters their personal identification Number (PIN). After the card holder enters his/her PIN then

the terminal will communicate with the card issuer's computer. This is to check whether the card is valid or not, and also to check whether the specified account contains sufficient funds for the expected transaction and if the PIN entered corresponds to that card or not. The PIN is used as a signature to authenticate the card holder to the card issuer (Maeda and Yokoyama, 1983).

Payments are the basic part of any commercial transaction in different organizations, which involve transfer of funds between two participating parties. Payments play an essential role in most of the e-commerce applications. With an increased growth in E-commerce, there is a major need for infrastructure and frameworks that can be used to implement e-payment services across different organizations (Dani et al., 2001).

According to Dani et al. (2001), the entities contained in the proposed model for the electronic payment service system are the payer, payee, payer's bank, payee's bank and the root certifying agency. The Electronic Payment Service (EPS) system basically acts as the interface between the users of the banks and the banking system. The proposed model for Electronic Payment Service (EPS) supports different types of payments, payment of taxes and duties, utility payments, insurance premiums, investments in approved securities etc. Using these services, the payer can write e-check, digitally sign it and effect the payment. Alternatively digitally signed Electronic Funds Transfer (EFT) instruction can be passed to the participating bank. Electronic Payment Service forwards the requests to respective banks and then it processes through the interbank payment mechanism.

Data mining supports banks to optimize their portfolio of services, and delivery channels. The record of past transactions can give helpful direction to the bank. Availability of information allows enterprises to improve their flexibility worldwide. Considerable developments in information technology have led to huge demand for sustainable analysis of resulting data in financial institutions and banking industries (Rajanish, 2002).

3.2. Customer Relationship Management (CRM)

3.2.1. Overview

Many industries focus on creating long lasting relationships with customers. This is so because they have to identify what satisfies customers best. Thus, the objective of the customer-centred industry becomes increasingly the share of wallet for each individual customer and costs by focusing on more targeted promotions. Nowadays, customers are the life line of any successful venture in the face of today's business competition both from the local and international markets. Therefore, organizations, specifically service oriented ones have shifted towards a long-lasting customer relationship management to gain a competitive advantage over their business competitors. Such organizations exert their maximum effort to keep their existing customers as well as attracting new customers.

Mudimigh et al. (2009) found that, CRM starts with the deep analysis of customer behaviour. The company that needs to create a personal interaction with customers should collect all the customer related data. As a result, they can achieve their customer satisfaction and behaviour by using IT infrastructure using the gathered data. CRM is very crucial for any organization through which the companies wanted to know the relationship between their customers and organization.

The necessity of CRM and identifying potential customers have been practiced by many businesses such as banks, insurance companies, and other service providers in order to acquire, retain, and maximize the value of customers. The relationship between IT and marketing helps organizations in creating strong coordination with customers. At this time CRM is a sound business strategy to identify the organization's most profitable customers and prospects, and devotes time and attention to expanding relationships with those customers through individualized marketing, decision making, and customized service-all delivered through the various sales channels that the organization uses (Onut and Erdem , 2000).

Based on Thearling (2004), data mining helps marketing users to target marketing campaigns more accurately, and align companies more closely with the needs, wants, and

attitudes of customers and prospects. If the necessary information exists in a database, the data mining process can virtually model any customer activity enabling to successfully implement better CRM.

Customer clustering and segmentation are two of the most essential data mining techniques used in all phases of CRM. The observation of Saarenvirta (1998) consolidates this idea by saying "Banks can use customers' data to divide them into segments based on 'shareholder value' variables as current customer profitability." Therefore, creating customer segments based on such variables highlights obvious marketing opportunities.

By and large, CRM is the only way for businesses to create a long-lasting attachment with their customers. Saarenvirta (1998) states, CRM is a process that manages the interaction between a company and its customers. The purpose of CRM initiatives is to allow customers to be treated individually during the marketing process. This differentiated marketing improves marketing results by making each marketing touch more appropriate and effective than is possible under any one size all marketing scheme.

To this end, the products and services of the organization are tailored to the needs and wants of individuals. The idea here is that an organization can't go after the whole market with the same strategy, rather identify a place in it to effectively and efficiently address the products and services it intends to offer. This change in focus from broad market segments to individual customer requires changes through the enterprise, but nowhere more than CRM, marketing, and sales (Saarenvirta, 1998).

Srivastava (1991), noted that, CRM is a broadly used term, and includes different functions such as 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 centred management (e.g., call management, integration of multiple contact channels, problem escalation and resolution, metrics and monitoring, logging interactions and auditing), among others.

Marketing is the process of planning and executing the conception, pricing, promotion and distribution of ideas, goods, and services. This is mainly to create exchanges that satisfy individual and organizational objectives. These days, customers that have real value to a company are the hub of marketing strategies. Accordingly, businesses have found it essential to acquire new customers as well as to keep those that have high value. Hence, one of the basic focuses in marketing is customer segmentation (Shiffaman and Kanuk, 1991).

Customer segmentation refers to an attempt to partition customers into several groups for the purpose of categorizing customer behaviours for a certain business strategy. Customers belonging to the group share similar characteristics. This in turn helps marketers to design the same promotion campaigns for customers that fall in a specific segment. However, segmentation requires the collection, organization and analysis of customer data (Bounsaythip and Rinta-Runsala, 2001). Especially, the banking industry should deal with individual level market segments in order to increase its services and profits, and to gain insight into the individual needs. Customer segmentation is at the heart of its effective and efficient marketing.

On top of this, Edelstein (2000) defined CRM in its broadest sense as a means of managing all customer interaction. CRM requires using information about the customers and to more effectively interact with customers in all stages. He has outlined the following as CRM phases:

- Acquiring customers
- Increasing the value of the customer and
- Retaining good customers

In all of the phases, data mining plays an essential role. For example, a certain bank can analyze its customers' database to identify their patterns and trends so as to target its particular service to a certain market group. It has become evident that data mining has a lot to contribute from the acquisition and retention of customers to business organizations achievement.

Due to the global business competition, many organizations today have identified the need to become more customers facing, as a result, CRM has become the issue of many organizational strategies. It is regarded as a business strategy targeted at obtaining a long term competitive advantage of optionally delivering customer value and extracting business value simultaneously. For the vast majority of business, the ability to acquire, retain and enhance customer relationship is the last place left to find as advantage (Bull, 2003).

It was, therefore, the aim of this investigation to build a data mining model that would enable to classify a new customer into one of the clusters. To come up with the data mining model, it was important to find the quantitative values of the cluster centroids of the clusters defined by qualitative values. To accomplish the clustering, the K-means algorithm was used. And the decision tree J48 and Multilayerperceptron ANN has been used for classification model.

3.2.2. Effective CRM and Reasons for Adopting CRM

According to Spartan InfoTech Co.WLL (2008), CRM is a many sided process, mediated by a set of information technologies. CRM is considered a business philosophy, in its ability to aid company strategy by adopting a customer centric approach. Therefore, being aware of CRM and what it means to their specific organization will go a long way in order to help them become competent in the marketplace of the world.

Gray and Byun (2001) stated, there are three components of an effective CRM, these are:

- Customer- The customer is the only and the major source of the company's present profit and future growth and development.
- Relationship: The relationship between a company and its customers involves continuous bi-directional communication and interaction. The relationship can be short-term or long-term, continuous or discrete, and repeating or one-time.
- Management: CRM is not an activity only within a marketing department rather it involves continuous corporate change in culture and processes. The customer

information collected is transformed into corporate knowledge that leads to activities that take advantage of the information and of market opportunities.

Consequently, an effective CRM puts these components together, equipped with tools, to put together a long term beneficial strategy of growth for the organization. Spartan InfoTech Co.WLL (2008) has listed the benefits of CRM in the following way:

- It develops better communication channels (i.e. the company and customers).
- It collects important data such as customer details and order history.
- Creates detailed profiles such as customer needs
- Delivers instant access to customer history
- Identifies new sales opportunities
- Simplifies marketing and sales processes by understanding customer needs

Overall it helps to:

- Improve the organization's ability to keep and acquire customers
- Maximize the lifetime value of each customer
- Improve service without increasing cost of service.

3.2.3. CRM implementation issues

CRM, as Girishankar's (as cited in Bull, 2003) statement, can be applied into different parts of the business. It has been suggested that organizations should adopt a holistic (analyzing whole system) approach, because it puts CRM at the heart of the organization with customer orientated business processes and the integration of CRM systems.

More to the point, outsourcing is another implementation issue of CRM. Since many organizations have few alternatives, they outsource a significant proportion of their CRM solution as they lack the resources to develop CRM software. Developing CRM software in-house can be a lengthy process and there are rewards to those that can respond rapidly and appropriately. CRM is also reputed to be facilitating the outsourcing of more business operations directly to the customer (Bull, 2003).

3.3. Customer Segmentation

CRM contains two most well-known applications. These are segmentation and customer profiling. Segmentation is defined as a word used to describe the process of partitioning customers into similar groups on the bases of shared and common attributes (like habits, taste etc), while customer profiling is describing customers by their attributes such as age, income, and lifestyles. This is done by building a customer's behaviour model and estimating its parameters. In other words customer profiling is a way of applying external data to a population of possible customers. Based on the data available, they can be used to prospect new customers or to drop out existing bad customers (Bounsaythip and Rinta-Runsala, 2001).

Customers' do have diversified preferences and behaviours. The transition from general to target marketing that is currently in progress creates demands for more sophisticated customer segmentation techniques. Customer data which is the key enabler of any segmentation strategy are the raw material that must be captured, integrated, and effectively analyzed in order to achieve the goal of profiling customers. Static customer profile data (example lifestyle, preferences) and customer behaviour data (example, usage, loyalty, and profitability) are both basic requirements for effective segmentation (Kelly, 2002).

The diversity and availability of communication technologies provide customers the power to access information on competitors, products, availability, and prices. Due to these reasons, business should become a customer centre approach. With the new challenges and competition in the world today, companies need to realize and understand their customers and to quickly respond to their preferences and needs. Companies have to identify the most valuable customers and the appropriate strategies to use in developing relationships with these customers. Such strategies would include developing one-to-one relationship with customers using market segmentation and Customer Relationship Management (Maalouf and Mansour, 2006).

Several segmentation approaches have been devised and each has advantages and disadvantages. Many experts suggest that most company's use a combination of

approaches for optimal benefit. The capability to use sophisticated segmentation techniques to the business enhances the ability of the enterprise to deliver more tailored marketing programs, to identify segments that are more important to the business, to identify segments that have been neglected and to become more attentive to previously unrecognized customer preferences (Kelly, 2002).

Market segmentation is the process of partitioning the heterogeneous market into separate and distinct homogeneous segments. A segment consists of a group of consumers who react in a similar way to a given set of marketing stimuli (Kelly, 2002).

According to Bounsaythip and Rinta-Runsala (2001), customer segmentation is a way to have more targeted communication with the customers. The process of segmentation describes the characteristics of customer groups called segments or clusters within the data. Segmenting is defined as putting the total population into segments according to their similarity. Customer segmentation is a preparation step for classifying each customer according to the customer groups that have been defined. Segmentation is crucial to cope with today's dynamically fragmenting consumer market place.

Kelly (2002) explained cluster segmentation as one of the customer segmentation strategies that seek to find out naturally occurring clusters of customers who share common characteristics or behave in the same way. Regardless of the segmentation technique used, the starting point is the collection of the data that provide the variables to construct the segments.

3.6. Data mining in CRM and Customer Segmentation

Data mining is a process that applies the techniques of AI to the task of finding useful and hidden patterns in data, and is providing particularly powerful in the identification of customers sharing the same characteristics. This segmentation of customers into affinity clusters presents new possibilities for customer segmentation (Kelly, 2002).

Two Crows Corporation (2005) explained that, data mining is becoming popular as it can be used to control costs as well as contribute to revenue increases. Due to this, many organizations are using data mining to support them manage all phases of the customer

life cycle, including acquiring new customers, increasing revenue from existing customers, and keeping good customers by identifying their characteristics.

It is known that CRM is a widely used term with a broad variety of functions. Not all CRM require data mining. Some of the functions of CRM are marketing automation, Sales force automation and contact centre management. But the focus is how data mining and analytics make these functions more effective. An increment of the lifetime values of the customer base in the context of a company's strategy is the basic objective of CRM. Understanding of the customer characteristics is the core activity of CRM. Customer understanding is the basis for maximizing customer lifetime value, which in turn encompasses customer segmentation and actions to increase customer loyalty, profitability, retention etc (Srivastava, 1991)**.**

Srivastava (1991) described, correct customer understanding and action leads to maximize customer lifetime value, where as if the customer is not wisely and correctly understood by the organization, it leads to hazardous actions. Therefore, focus should be given for correct customer understanding and intensive actions derived from it.

3.7. Data Mining Methods for Customer Segmentation

Clustering and classification are the two mostly used for customer segmentation data mining techniques. According to Saarenvirta (1998), customer clustering and classification are two of the most important data mining methodologies that are used in marketing and CRM. In addition to this, the business can divide customers into segments based on variables as current customer profitability, a measure of the life time value of a customer (Saarenvirta, 1998).

3.7.1. Clustering Techniques

In this study, the classification (supervised), and clustering (unsupervised) data mining tasks are experimented. The clustering (unsupervised) task is carried out to investigate the behavioral and optimal segments of customers. Clustering is a widely used technique in data mining application for discovering patterns in underlying data. In marketing, clusters formed for a business purpose are usually called segments, and customer

segmentation is the popular application of clustering. Segmentation and customer profiling are two of the most well known application of CRM (Berry and Linoff, 2004).

Clustering identifies natural clusters of records that share similar characteristics. In early days, market segmentation were done not based on customers' product or service usages, but clustering technique of data mining is at the heart of identifying the natural customer segmentation.

As Han and Kamber (2001), described, clustering is defined as the process of grouping a set of physical or abstract objects into classes of similar objects. A cluster is a collection of data objects that are similar to one another within the same cluster and dissimilar to the objects in another cluster.

Different algorithms can be used to do the clustering task of the data mining. Han and Kamber (2001) categorized these different algorithms into partitioning, hierarchical, density-based, grid-based, and model-based methods. But from these, the most widely used is the K-means algorithm which is the partitioning method and different researchers have successfully been using the k-means algorithm to customers' segmentation (Han and Kamber, 2001). This research also uses the K-means algorithm to undertake the clustering task.

The reasons of using K-means algorithm is because of its betterment and easiest way of using it. Customer segmentation is best achieved using K-nearest neighbour technique clustering algorithm. The K-means algorithm is relatively efficient in processing large datasets. The researcher had also tried to study other clustering algorithm, which is the Expectation Maximization (EM). According to Han and Kamber (2001), EM extends from the K-means algorithm and it is suggested when there are too much missing values of the variables. Besides this, the reason of choosing K-means algorithm in this research is, it is easy to interpret and visualize the results obtained. The k-means algorithm has basically different steps to do the clustering task. The first step is, the algorithm randomly selects K data points to be the seeds. And secondly, it assigns each record of the closest seed; one way to do this is by finding the boundaries between two clusters. The boundaries between two clusters are the points that are equally close to each other.

Lastly, it calculates the centroids of the clusters and the centroids become the seeds for the next iteration of the K-means algorithm.

3.7.2. Decision Trees

After the clusters have been created using the clustering algorithms, they need to be interpreted. Even though there are different approaches to perform this, one of the approaches widely used for understanding clusters is building a decision tree with the cluster label as the target variable, and using it to derive rules explaining how to assign new records to the correct cluster (Berry and Linoff, 2004).

According to Two Crows Corporation (1999), decision trees are a way of representing a series of rules that lead to a class or value. Classification trees label records and assign them to the appropriate class.

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

The training process that creates the decision tree is called induction and requires a small number of passes through the training set. Most decision tree algorithms go through two phases, a tree growing (splitting) phase followed by a pruning phase (Bounsaythip and Rinta-Runsala, 2001). In splitting (tree growing) phase, the tree growing is an iterative process which involves splitting the data into progressively smaller subsets. The first iteration considers the root node that contains all the data.

After a tree is grown, one can explore the model to find out nodes or sub trees that are undesirable because of overfitting or rules that are evaluated inappropriate. Pruning removes splits and the sub trees created by them. Pruning is a common technique used to make a tree more general. Algorithms that build trees to maximum depth will automatically invoke pruning (Bounsaythip and Rinta-Runsala, 2001).

Decision tree is selected for different reasons. Some of these reasons are: Decision trees are easy to understand, decision trees are easily converted to a set of production rules, and so on. There are different decision tree algorithms, and C4.5 is among one of the

algorithms that includes methods to generalize rules associated with a tree; this removes redundancies. J48 is an improved version of the C4.5 (Han and Kamber, 2001).

The processes of the J48 algorithm to build a decision tree are to:

- Choose an attribute that best identifies the output.
- Create a separate tree branch for each value of the chosen attribute
- Divide the instances into subgroups so as to reflect the attribute values of the chosen node.
- For each sub group, terminate the attribute selection process.

The algorithm is applied to the training data. The created decision tree is tested on a test data set, if available. If test data is not available, J48 performs a cross-validation using the training data itself.

3.7.3. Artificial Neural Networks (ANNs)

Neural Network is a field in Artificial Intelligence (AI) where we use data structures and algorithms for learning and classification of data, by inspiration from the human brain. Many tasks that humans perform naturally fast, such as the recognition of a familiar face, proves to be a very complicated task for a computer when conventional programming methods are used. By applying Neural Network techniques, a program can learn by examples, and create an internal structure of rules to classify different inputs, such as recognizing images (Nielsen, 2001).

According to Berry and Linoff (2004), ANN started to allow machines think and reason out as human beings do. Neural networks used a means of modeling large and complex problems in which there may be hundreds of predictor variables that have many interactions. It is used in classification problems of data mining such as customer profiling, financial predictions and so on.

Neural networks essentially comprise the architecture or model, the learning algorithm, and the activation functions. Neural networks are programmed or trained to store, recognize, predict, and associatively retrieve patterns or database entries to solve combinatorial optimization problems to filter noise from measurement data. In general, it

is used to estimate sampled functions when we do not know the form of the functions. Neural nets are best distinguished from other intelligent techniques in that they are non-rule based and can additionally be made stochastic so that the same action does not necessarily take place each time for the same input. Neural networks have emerged as advanced data mining tools in cases where other techniques may not produce satisfactory predictive models (Nielsen, 2001).

A neural network starts with an input layer as shown in **figure 3.1**, where each node corresponds to a predictor variable. These input layers are connected to different nodes of the hidden layer. The hidden layer nodes may also be connected to other hidden layers or to an output layer. Each node takes in a set of nodes, multiplies them by a connection weight, adds them together then applies a function to them, and passes the output to the node in the next layer.

As Han and Kamber (2001) stated, training is defined as the process of iterating through the training set to adjust the weights. The neural network is used to estimate the connection weights so that the output of the neural net accurately predicts the test value for a given input set of values. Back propagation is the most common training method, and each training method has a set of parameters that control various aspects of training such as adjusting the speed of conversion. Even though the neural net can provide reasonable predictions for any data quality, it needs a careful data pre-processing, preparation selection and data cleaning.

Neural network is a set of connected input and output units where each connection has a weight associated with it as depicted below:

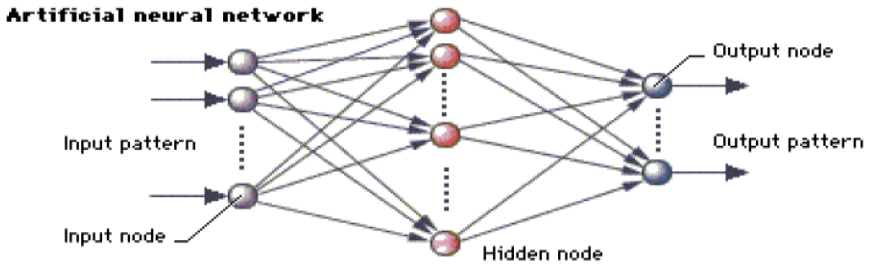


Figure 3. 1: A multilayer feed-forward neural network

A neural net has high tolerance to noisy data as well as ability to classify patterns on which they have not been trained. Moreover, several algorithms have recently been developed for the extraction of rules from trained neural nets. The conditions contribute towards the usefulness of the neural networks for classification in data mining.

3.7.3.1. The multilayer perceptron (MLP) or Multilayer feedforward network

Han and Kamber (2001) described the feed-forward back-propagation as one of the common types of the neural network. The Multilayer perceptrons (MLPs) is one of the feed-forward architecture which uses inner products. Building on the algorithm of the simple perceptron, the MLP model not only gives a perceptron structure for representing more than two classes, it also defines a learning rule for this kind of network. The MLP is divided into three layers: the input layer, the hidden layer and the output layer, where each layer in this order gives the input to the next. The extra layers give the structure needed to recognize non-linearly separable classes.

The perceptron algorithm can be trained by adjusting the weights of the inputs with supervised learning. In this learning technique, the patterns to be recognized are known in advance, and a training set of input values are already classified with the desired output. Before commencing, the weights are initialised with random values. Each training set is then presented for the perceptron in turn. For every input set the output from the perceptron is compared to the desired output and if the output is correct, no weights are altered. However, if the output is wrong, we have to distinguish which of the patterns we would like the result to be, and adjust the weights on the currently active inputs towards the desired result (Nielsen, 2001).

Han and Kamber (2001) explained that, back propagation learns by iteratively processing a set of training samples, comparing the network's prediction for each sample with the actual known class label. For each training sample, the weights are modified so as to minimize the mean squared error between the networks prediction and the actual class. These modifications are made in the backward direction, i.e. from the output layer, through each hidden layer down to the first hidden layer. Back first initializes the weights

in the network to small random numbers, then the network gets a training example and, using the existing weights in the network, it calculates the output. Back propagation then computes the error by taking the difference between the calculated result and the expected one. Finally, the error is feed- back through the network and the weights are adjusted to minimize the error.

3.8. Review of Related Works

Customer Relationship Management (CRM) is one of the principal parts in expanding company's customers because it helps to identify customers' behaviour and customer segmentation. Several researchs are conducted in the area of CRM, some of which are reviewed henceforth.

Kumneger (2006) did a research on "Application of data mining techniques to support customer relation management for Ethiopian shipping lines (ESL)". Kumneger used 60% of the dataset for training, 30% of the dataset for testing, and the remaining 10% for validation set. She has indicated that, her initial intension was to segment similar customers into one group according to their behaviour. Finally the result generated the possibility of segmenting similar customers regarding to their income generation. She has applied a CRISP-DM methodology to accomplish her research. She has indicated that domain experts of the shipping lines have appreciated the result. In her clustering model, she used the value of k=3, 4, and 5 for the clustering model and got a clustering accuracy of 98.37%, 98.62%, and 97.88% respectively and obtained an overall classification result of 98.55%. Hence, from the above result, the researcher concluded that the used models (i.e. k-means clustering and decision tree classification) are appropriate for customer segmentation in Ethiopian shipping lines.

Tilahun (2009) also conducted a research on "Possible Application of Data mining Techniques to Target Potential Visa Card Users in Direct Marketing at Dashen Bank S.C.". Tilahun studied the possible application of data mining techniques in identifying potential customers. In his research, he used a decision tree J48 classification algorithm and K-means clustering algorithm to help identify potential customers at the company. He applied the CRISP-DM methodology. In his research different CRM concepts were

discussed and CRM is described as the best marketing strategy for acquiring, and retaining customers to create potential customers' for the banking industry.

Tilahun used k-means clustering algorithm to segment dissimilar customers and decision tree J48 algorithm for classification. The study was focused on identifying potential visa card users in direct marketing contribution and about 5110 records and 8 attributes obtained from Dashen Bank S.C. were used for his research. For the purpose of data understanding, data pre-processing, and model building WEKA data mining tool was used. Different values of K were tested, i.e. k=5, 4, and 3. The best result was found when k=3, where all the three clusters were different and meaningful. He selected the clustering as good based on the domain experts' opinion and dissimilar segment achievement.

Finally, the decision tree was used to classify the customers into one of the three clusters and 96.14% overall classification accuracy was obtained and considered as a better classification accuracy. In conclusion, he recommended and emphasized the importance of taking further data mining research in the banking industry with different classification like ANN and clustering techniques and parameter values.

Henok (2002) also worked on the "Application of data mining techniques to support Customer Relationship Management at Ethiopian air lines (EAL)". This is also a paper which has been done on the area of CRM, and he has tried to apply a data mining technology in order to increase the productivity of EAL. The modeling used were the decision tree and k-means clustering algorithm in segmenting the customer data into similar groups. Generally the result he has obtained was satisfactory and shows the possibility of clustering customers by using the k-means clustering algorithm. The researcher concluded that knowledge of data mining, marketing strategies, and business organizations should be integrated in order to implement CRM.

There are different researchs that have been conducted in the banking industry. Some of these are the work of Askale (2001) and Meretework (2004). These two researchs showed applying data mining technology in the banking industry is important because banks maintain large amount of data in different operations.

CHAPTER FOUR

DATA PREPARATION AND DATA PREPROCESSING

This chapter discusses about data collection, data preparation or pre-processing tasks of the data mining. This embraces detail description from which the research data has been collected. The data selection processes, data cleaning, transformation, and integration activities are explained in this chapter. This chapter also describes the methods and tools used in this study.

4.1. Data Collection and preparation

The data employed in this research was collected from Dashen Bank S.C. head office. The original data of the card-holder customers' profile as showed in **Appendix 1** of the Dashen Bank S.C. was taken and all the necessary operations of data selection and data preparation were carried out. Originally, information about the card-holder customer is recorded when the individual is registered and took the card as well as the transaction of the customer is also recorded whether he/she uses the ATM or EFT of POS service. The database of the Dashen Bank contains more than 110,000 card holder customers; still there are a number of records entered to the automated system all the time, however, after discussions and survey of the data with the domain experts and database administrators, the data taken were constructive as well as sufficient and could be used for a realistic data analysis.

4.1.1. Description of the data collected

The relevant data to carry out this research have been collected from the customers profile database. The collection has about 110,000 records and 20 attributes. The fields that were obtained from the collection are stated below in **Table 4.1.**

Attribute Name	Data type	Description
Card_holder_ID	Text	The customers card holder ID
Available Balance	Number	This is the balance of the customers in the bank
Office tel	Number	This is the office telephone number of the customer
Date of birth	Date	This is the date of birth of the customer
Occupation	Text	The type of job of the customer
ATM	Text	This are Automatic Teller Machine (ATM) users of the bank
POS	Text	This are EFT users on Point of Sale terminal service users
Marital_Status	Text	This is a field that shows whether the customer is married or not
Year_of_Join	Text	This is the date in which a customer joins
Saving Account	Number	The type of account in which the customers creates for his/her personal use (Non PLC account)
Current Account	Number	An account type created by customers that is used for business purpose
Kebelle	Text	Kebelle of the customer
House Number	Text	This is the house number of customer
Home Tel Number	Number	This is home tel number of a customer
Mobile_Number	Number	This is mobile number of customers
Address	Text	This is the address of the customer by city
Nationality	Text	Nationality of the customer
Sex	Text	This is the gender of the customer
Branch	Text	The branch in which the customer uses
Position	Text	The position of the customer in his/her company

Table 4. 1: The attributes of customers table

4.2. Method Selection

In doing this project, CRISP data mining method is used. The CRISP-DM method is described in terms of a hierarchical process model, consisting of sets of tasks described in six levels: business understanding, data understanding, data preparation, modeling, evaluation and deployment. This model is selected because it is flexible and easy to work with (Pete, 2000).

4.3. Data Mining Goals

The first important step in the whole data mining process is to understand the need to do data mining. Identifying the goal of the data mining process is a prerequisite to discover knowledge from the database. The goal of the data mining process depends on the type of problem to be solved using data mining technology. So, before starting the actual data mining task, we should be able to clearly define our problem and also have a good understanding of our data to be used for the data mining task.

As mentioned in chapter 1 section 1.4, the main objective and output of this research is to find interesting and meaningful patterns and relationships in Electronic Fund Transfer (EFT) of POS service at Dashen Bank S.C. Provided that meaningful relationship among attributes were to be established, Expansion of Electronic Fund Transfer could have a better service to customers in the banking industry of Ethiopia and thus could develop strategic solution to spread out the usage of EFT in POS terminals. After we define the goal of our data mining task, we should be able to select an appropriate data mining tool which can perform the expected functions.

The business survey undertaken by the researcher has discovered that current customer value is based on a reliable revenue value and individual visa card or electronic transaction usage activities of each customer. Therefore, the first data mining goal was to identify the customer segments and describe the resulting clusters. In addition, the variables that determine customer value will be used to derive customer segments, and the subsequent classification rules. The most appropriate data mining techniques, which are clustering (or segmentation) and classification has been used for this purpose.

In order to provide customer classification rules the researcher have consulted with the domain experts and the main focus was on the inspection and use of the important attributes for customer segmentation. This process allows obtaining meaningful clusters where each member customer of a cluster will be identified by the cluster index of the group. This cluster index of each customer will in turn be used as input to the classification algorithms to generate the required classification rules.

4.4. Data Mining Tool Selection

The basic aim of using data mining tool is to discover hidden knowledge from a large database. The selection of an appropriate data mining tool for this research was done based on a certain criteria, such as the algorithm supported, the operating system on which the software runs the possible formats for the data that is to be analyzed, the maximum number of records the software can handle, and visualization capabilities. To this end, the researcher used an open source data mining tool, WEKA, which is developed by the University of Waikato (New Zealand) (Shigeki, 2006). WEKA includes J48 which is an improved version of C4.5. WEKA was selected since the researcher is familiar with this tool.

4.5. Business Understanding

To understand the banking business and the data mining problems, the researcher observed Dashen Bank S.C. head office and some area banks as well as some super markets in which EFT of POS systems are deployed. Domain expert are consulted to have brief understanding on the problem area. Besides, the dataset is also thoroughly examined by the researcher and domain experts.

4.6. Data understanding

After setting up the problem and building a simple plan for its solution, the researcher proceeded with the central item in data mining process – data understanding. There are several things to be learned about the data before the actual application of data mining techniques. Data is the principal part of data mining tasks, methods, and tools. In ideal cases, the required data would already be resident in a corporate data warehouse,

cleansed, available, historically accurate, and frequently updated (Berry and Linoff, 2004). However, this is not the case in most situations.

The researcher collected data of all card-holders starting from the year the ATM or electronic fund transfer service has been introduced i.e. 2006 up to the year 2011. Hence, for the present research five years (2006 to 2011) card-holders has been considered. These five years data contains more than one hundred ten thousand records.

4.7. Data pre-processing

The data pre-processing task is one of the challenging parts of this study. It is time consuming and it has taken much effort. Before feeding data to data mining, one has to make sure the quality of data by measuring its accuracy, completeness, consistency, believability and interpretability. As stated by Han and Kamber (2006), the data processing task of data mining includes data cleaning, data integration, data reduction, and data transformation. The card-holder customers' record database system of Dashen Bank S.C., which is employed for the purpose of this study, suffers from a number of limitations. These include missing values, outliers and encoding inconsistency in some of the attribute values. In fact, as stated by Witten and Frank (2000), one of the critical problems in building data mining models is limitations in the data itself. An optimal model could only be constructed after a comprehensive, clean and automated data is prepared.

4.7.1. Data cleaning

The basic and major task in data mining is preparing the data in a way that is suitable for the specific data mining tool or software package. Real world databases contain incomplete, noisy and inconsistent data. Thus, such unclean and incomplete data may cause confusion for the data mining process (Han and Kamber, 2001). As a result data cleaning has become a must in order to improve the quality of data and the accuracy and efficiency of the data mining techniques.

In order the data mining algorithm to perform quality mining, there should be a cleansed data that is free from outliers, noise, and with no missing values of an attribute. Some of

the attributes used in this research are derived from the base attributes, for example, Age_classification attribute was derived from date of birth attribute (Age_classification attribute is obtained after calculating age and used binning methods for age _classification field) through the Excel aggregate function. However, the researcher tried to do the data cleaning such as handling missing attribute values, and noise removal using Microsoft Excel filtering mechanism before deriving the attributes that are used for analysis. The following points briefly describe the methods employed in handling missing, noisy, inconsistent, and outlier data.

- **Missing values:** Different methods can be used to handle missing values, such as discarding the record, filling in the values by the most probable values (for nominal values) and mean average (for numeric values).
    - After a thorough look at the card-holder customers' data, the researcher identified very small data that had missing values in "Date of birth" attribute. The "Date of birth" field contained about 56 missing values from the total of 110,000 records. This was easily identified when a derived attribute "Age" from date of birth was calculated. Hence, the researcher has filled the dataset manually by taking the modal value of Age_classification which was"Age_group_One".
    - There were also 143 missing values in the available balance field. Then, the researcher used the mean value of the monthly available balance of the EFT or payment card system customers by their type of occupation, and thus the missing values have been replaced by the calculated result.
- **Outliers:** An outlier is a single, or very low frequency, occurrence of the value of a variable that is far away from the bulk of the values of the variable. Outliers may distort the remaining data to the point of uselessness (Dorian, 1999). In this study after the field "Age" is derived, there were about 40 records whose ages were less than 18 (Which is not allowed by the bank to be a bank customer) and 80 records whose age were greater than 65 (according to the payment card payment system consultation, such age is rarely existing as a result the researcher considered such age as an outlier). Therefore, the researcher discussed with the

domain experts and decided to change these outliers by calculating the mean average of the total records. Then these values (outliers) are replaced by the new calculated result as described below.

Average Mean= Sum (all values of age)/Total number of records, this was calculated in Microsoft Excel aggregate functions and result found is 38.778 which is rounded to 39. All the age outliers are replaced with this value using the filtering mechanism of Microsoft Excel.

- **Noisy and inconsistent data:** Noise is a random error or variance in a measured variable (Han and Kamber, 2001). In this research the dataset was very large, and it was time taking to handle the noisy data. The following were the fields with noisy data.
    - The Date_of_birth field was written in Ethiopian and Gregorian system. The researcher then identified the problem and accordingly about 120 records were in Ethiopian system. This was identified easily because the Ethiopian system was written by including **E.C.** at the end of the date, and the remaining records were in Gregorian system. The researcher has decided to change the Ethiopian system to Gregorian type manually and age was calculated accordingly. Moreover, this attribute was written using different standards and formats. This was identified when age field was derived from it and the researchers have changed it to a common format.
    - There were also inconsistent values in the sex field. Most of the records were written as M and F values and some of the records in the dataset were written as 000 and 001 to represent Male and Female values respectively. The data encoders entered the values as per their understanding as it was stated in the field part of "Sex" as **Sex 000 M, 001 F.** Thus some of them entered in terms of M and F and some of them entered as 000 and 001. Generally, the field leads them to enter different values. The researchers changed 000 and 001 as M and F respectively assuming that these values could be easily understood when the result is analyzed.

4.7.2. Data Integration

Data mining usually requires merging of data (data integration) from multiple data stores. The data may also need to be transformed into forms suitable for mining (Han and Kamber, 2001). The datasets which were used for the subsequent model building are prepared and derived from one source, the cardholder customers' profile. Hence, no need of merging of databases.

4.7.3. Data Reduction

Commonly the data collected for analysis from different organizations is very huge, which is sure to slow down the mining process. Hence, the size of the data set should be reduced before the data mining results. Data reduction obtains a reduced representation of the dataset that is much smaller in volume, yet produces the same (or almost the same) analytical results (Han and Kamber, 2001).

4.7.3.1. Dimensionality reduction (Attribute Selection)

In theory, relevant attributes could be determined by decision tree for classification automatically using the concept of information gain or entropy without manual efforts. Nevertheless, it is common to find many attributes that could not be helpful for some task of classification or any other purpose before adopting any technique. In this research, certain attributes were excluded from being considered in the data mining in order to simplify the task of decision tree. The cardholder customer database originally contains about 20 attributes and the researcher discussed with domain experts to decide relevant attributes for this study. The discussion resulted in 10 relevant fields after deleting attributes that are not relevant for analysis. The attributes that were removed can be viewed as belonging to one of the following three groups.

- ✓ **Fields with similar information content:** When fields with the same information content were encountered, only one of the fields was considered. For instance, Age_classification was chosen over Date of birth, as well as Year_Of_Join was chosen over Date_of_Join which holds the same information.

- ✓ **Non-variant fields:** Attributes that take a value that holds the same value for all the records in the database were also dropped, for example, the attribute Nationality. There were about 98 records with other citizens and the researcher and the bank experts removed this attribute as it is less relevant when compared with the total number of records. Almost all the card holder customers' data collected were of Ethiopian nationality. Thus, since Ethiopian was applied to all the cases, the attribute was considered as being irrelevant.
- ✓ **Fields taking many different values:** Attributes that take different values for all records were excluded. Accordingly, the following attributes were also disregarded since they take many different values. This is done with the intention to improve the speed, accuracy of analysis and training. For example, the following fields contain different values for all records :
    - o Card_holder_ID, Telephone-Number (Home-tel and office-tel), House-Number, Kebelle, and Mobile-Number.
    - o Address: This attribute were also excluded because it contained very detailed values which was a challenge for analysis.
    - o Kebelle: there are unmanageable Kebelle's in the cities. This is too detail, a data for rule mining, which is computationally expensive. Having more attributes will make the exploration task difficult.
    - o House_Number: There was House_Number field in the database and this is very detailed and difficult for analysis. Hence the researcher discussed with domain experts and decided to exclude this field from further considerations.

In this research, the standard of considering attributes as being complete as long as at least 70% of the records comprise values is adopted (SPSS INC, 1991). Accordingly, all the 20 attributes satisfy this criterion. Even though some of the attributes (i.e. Date of birth and Monthly Available Balance) had missing values, the attributes Date of birth and Monthly Available Balance constitutes 5% and 13% missing values respectively and thus these missing values could not lead to exclude this attribute as stated by SPSS INC (1991).

Nevertheless, the attribute Date of birth was excluded from the dataset because Age-classification field were derived from Date of birth. Age was calculated as current time from the system minus date of birth for classifying age as Age_Group_One, Age_Group_Two, and Age_Group_Three. The age attribute with real/numerical data type was converted into a nominal data type and it was derived for that purpose. Finally, the researcher and domain expert's selected 10 attributes that were deemed to be relevant for the exploration purpose. The selected attributes are: Sex, Occupation, Saving-Account, Current-Account, Marital_Status, Year_Of _Join, Monthly-Available-Balance, Age-classification, ATM, and POS

4.7.3.2. Numeriousity Reduction (Size Reduction)

As Han and Kamber (2001) described, in numeriousity reduction, the data are replaced or estimated by alternative, smaller data representations such as parametric models (which need to store only the model parameters instead of the actual data) or nonparametric methods such as clustering, and sampling.

In this research, the payment card system customers' record of the Dashen Bank S.C. was the target data set. These data are collected from the Dashen Bank S.C.

To select samples of payment card system customers, a stratified simple random sampling method was used with regard to their type of occupation. To select proportional samples of card holder customers' from their type of occupation, the total customers were stratified into two strata (i.e. Hired and Private) and from each strata, 10% sample records from a total of 56,500 records of hired customers and 10 % records from a total 53,500 records of private customers were selected randomly using SPSS. Each stratum was imported from Microsoft Excel into SPSS separately. Then, simple random sampling from each stratum has been used to select a total of 5650 hired (type of occupation) records and 5350 private card holder records randomly. The researcher has obtained a total of 11,000 (from both hired and private) records randomly for this study from the whole 110,000 records by using SPSS. Even though theoretically large volume of data is important to train data mining models, due to time constraint to visualize, interpret and to analyze results of the data, sample data were selected. Simple random sampling was

preferred in order to give each record an equal chance of being selected from each stratum.

4.7.4. Data Transformation

According to Han and Kamber (2001), Discretization is one of the transformation methods which can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace the actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels reduces and simplifies the original data. This leads to a concise, easy to use, knowledge-level representation of mining results. Discretization can be performed recursively on an attribute using Binning (divides values into intervals) and Concept hierarchy generation (organizes concepts i.e. attribute values hierarchically) methods. Concept hierarchies can be used to reduce the data by collecting and replacing low level concepts for example numerical values for the attribute "age" with higher level concepts such as youth, middle-aged, senior. Binning is one of the discretization methods, and is a top-down splitting technique based on a specified number of bins. These methods are also used as discretization methods for size or numeriousity reduction and concept hierarchy generation.

In this study, Binning was used as discretization methods for concept hierarchy generation. The derived attribute "Age" stored detailed numerical values (like 18, 19, 20, 21, 22,….65), as a result the researcher and domain experts have decided to change these low level detailed numerical values into higher and generalized concepts such as Age_Group_One, Age_Group_Two, and Age_Group_Three using equal-width (distance) partitioning which is method of simple discretization Binning. The researchers have chosen discretization (binning) due to the following reasons.

- The generalized data may be more meaningful and easier to interpret.
- Mining on a reduced data set requires fewer input/output operations and is more efficient than mining on larger, ungeneralized dataset as well as reduces computational value.
- If the values are detailed or distinct, it minimizes pattern quality

- The generalized data set has special pattern identification

Due to these benefits, discretization techniques and concept hierarchies are typically applied before data mining as a pre-processing step.

Dorian (1999) states, equal-width (distance) partitioning divides the range into N intervals of equal size. If A and B are the lowest and highest values of the attribute respectively, the width of intervals will be given as: W= (B-A)/N and accordingly the researcher has computed the width as:

> Width= (Max-Min)/Number of intervals. In this study, the maximum Age was 65 and the minimum Age was 18 and the number of intervals were 3; Age_Group_One, Age_Group_Two, and Age_Group_Three. Therefore, width= (65-18)/3=16. After this, the researchers have sorted the values of "Age" in ascending order and applied **Equi-width binning** – for bin width of 16. Hence, the concepts (Age_Group_One, Age_Group_Two, and Age_Group_Three) has been assigned as
>
> - Bin 1(Age_Group_One) ----------> [18, 34) bin-this ranges of values of Age have been replaced by Age_Group_One.
> - Bin 2(Age_Group_Two) ----------> [34, 50) bin-this ranges of values of Age have been replaced by Age_Group_Two.
> - Bin 3(Age_Group_Three) ----------> [50, +) bin-this ranges of values of Age have been replaced by Age_Group_Three .

The same procedures were done for the Monthly-Available-Balance field to change the numeric values to generalized concepts as Low, Medium, High, and Veryhigh. The dataset were in Microsoft Excel (.xls) format and this data format can't directly be processed by WEKA data mining tool. Hence, transformation was performed. The Microsoft Excel .xls dataset were saved in CSV (Comma Separated Value) format. Sample data is shown below:

Sex,Occupation,Saving_Account,Current_Account,Marital_Status,Monthly_Available_Balance,Age_Classification,Year_Of_Join,ATM,POS

M,Hired,NO,YES,Unmarried,Low,Age_Group_Three,Joined_2010_11,YES,NO

M,Hired,NO,YES,Unmarried,Medium,Age_Group_Three,Joined_2010_11,YES,NO

M,Hired,NO,YES,Unmarried,Medium,Age_Group_Three,Joined_2010_11,YES,YES

M,Private,YES,YES,Married,High,Age_Group_Three,Joined_2010_11,NO,NO

M,Private,YES,YES,Married,Medium,Age_Group_Three,Joined_2010_11,NO,NO

M,Private,YES,YES,Married,Low,Age_Group_Three,Joined_2010_11,YES,NO

M,Private,YES,NO,Married,High,Age_Group_Three,Joined_2010_11,YES,NO

In order to make the dataset convenient for the data mining tool, the dataset was prepared in ARFF (Attribute Relation File Format) file format. In doing so, declarations, such as *@relation <relation name>*, *@ attribute <attribute names>* together with the data type and *@data* are added in the CSV file. Bouckaert et al. (2008) has defined Relation, Attribute, and Data as below:

- @Relation is defined as the first line in ARFF file and represents the database title.
- @Attribute defines the name of the attributes and its data type
- @data declaration is a single line denoting the start of the data segment in the file.

The data that were converted into ARFF file format has been used for the experimentation. The predefined dataset is printed below.

@relation bank

@attribute Sex {M, F}

@attribute Occupation {Hired, Private}

@attribute Saving_Account {YES, NO}

@attribute Current_Account {YES, NO}

@attribute Marital_Status {Married, Unmarried}

@attribute Available_Balance {High, Low, Medium, Veryhigh}

@attribute Age_Classification {Age_Group_One, Age_Group_Two, Age_Group_Three}

@attribute Year_Of_Join {Joined_2010_11, Joined_2006_07, Joined_2008_09}

@attribute ATM {YES, NO}

@attribute POS {YES, NO}

@data

M,Hired,NO,YES,Unmarried,Low,Age_Group_Three,Joined_2010_11,YES,NO

M,Hired,NO,YES,Unmarried,Medium,Age_Group_Three,Joined_2010_11,YES,NO

M,Hired,NO,YES,Unmarried,Medium,Age_Group_Three,Joined_2010_11,YES,YES

M,Private,YES,YES,Married,High,Age_Group_Three,Joined_2010_11,NO,NO

M,Private,YES,YES,Married,Medium,Age_Group_Three,Joined_2010_11,NO,NO

M,Private,YES,YES,Married,Low,Age_Group_Three,Joined_2010_11,YES,NO

M,Private,YES,NO,Married,High,Age_Group_Three,Joined_2010_11,YES,NO

In data formatting, the data was converted into .ARFF format, which is suitable for the WEKA data mining tool and thus the above was saved as .ARFF file extension and submitted to the tool.

CHAPTER FIVE

EXPERIMENTATION

5.1. Overview

In this chapter, the researcher describes the techniques that have been used in pre-processing and model building. Thearling (2004) discusses, in order to enable successful CRM; the initial task is to identify market segments containing high profit potential customers'. Accordingly, the main objective of this research was to provide a model that classifies customers with respect to the important dimensions of EFT of POS service customers' usage behaviour and corresponding value. This research project incorporated the typical stages that characterize a data mining process. Hence this study is organized according to the CRoss-Industry Standard Process for Data Mining (CRISP-DM) process cycle, which is described and discussed in section 2.3.1 and shown in Figure 2.1 of chapter two. In this section of the study, the researcher discusses the experimentation process by relating the steps followed, the choices made, the tasks accomplished, the results obtained, the evaluation of the model and results, and presented the result in a way that the company can easily understand and use it.

5.2. Modeling

Modeling is one of the basic phases of the CRISP-DM process. In this phase, various modelling techniques are selected and applied and their parameters are adjusted to optimal values. Typically, there are several techniques and tasks for the same data mining problem type. Some of the tasks include selecting the modeling techniques, generating the test design, Building a model, and assessing the model used.

5.2.1. Selecting the modeling techniques

According to Huang (1998), partitioning a set of objects in databases into homogeneous groups or clusters is one of the basic operations in data mining. It is useful in a number of tasks, such as classification and segmentation. For example, by partitioning objects into clusters, interesting object groups may be discovered. To this end, clustering is a popular approach to implementing the partitioning operation and it was selected as it is a direct data mining technique where there are no predefined classes to be predicted. Clustering

methods partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters. After the segments are identified, new customers should be classified to one of these cluster indexes. For the classification purpose the decision tree was used because it has a powerful visualization features, it is easy to understand and interpret, and finally it is one of the widely used classification method.

There are many clustering algorithms that can be used in WEKA 3-7-2. Some of these include CLOPE, DBScan, Expectation Maximization (EM), SimpleKMeans, Optics, Cobweb and so on. However, the researcher selected SimpleKMeans (K-means) algorithm due to its popularity and simplicity. The k-means algorithm is a well known and widely used clustering algorithm, and is classified as a partitional or nonhierarchical clustering method (Huang, 1998). As Xiong et al. (2006) described, People have identified some characteristics of data that may strongly affect the K-means clustering analysis including high dimensionality, the size of the data, noise and outliers in the data, types of attributes and data sets. But in this particular research, missing values, outliers and noisy data are controlled as much as we can in the pre-processing part. Customer segmentation task is done with k-means algorithm. Determining the value of K is the difficult task in the k-means algorithm, because it determines the optimal clustering model that creates the dissimilar segments of customers. The optimal number of cluster size, that means the value of k, is obtained through trial of different models (Berry and Linoff, 2004). This is because K is user defined number. Srivastava (1991) advises that the number of clusters chosen should be driven by how many clusters the business can manage. Accordingly, the business experts have been consulted in setting the optimal value for the same. Due to their consultation, different experimentation with different values of k has been used, and thus about 4 cluster models have been experimented and evaluated.

Classification is a different technique than clustering. Classification is similar to clustering in that it also segments customer records into distinct segments called classes. But unlike clustering, a classification analysis requires that the end-user/analyst know ahead of time how classes are defined. For the purpose of classification, there are a

number of decision tree algorithms available in WEKA 3-7-2. Some of these include NB Tree, ID3, J48, FT, ADTree and so on. J48 algorithm was selected because it is most widely used algorithm, and it has a visualized description of the output etc. Similarly, Multilayerperceptron algorithm was selected in a neural net algorithm.

5.2.2. Generating the Test Design

In the test design part, before we actually build a model, we need to generate a mechanism to test the model's quality and validity. For example, in supervised tasks such as classification, it is common to use error rates as quality measures for data mining models. Hence, we should partition the dataset into train and test set. Creating the training dataset, which is used to build the model and a test dataset, which estimates the quality of the model is essential. It is also important to decide how to divide the available dataset into training data, test data and validation datasets. Large proportion of the total data available is used for training, where as testing is done on a small amount of data when compared with training in classification. After running the modeling tool on the prepared dataset to create one or more models, the resultant model should be described as well as the interpretation of the qualities of the models (e.g. in terms of accuracy) and any difficulties encountered should be reported and documented.

Another model is clustering; it is an automated process to group related records together. Related records are grouped together on the basis of having similar values for attributes. This approach of segmenting the database via clustering analysis is often used as an exploratory technique because it is not necessary for the end-user/analyst to specify ahead of time how records should be related together. In fact, the objective of the analysis is often to discover segments or clusters, and then examine the attributes and values that define the clusters or segments. As such, interesting and surprising ways of grouping customers together can become apparent, and this in turn can be used to drive marketing and promotion strategies to target specific types of customers (Colet, 1997). As clustering is unsupervised learning, all the dataset records were used for the training purpose, where the algorithm is provided with data points without labels, the purpose is to find appropriate representation of the underlining distribution of the data.

After selecting and testing the modeling techniques, we should build the clustering model followed by decision model using the selected tools.

The following is the figure that shows the basic configuration parameters of the WEKA 3-7-2 for the K-means algorithm.

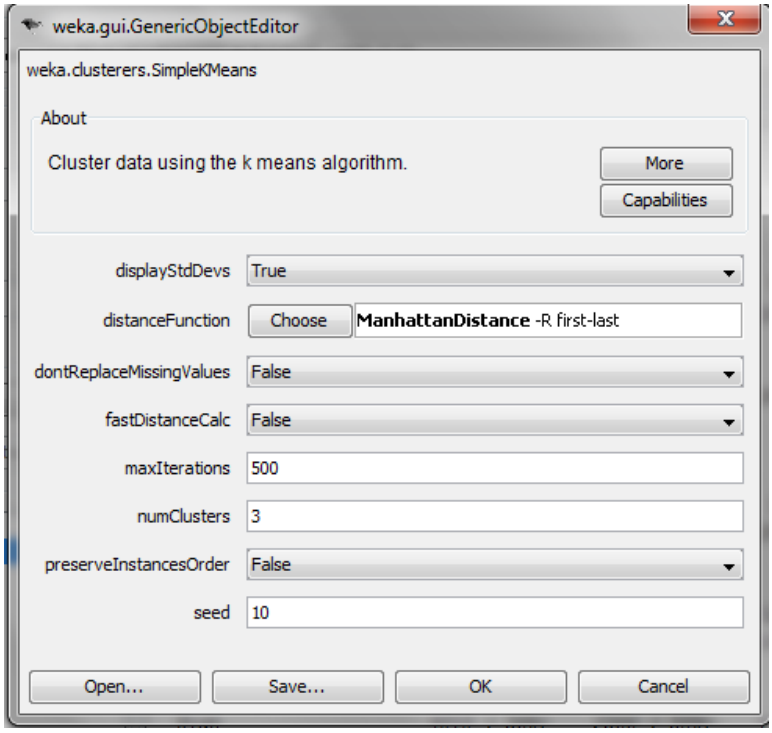


Figure 5. 1: Simple k-means Algorithm dialog box

Some of the functions displayed above in the dialog box are defined below:

- ✓ DisplayStdDevs: display standard deviations of numeric attributes and counts of nominal attributes with a True and False choices.
- ✓ DistanceFunction: This allows choosing the type of function supported by K-means you want to use like ManhattanDistance, EuclideanDistance.
- ✓ DontReplaceMissingValues: This is used to replace or not missing values globally with mean/mode with an available choices of True and False.
- ✓ NumClusters: This textbox is used to set the number of clusters (k in K-means) that we need to create and this value is needed to be input manually

into the system. This has to be defined by the number of segments that the business can successfully handle or control.

- ✓ Seed: The random number seed to be used. This defines the number of data tuples the cluster must start with.

Four clustering experiments have been conducted by changing the values of K as stated in section 5.2.1.

Experimentation 1

The researcher focused on identifying the dominant factors for being an Electronic Fund Transfer (EFT) customer before experimentation. Hence, discussing with the experts of the bank was the first task. Domain experts explain essential variables that are used to identify potential Electronic Fund Transfer (EFT) customer of the POS service. Consequently, the researcher focused on the points raised and discussed by the bank experts and acted accordingly.

According to Pritscher and Hans (2008) as well as based on the Dashen Bank S.C. experts suggestion, finding the clusters based on certain influential attribute is reasonable as there is no actual and concrete definition of a good segmentation output. Therefore, "ATM","Age_classification", "Monthly_Available_Balance", and "Occupation", were given a higher emphasis by the bank experts. Whereas, according to the GainRatioAttributeEvaluation attribute evaluator using Ranker search method from the WEKA tool, "ATM", "Sex", "Age_Classification", and "Monthly_Available_Balance" were the ranked and significantly relevant attributes, and thus, the domain experts have advised the researcher to take the attribute "Occupation" instead of attribute "Sex". Finally, the researcher agreed with the domain experts and replaced the "Sex" attribute by "Occupation" and this attribute was taken for segmentation purpose. Therefore, even though the remaining variables were not ignored totally, more attention was given on the specified variables on experimentation during analysis and interpretation of each and every cluster. The experts also put a clear opinion on the condition of being a potential customer of the Electronic Fund Transfer (EFT) of the POS service based on the above

stated attributes. They have stated a user of being a higher probability customer and of being a low probability customer as follows:

> If a customer is an ATM user, if the age classification of the customer is Age_Group_One and Age_Group_Two, if his/her Monthly_Available_Balance score is high and very high, and if the customer is working in his/her private business, then the customer is treated as a higher probability of being a potential EFT user. On the other hand if the above conditions are not satisfied, the customer is treated as having a low probability of the EFT of POS service customer.

Consequently, customers who satisfy the above criteria are treated as being a potential EFT user and those of who didn't satisfy the above stated condition are low value customers. Therefore, the interpretation and analysis of the different clusters is dependent on the points described above. Moreover, the researcher also consulted the experts for additional suggestions during interpretation of the clusters.

The final variables taken for the experimentation purposes after all pre-processing tasks are shown in the following table **(Table 5.1)**

Attribute Name	Data Type
Sex	Text
Occupation	Text
Saving Account	Text
Current Account	Text
Marital status	Text
Available Balance	Text
Age classification	Text
Year of Join	Text
ATM	Text
POS	Text

Table 5. 1: List of Attributes Taken For Experimentation

As clustering is unsupervised data mining technique, all the selected attributes were set as independent variable. Hence, the researcher and the bank experts reached consensus to exclude such variables and decided to trace the experimentation with the specified attributes. To see the pattern discovered the researcher used the dataset mean, maximum and minimum values using binning to divide values into intervals together with the judgment of the domain experts to determine the threshold values of Age_Classification and Available_Balance which is found at **Appendix 3.**

Some short forms have used for the attribute values for easy representation are shown below:

Values of variables	**Short Form**
Married	MA
Unmarried	UM
Age_Group_One	AGO
Age_Group_Two	AGT
Age_Group_Three	AGTH
Joined_2006_07	J/06/07
Joined_2008_09	J/08/09
Joined_2010_11	J/10/11
Hired	HR
Private	PR
High, Veryhigh, Low, Medium	H, VH, L, M respectively

The following figure (**Figure 5.2**) shows the training result of the clustering model including the number of attributes used, the number of instances used, the clustering

algorithm used, the test mode and other additional information. The WEKA 3-7-2 explorer window, as represented in **Appendix 4**, lists all attributes ready for the first cluster run. In addition, it is possible to know the minimum, maximum, mean, standard deviation, data type, number of missing values, number of distinct values, and currently selected attribute. One can see the values of all attributes by selecting one by one. These values in turn enable the researcher together with domain experts to determine the threshold values of each variable for the analysis of the result. The following table **(Table 5.2)** shows the result of the cluster run and the resulting segments. The algorithm is instructed to segment the dataset into six clusters. However, cluster6 has only 300 (3%) instances. The researcher assumed that there would be outliers in the training dataset as cluster6 contains very few records with the other clusters. The algorithm might have treated these outliers as records having different behaviour from other clusters and put them in a separate cluster.

Cluster Index	Freq. records	Sex	Occup ation	Saving Account	Current Account	Marital Status	Available Balance	Age classification	Year of Join	ATM	POS
1	2462	M	PR	YES	YES	MA	H	AGT	J/08/09	YES	NO
2	2181	F	PR	YES	NO	UM	H	AGO	J/08/09	NO	YES
3	2171	M	HR	NO	YES	UM	M	AGO	J/10/11	YES	YES
4	1787	M	HR	YES	YES	UM	VH	AGTH	J/10/11	NO	YES
5	2099	M	HR	YES	YES	MA	M	AGT	J/10/11	YES	NO
6	300	F	PR	YES	NO	UM	H	AGO	J/06/07	YES	NO

Table 5. 2: Summary result of the first cluster with k=6 and default seed (10)

Searching for the presence of outliers, it is found that Current_Account and Marital_Status contain some records that are out of the normal values. This is the reason why it is displayed in a separate cluster. The outliers are replaced with the Modal values of each field manually. In addition, it was found that cluster2 and cluster6 with similar

values almost in all variables and this shows that this is not a good cluster as it puts these two clusters separately.

Hence, after the outliers are replaced by the Modal values of the fields, the next task is to statistically identify important variables from the dataset and to use these attributes in model building process to improve the performance of the algorithm. As stated by Han and Kamber (2001), the attributes that appear at the top of the decision tree provide more information to classify the instances of the test dataset. Accordingly, in this study attributes are selected using the J48 decision tree algorithm and based on the condition stated by the domain experts and researcher to be a potential customer of the Electronic Fund Transfer (EFT) of POS service as stated at the beginning of this experimentation. The decision tree is built by selecting the cluster index as the dependent variable and the rest of the attributes as independent variables. Attributes that appear at the top of the decision tree and those attributes upon which many records are predicted into different class labels. The attributes that are considered statistically relevant and which are going to be used by the next experimentation are: Sex, Occupation, Saving_Account, Marital_Status, Available_Balance, Age_Classification, Year_Of_Join, ATM, and POS

Experimentation 2

After all the outliers are removed and important attributes are selected with the help of the J48 decision tree, as listed and described above, the second experiment with the same number of K (6) and default seed is conducted. **Table 5.4** summarizes the records distribution in segments of the second cluster run with k=6 and the default value of the seed. The specific values and percentage of composition for each attribute in each of the clusters are shown in **Figure 5.2.**

== Run information ===

Scheme: weka.clusterers.SimpleKMeans -V -M -N 6 -A "weka.core.EuclideanDistance -R first-last" -I 3 -S 10

Relation: bank-weka.filters.unsupervised.attribute.Remove-R4

Instances: 11000

Attributes: 9: Sex, Occupation, Saving_Account, Marital_Status, Available_Balance, Age_Classification, Year_Of_Join, ATM, POS

Test mode: evaluate on training data

Number of iterations: 3

Within cluster sum of squared errors: 22696.0

Figure 5. 2: First Cluster Run of the training dataset result with k=6

Cluster Distribution					
Cluster1	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**	**Cluster6**
2676 (24%)	1809 (16%)	1606 (15%)	2412 (22%)	1148 (10%)	1349 (12%)

Table 5. 3: Cluster distribution with k=6 and with the seed value= 1000

Different seed values have been tried in order to improve the data distribution of the segments but the default seed value (10) depicted at **Table 5.4** below gives a better data distribution in the segments. For example, where seed was set to 1000, cluster5 has only 1148(10%) instances as shown in **Table 5.3** above where as with the default seed value 10; its size is 1394(13%). The optimal seed value has been obtained after many experimentations and the default seed value has a better cluster distribution when compared with others.

Cluster distribution					
Cluster1	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**	**Cluster6**
2361(21%)	1852(17%)	2385(22%)	1690(15%)	1394(13%)	1318(12%)

Table 5. 4: Clustering distribution of the second experiment with k=6 and default seed =10

Different seed sizes are tested on each of this cluster formation to see whether the distribution of the segments could be improved. All the clusters with k=6, 5, and 4 and the default seed value has not showed a significant difference on the segments data distribution. The clustering model was built with the thresholds described by the marketing experts and by the WEKA's minimum, maximum, and mean values for each attribute as is illustrated in **Appendix 3.** This is used to determine what patterns are discovered for each subsequent cluster models with k (6, 5, and 4) and the best cluster seed. The detailed result of the second experimentation (Cluster run) is presented in **Table 5.5** below**.**

Cluster Index	Freq of records	Sex	Occupation	Saving Account	Marital Status	Monthly Available Balance	Age classification	Year of Join	ATM	POS
1	2361	79%M 20%F	4%HR 95%PR	77%YES 22%NO	83%MA 16%UM	36%H 25%L 3%M 34%VH	13%AGO 79%AGT 7%AGTH	28% J/10/11 20% J/06/07 51% J/08/09	76%YES 23%NO	36%Yes 63%NO
2	1852	7%M 92%F	20%HR 79%PR	98%YES 1%NO	52%MA 47%UM	60%H 21% L 17% M 0%VH	55%AG 35%AGT 8%AGTH	19% J/10/11 18% J/06/07 61% J/08/09	5%YES 94%NO	98%Yes 1%NO
3	2385	79%M 20%F	94%HR 5% PR	34%YES 65%NO	48%MA 51%UM	0%H 17% L 81% M 1%VH	74%AGO 23%AGT 1%AGTH	57% J/10/11 5% J/06/07 36% J/08/09	83%YES 16%NO	78%Yes 21%NO
4	1690	75%M 24%F	80%HR 19% PR	99%YES 0%NO	25%MA 74%UM	0%H 4% L 0% M 94%VH	1%AGO 36%AGT 61%AGTH	99% J/10/11 0% J/06/07 0% J/08/09	47%YES 52%NO	57%Yes 42%NO
5	1394	61%M 38%F	100%HR 0% PR	82%YES 17%NO	64%MA 35%UM	0%H 9% L 71% M 19%VH	28%AGO 59%AGT 12%AGTH	31% J/10/11 16% J/06/07 52% J/08/09	85%YES 14%NO	3%Yes 96%NO
6	1318	77%M 22%F	5%HR 94%PR	78%YES 21%NO	12%MA 87%UM	72%H 21% L 0% M 6%VH	82%AGO 14% AGT 3%AGTH	47% J/10/11 23% J/06/07 29% J/08/09	93%YES 6%NO	63%Yes 36%NO

Table 5. 5: Clustering result of the second experiment with k=6 and default seed =10

The above table (**Table 5.5**) shows the discrete values of each attribute. After the values of each attribute in each cluster have been generated, a description for each segment of cluster has been done as presented in **Table 5.6.** The description is determined based on EFT marketing strategies of customers in different segments.

Cluster Index	**Description based on the result**	**Rank**
1	The Sex that dominate this cluster is male (79%), A great number of customers (95%) run their private business while the remaining customers are hired by others; Most of the customers have an account that is used for their own purpose not for business purpose, Most of them are married. They have 25%, 36%, 34% for low, high and very high attribute values for monthly available balance. Most of them are at the age of greater than 30 and less than 50 (AGT), Mostly Joined in 2008/09, 76% of them are ATM users, and 63% of them are not POS users	2
2	High percentage of Females, 79% of the customers are working in their private business, Almost all (98%) of the customers have an account that is used for their own purpose not for business purpose, More than half of them are married, They have 60%, 17%, 0% for High, Medium and very high attribute values for monthly available balance, Most of them are at the age of Age_Group_One (55%) and Age_Group_Two (35%) , Mostly Joined in 2008/09, 94% of them are not ATM users, and 98% of them are POS users	4
3	This cluster is the highest cluster that contains 2385(22%) instances of the total customers. In this cluster there are high male users, High hired customers, Low percentage of saving account users, most of them are unmarried,81% of them have a medium monthly available balance, 74%AGO and 23%AGT is their age classification, Most of them are	5

	recent card customers, 83% , 78% are ATM and POS customers of the bank	
4	High male users, High hired customers, All customers are saving account users, most of them are unmarried, 94% of them have a very high monthly available balance, 61%AGTH and 36%AGT is their age classification, All of them are recent card customers, 47% , 57% are ATM and POS customers of the bank respectively	6
5	High male users, All are hired customers, most of them have saving account users, most of them are married, 19% of them have a very high monthly available balance and 71% have Medium monthly deposit, 28%AGO and 59%AGT is their age classification, Most of them joined 2008/09, 85% are ATM users but 96% are not POS service users.	3
6	This cluster is the smallest cluster that contains (12%) instances of the total customers. It contains high male users, 94% of them have private job, Majority customers are saving account users, most of them are unmarried, 72% of them have a high and 6% very high monthly available balance, 82%AGO and 14%AGT is their age classification, Most of them are recent card customers, 93%, 63% are ATM and POS customers of the bank respectively.	1

Table 5. 6: Summary of Cluster for k=6 and default seed (10) and the possible rank of cluster

As it is described in the first experimentation of this research, potential Electronic Fund Transfer (EFT) of POS service customers' are those who can satisfy the following criteria:

i. Customers who use ATM in the company and whose monthly available balance is high or very high as well as if these customers are working in their private business, then they are treated as high EFT customers.

ii. Those customers with lower age group (Age_Group_One and Age_Group_Two) are also considered potential EFT customers of the company next to the above or formerly described customers.

Each cluster has been assigned rank based on the above mentioned facts of the business. As it is represented in **Table 5.6**, the sixth cluster (Cluster6) is ranked first. This is due to the reason that, customers in this cluster use ATM service, their monthly available balance is grouped under high and very high, most customers have their own private job, most of them are grouped in a lower age (AGO and AGT) and most of them have a saving account, of the company and these are the potential customers of EFT of POS service. Cluster6 also includes high percentage male customers and high percentage POS service customers whose marital status is single in majority.

Besides, the first cluster ranked second as customers in this segment contain high number of ATM users and medium numbers of customers are grouped under Age_Group_Two (AGT) and saving account. In addition to this there are large percentages of high and very high numbers of customers who deposit a monthly available balance and most of them are engaged in their own business. This cluster also consist high number of male and married customers with least POS service usage.

Cluster5 contains high percentage of AGO and AGT and most of them have large number of saving account users. Similarly customers in this group generate very high number of ATM potential customers. On top of this, Majority of customers in this cluster also contains high and medium monthly available balance. And therefore, this cluster is ranked third. The second cluster is also ranked fourth in which it contains very low ATM customers. But it contains very large percentage of monthly available balance and Lower age. There are very large numbers of saving account customers and most of them are engaged in their personal business. The third and fourth clusters are ranked fifth and sixth respectively and are considered low level customers as can be seen from the generated output in **Table 5.6** above**.** In the third cluster medium level monthly available balance and the Age_Classification is under AGO and AGT group and therefore ranked fifth.

Whereas cluster4 contains least value customers because it consists of low ATM customers and most customers are employed in public or private organizations.

The experimentation with k=6 and default seed value (10) appears to have created dissimilar segments of clusters (i.e. seems to have a good cluster distribution). Because it clearly identified the low, medium and high level customers in the different clusters generated above. Even though this experimentation seemed better, another experimentation have been conducted with k=5 and default seed value (10) in order to compare the segmentation models.

Experimentation 3

This experimentation has been run with value of k=5 (number of clusters) and default seed value=10. The generated result of the WEKA clustering is shown below in **Table 5.7.**

Cluster Index	Freq. records	Sex	Occupation	Saving Account	Marital Status	Available Balance	Age classification	Year of Join	ATM	POS
1	1650	76%M	7%HR 82%PR	74%YES 25%NO	71%UM	65%H 17%L 12%M 3%VH	64%AGO 31%AGT 4%AGTH	46% J/08/09	89%YES 10%NO	26%Yes 73%NO
2	1899	87%F	14%HR 85%PR	96%YES 3%NO	55%MA	66%H 22% L 10% M 0%VH	61%AGO 30% AGT 8%AGTH	59% J/08/09	11%YES 88%NO	98%Yes 1%NO
3	2167	80%M	91%HR 8% PR	36%YES 63%NO	56%MA	2%H 15% L 80% M 1%VH	67%AGO 30% AGT 1%AGTH	64% J/10/11	85%YES 14%NO	85%Yes 14%NO
4	1587	75%M	77%HR 22%PR	99%YES 0%NO	73%UM	0%H 9% L 0% M 90%VH	1%AGO 39% AGT 58%AGTH	99% J/10/11	44%YES 55%NO	59%Yes 40%NO
5	1733	65%M	49%HR 50%PR	82%YES 17%NO	82%MA	3%H 17% L 58% M 20%VH	11%AGO 76% AGT 11%AGT	48% J/08/09	78%YES 21%NO	10%Yes 89%NO

Table 5. 7: Clustering result of the third experiment with k=5 and default seed =10

In this experimentation, five clusters are created. The result and interpretation of this clustering run with the above experimentation is presented below in **Table 5.8** on the facts stated above. Here, cluster1 is ranked first based on the facts described by the domain experts and the researcher. This cluster (Cluster1) segment generate high number of ATM users, they have a total percentage of 80% of high and medium monthly available balance and most of them are engaged in their private business. This segment also contains customers who are not married; most customers are males and use their account for their personal use, and 73% of them are not users of the POS service. Customers in the fifth cluster also ranked second from the fact that they are highly ATM users and 81% of them are high, middle and very high level customers according to their level of monthly available balance. Male and private customers are the dominating customers in this segment. Besides, most customers in this cluster used saving account as an account type and they are not often customers of the POS service.

Cluster Index	Description based on generated result	Rank
1	There are high male customers, most of them are engaged in their private job, There are high number of customers who use saving account, large number of single customers, High(65%), Medium (12%) and 3% very high monthly available balance, High AGO (64%), 31% AGT, high (46%) is their age classification, and they have joined in 2008/09, 89% of them are ATM users and majority 73% of them are POS customers.	1
2	Very high Female customers, large number of users are committed in their private work, large number of saving account, majority of them are married, High(66%), and Medium (10%) monthly available balance, the percentage of age classification is of AGO (61%), 30% AGT. 59% customers joined in 2008/09, 11% ATM users and 98% Pos customers	5
3	This cluster contains greater number of customers than any other cluster (19.7%). This also contains large number of males, high number of hired customers, less number of saving account, large number of married customers, High(2%), Medium (80%) and 1% very high monthly available balance, Age is AGO (67%), 30% AGT, 64% joined in 2010/11, 83% ATM users and 85% POS customers	3
4	This cluster contains few numbers of instances. Very high male, high number of hired customers, all customers have saving account, large number of single customers, Very high(90%) monthly available balance, 58% AGTH of age, 99% joined in 2010/11, 55% ATM users and 59% POS customers	4
5	65% males, high number of private customers, high number of saving account, large number of married customers, High(3%), Medium (58%) and 20% very high monthly available balance, AGO (11%) and 76% AGT age groups, 48% joined in 2008/09, 78% ATM users and 89% POS customers	2

Table 5. 8: Summary of Cluster for k=5 and default seed (10) and the possible rank of cluster

Customers in the third cluster are medium level customers as they generate medium monthly deposit amount, medium number of customers are grouped at Age_Group_One (AGO) and most of them are ATM customers. Though most of them are not working in their personal private business, this cluster is ranked third when compared with cluster4 and cluster2. These two clusters contain low level customers. But cluster4 is ranked fourth when compared with the later one as it consists customers whose monthly balance is very high (90%) and all of them used their account type for their personal use. In this cluster segment the customers of the card system are increasing from time to time. Where as in cluster2, the monthly balance of the customers is low and there are very few ATM customers when compared with cluster4. As can be seen from the above table, this clustering run also creates dissimilar clusters but when compared with the previous cluster run with k=6 , the clustering run with k=6 clearly differentiate between the high value and low value customers better than when k=5.

Experimentation 4

The researcher has conducted another experimentation with a cluster run of k=4 and default seed size (10). The following table (**Table 5.9**) generates the summarized result of the fourth experimentation and the detailed description of this result is depicted in **Table 5.10.**

Cluster Index	Freq. records	Sex	Occupation	Saving Account	Marital Status	Available Balance	Age classification	Year of Join	ATM	POS
1	3824	75%M	27%HR 72%PR	68%YES 31%NO	67%UM	57%H 18%L 24%M 0%VH	65%AGO 31%AGT 3%AGTH	57% J/08/09	85%YES 14%NO	24%Yes 75%NO
2	2299	78%F	16%HR 83%PR	93%YES 6%NO	50%UM	58%H 22% L 12% M 7%VH	57%AGO 34% AGT 7%AGTH	56% J/08/09	9%YES 90%NO	97%Yes 2%NO
3	3435	72%M	84%HR 15%PR	67%YES 32%NO	76%MA	4%H 15% L 59% M 19%VH	32%AGO 61% AGT 5%AGTH	68% J/10/11	77%YES 22%NO	64%Yes 35%NO
4	1442	76%M	89%HR 10%PR	93%YES 6%NO	88%UM	2%H 8% L 6% M 82%VH	5%AGO 19% AGT 74%AGTH	93% J/10/11	75%YES 24%NO	29%Yes 70%NO

Table 5. 9: Clustering result of the fourth experiment with k=4 and default seed =10

Cluster1 is also ranked first as it has large number of ATM customers, and customers of this cluster segment have a monthly available balance of 81%. In addition to this, majority of them are engaged in their personal job and used their account for their personal use Most of the customers are male and single. Cluster3 is also ranked second when compared with the remaining cluster segments other than cluster1 based on the facts defined at the first experimentation by the domain experts. This cluster segments consists of high number of ATM customers, 82% of them have high, middle, and very high monthly deposit when their available balance is checked by the marketing department of the bank. On top of this, their age classification is mostly grouped in AGO and AGT.

Cluster Index	Description based on generated result	Rank
1	This cluster segment contains large number of instances. Large male customers, high number of private customers, high number of saving account, large number of single customers, High(57%), Medium (24%) and 0% very high monthly available balance, age is in group AGO (65%) and 31% AGT, high (57%) joined in 2008/09, 85% ATM users and 75% are not POS customers	1
2	Very high Female, large number of private customers, larger number of saving account, equal number of single and married customers, High(58%), and Medium (12%) and 7% very high monthly available balance, large percentage AGO (57%), 34% AGT, 56% joined in 2008/09, 9% ATM users and 97% POS customers	4
3	Large number of males, high number of hired customers, large number of saving account, large number of married customers, High(4%), Medium (59%) and 19% very high monthly available balance, Age is AGO (32%), 61% AGT, 68% joined in 2010/11, 77% ATM users and 64% POS customers	2
4	This cluster has very less number of instances when compared with other clusters. Very high male, high number of hired customers, most customers have saving account, large number of single customers, Very high(82%) monthly available balance, 73% AGTH of age, 93% joined in 2010/11, 75% ATM users and 29% are POS customers	3

Table 5. 10: Summary of Cluster for k=4 and default seed value=10 and the possible rank of cluster

Cluster4 and cluster2 are the least potential EFT customers of POS service. They differ in that; cluster4 contains higher number of ATM users, and the monthly balance is relatively high when compared with cluster2, as well as the card holder customers are increasing from time to time in this cluster segment for that reason it is ranked third. While cluster2

consists of very low number of ATM users and low level monthly balance when compared with other clusters. Generally, the cluster segments in this cluster run does not show customers with high interest of being EFT users of the POS service like the cluster model of k=6.

After the output of the research is generated, the researcher observed that the variable "Sex" and saving account plays a great role in identifying potential EFT of POS service. In this, if the customers' sex is "Male", and if the customers have an account for their personal use, then the probability of being a potential customer is high. From this result, the experts critically observed and concluded that the bank should not be limited to the variables that were identified as best attributes for obtaining best customers' which is defined at experimentation 1.

5.2.4: Selecting the best clustering model

In the above, four different clusters model experimentations is done in order to find appropriate segmentation model. The researchers have taken different experimentation with k values of 6, 5 and 4 with a default seed value (10). At experimentation 1 with k=6, some few outliers have been found in Current_Account and Marital_Status fields and they replaced by the Modal values manually for both variables.

The second experimentation with the same value of k and seed, but after replacing the outliers is done and dissimilar clusters or improved cluster distribution are generated ; all the created clusters were with different behaviours and such clustering is appropriate.

Lastly, the best cluster model with best cluster distribution has been evaluated based on:

1. Number of iterations the algorithm uses (This shows the algorithm has reposition and all misplaced data items in their correct classes within a few looping. And the minimum value shows k-means algorithm has converged very soon).
2. Within cluster sum of squared errors (This is the measure of the goodness of the clustering and tells how tight the clustering is overall, that means lower values of squared errors are better) , and

3. The judgment of the domain experts based on the facts stated in the first experiment.

Therefore, based on these criteria's the three cluster models (i.e. at k=6, 5, 4 values) are compared.

- Cluster model at k=6, consists of: Number of iterations= 3 and within cluster sum of squared errors= 22696.0.
- Cluster model at k=5 consists of: Number of iterations= 4 and within cluster sum of squared errors= 27046.0, and
- Cluster model at k=4 consists of: Number of iterations= 4 and within cluster sum of squared errors= 30295.0.

Consequently, as stated above, cluster model at k=6 shows the least value within cluster sum of squared errors and in number of iterations than cluster model at k=5 and 4. Hence due to the opinion of domain experts and least value of within cluster squared errors as well as number of iterations, the cluster model formed at k=6 is selected as the best model which creates dissimilar segments of EFT of POS service customers of the bank and this cluster model is used as an input for the decision tree and ANNs creation.

5.2.5. Classification Modeling

The main purpose of the clustering model is to identify segments of customers that share high similarity within class. J48 decision tree classification algorithm and Multilayerperceptron ANN is further used to compare its performance against other methods. The decision tree and ANN model is tested by passing different parameters and the result of each classification accuracy is reported and its performance is also compared in classifying new instances of records.

Before focusing on the classification models building, it is important to discuss about the default parameter for the model training and testing which is 10-fold cross validation. In 10-fold cross validation, the initial data are randomly partitioned into 10 mutually exclusive subsets or folds, 1,2,3,4 ...10, each approximately equal size. The training and testing is performed 10 times. In the first iteration, the first fold is reserved as a test set,

and the remaining 9 folds are collectively used to train the classifier. The classifier of the second iteration is trained on folds 1, 3, 4... 10 and tested on the 2nd fold etc. The accuracy estimate is the overall number of correct classifications from the 10 iterations divided by the total number of samples in the initial dataset (Han and Kamber, 2001).

5.2.5.1. Decision Trees Model Building

As a result of extensive data processing, 11000 customer records are chosen for the purpose of classification. Nominal value of" YES" and "NO" is used to classify the Point Of Sale usage by Dashen Bank S.C. customers.

In the experiment, 11,000 datasets having 10 attributes, 9 of them are independent variables and the tenth one is attribute for POS or EFT usage being dependent (class label attribute) are fed to the WEKA's explorer. Two classification algorithms were used to cross check the accuracy of the classification technique used. These are J48 tree algorithm classifier and NaiveBayes classifier. On both classifiers ten trials are made at different percentage splits. The following table shows the classification accuracy of these algorithms on each of the trials made.

Percentage split of the training data set	**NaiveBayes**		**J48**	
	Correctly classified	**incorrectly classified**	**Correctly classified**	**Incorrectly classified**
50 %	66.9818 %	33.0182 %	88.0909 %	11.9091 %
60%	67.7045 %	32.2955 %	88.3182 %	11.6818 %
65%	67.4805 %	32.5195 %	88.6753 %	11.3247 %
66%	67.3529 %	32.6471 %	88.7968 %	11.2032 %
70%	68.2121 %	31.7879 %	88.8182 %	11.1818 %
75%	68.0364 %	31.9636 %	88.4364 %	11.5636 %
79%	67.619 %	32.381 %	88.0519 %	11.9481 %
80%	67.6818 %	32.3182 %	88.5455 %	11.4545 %
82%	67.7778 %	32.2222 %	88.7879 %	11.2121 %
85%	69.1515 %	30.8485 %	89.0303 %	10.9697 %

Table 5. 11: Accuracy of NaiveBayes and J48

According to the result showed on the above table, the maximum accuracy difference between NaiveBayes and J48 is 21.4439% at percentage split of 66% and the minimum accuracy difference is 19.8788% recorded at percentage split of 85%. The J48 classification algorithms have higher accuracy than the NaiveBayes classification algorithm as displayed in **Table 5.11.**

In addition to this, to select the best classification algorithm (i.e. from J48 and NaiveBayes), an experimenter from the WEKA 3-7-2 tool has also been used to check their accuracy. The WEKA 3-7-2 tool experimenter first allows you to choose the algorithms you want to compare, and then it checks whether you didn't have any error or not, if it is successful, it shows the following run experimenter dialog box.

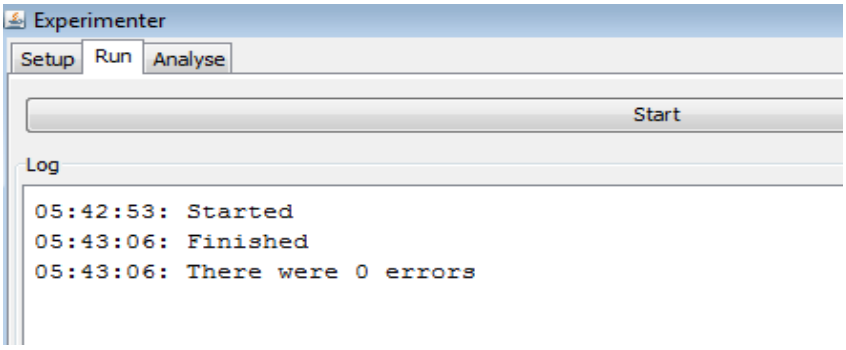


Figure 5. 3: The Run Experimenter Dialog Box

After a successful run of the experimenter, then it starts analyzing the accuracy of the selected algorithms and as shown in the following figure:

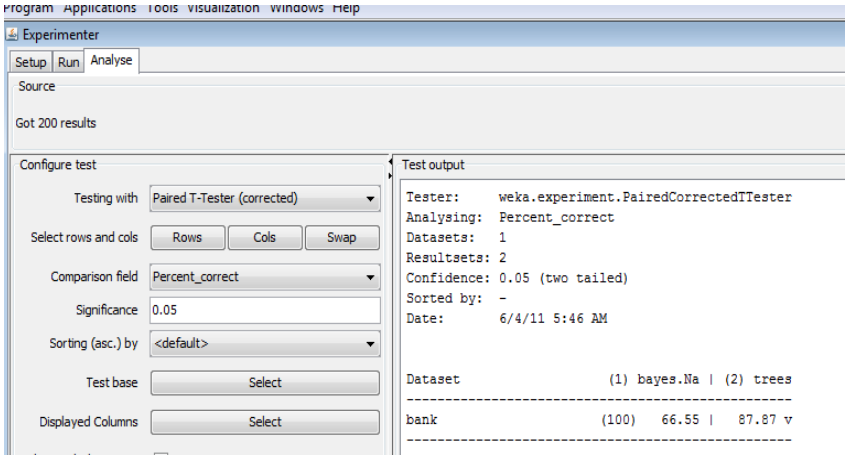


Figure 5. 4: The Analyze Experimenter Dialog Box to Analyze the Classifiers

As it is displayed in the above figure, the accuracy of the J48 and NaiveBayes is 87.87% and 65.55% respectively, and thus the J48 classifier has a better accuracy level than the NaiveBayes.

Therefore, due to the accuracy of results as well as easy result interpretation and tree visualization facilities, the J48 classifier was selected for this study. All variables, except the variable which has been removed at experimenation2 of the clustering model are used as independent variables and the cluster labels, which are assigned by the clustering, as dependent variable for the algorithm. The decision tree model algorithm which is the J48 contains some parameters (initially created with their default values) that can be changed manually can improve the classification accuracy. Some of these include:

- confidenceFactor parameter : is used for pruning (i.e. smaller values gain more pruning) and its default value is 0.25,
- minNumObj: which is the minimum number of instances per leaf (i.e. larger values minimizes the leaf) and its default value is 2, and unpruned with default value false parameter is used to check whether pruning is performed or not.

Figure 5.5. Shows the result of the first decision tree with J48 algorithm and default values of the parameters. The result contains 218 sizes of tree and 134 leaves. The confusion matrix generated from the decision tree using J48 algorithm shown below and its summarized result is depicted in **Table 5.12.**

```
=== Confusion Matrix ===

    a    b    c    d    e    f   <-- classified as
 2358    0    1    0    2    0 |    a = cluster1
    4 1841    2    2    3    0 |    b = cluster2
    4    4 2373    0    4    0 |    c = cluster3
    1    5    4 1677    3    0 |    d = cluster4
    1    2    1    1 1389    0 |    e = cluster5
    4    0    1    0    1 1312 |    f = cluster6
```

Figure 5. 5: Result of the decision tree J48 algorithm with default parameter values

As it is displayed in the confusion matrix **Table 5.12** below, the accuracy is 99.55%, which means from the total 11000 number of instances used, 10950 (99.55%) records are classified correctly while the remaining 50 (0.45%) records are classified incorrectly. The results of the misclassification might be created by the results of the outliers replaced by the Modal values of both Current_Account and Marital_Status attributes in the clustering model building.

Actual	**Predicted**						**Total**	**Accuracy rate**
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6		
Cluster1	2358	0	1	0	2	0	2361	99.87%
Cluster2	4	1841	2	2	3	0	1852	99.40%
Cluster3	4	4	2373	0	4	0	2385	99.49%
Cluster4	1	5	4	1677	3	0	1690	99.23%
Cluster5	1	2	1	1	1389	0	1394	99.64%
Cluster6	4	0	1	0	1	1312	1318	99.54%
Total	**2372**	**1852**	**2382**	**1680**	**1402**	**1312**	**11000**	**99.55%**

Table 5. 12: Result from the J48 decision tree learner with default parameter values

Furthermore, the results of the classification experiment using the decision tree J48 algorithm has shown 99.87 %(for cluster1), 99.40 %(for cluster2), 99.49% (for cluster3), 99.23% (for cluster4), 99.64% (for cluster5), and 99.54% (for cluster6) of the records are correctly classified. In addition to this, out of the total 2361 customers who are described in cluster1 of **Table 5.4** and **Table 5.5**, 2358 (99.87%) are correctly classified in its appropriate cluster (cluster1). Whereas about 3(0.13%) are misclassified in cluster3 and cluster5. Another is the high level potential EFT of the POS service customers (cluster6), who are described in **Table 5.4** and **Table 5.5 ,** out of the 1318**,** 1312 (99.54%) of them are correctly classified in their appropriate cluster (cluster6).

From this, about 6 (0.46%) of them are wrongly classified in cluster1 (4), cluster3 (1), and cluster5 (1). Here, there are no least EFT customer (cluster4), who belongs to cluster4 (least customers), which are wrongly classified into high level EFT customers segment which are identified with cluster6. However, out of the total 1690 of the least level customers (cluster4), about 1677 (99.23%) of them are correctly classified, while 13 (0.77%) of them are misclassified into cluster1 (1), cluster2 (5), cluster3 (4) and cluster5 (3). The decision tree J48 algorithm generated from this model is shown in **Appendix 2.**

Although, this decision tree model gives a good accuracy, it has displayed very lengthy tree to generate all relevant rules (i.e. with 134 Number of Leaves and 218 size of the tree). To alleviate this, some of the J48 parameter values to get a minimized tree size and number of leaves. As a result, the value of the minNumObj (the minimum number of instance per leaf) was traced with the values default value (2), 5, 10, 15, 20, 25, and 30. After these trials have took the above stated values, better and minimized tree size (78 numbers of leaves and 126 nodes) is found with relatively better classification accuracy (98.56%) at the value 15 and its result is represented at **Table 5.13.** Even though there are minimized number of leaves at value 20, 25 and 30, the classification accuracy becomes less and for that reason the value 15 is somewhat better in the tree length and accuracy. The accuracy level is less at value 15 when compared with the default value because the default has an accuracy of 99.55% and at the value 15; there is accuracy level of 98.56%.

The following generated result with a changed value of minNumObj to 15 has the outputs of the classification experiment using the J48 decision tree algorithm for each cluster as depicted in **Table 5.13** is 99.40%, 97.46%, 98.74%, 98.40%, 98.56%, and 98.48% of the records are correctly classified for cluster1, cluster2, cluster3, cluster4, cluster5, and cluster6 respectively. Moreover, an overall classification accuracy of 98.56%is resulted and this is reduced by 0.99% when compared with the default value of minNumObj displayed in the above table (**Table 5.12**). Besides this, out of the total 1318 high level EFT customers in cluster6, 1298 (98.48%) of them are correctly classified, which is 1312 (99.54%) in the previous model with default parameter values. As a result, its accuracy level has reduced by 1.06%. However, it has reduced the tree size into 78 numbers of leaves and 126 nodes.

Actual	Predicted						Total	Accuracy rate
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6		
Cluster1	2347	4	8	0	2	0	2361	99.40%
Cluster2	23	1805	8	9	7	0	1852	97.46%
Cluster3	4	3	2355	11	12	0	2385	98.74%
Cluster4	1	13	8	1663	5	0	1690	98.40%
Cluster5	1	5	1	9	1374	4	1394	98.56%
Cluster6	17	0	3	0	0	1298	1318	98.48%
Total	**2393**	**1830**	**2383**	**1692**	**1400**	**1302**	**11000**	**98.56%**

Table 5. 13: Result from the J48 decision tree learner with minNumObj=15

It is now simple to conclude that classification model built with the default parameter values have a better classification accuracy result from the above detailed trials, though it has a long tree through the generated useful rules. Another classification model experimentation has been done with default parameters of the J48 decision tree learner algorithm. A model was built with 70% (7700) records for training and 30% (3300) records for testing of the total 11000 records. Out of 3300 (30%) the dataset used for testing, this algorithm has classified 99.45 % (3282) correctly, while the remaining 0.55 % (18) are incorrectly classified testing dataset. Again, another classification model similar to the above experimentation has been conducted by taking the same training and testing dataset but by changing the default parameter values of the minNumObj value from 2 (default value) to 10 and 20. Out of the 3300(30% of the whole dataset) testing dataset, 98.70% of them are correctly classified with a minNumObj value of 10, and as well as 97.49% (3217) with minNumObj value of 20 are correctly classified. That is 2.5152 % (83) the test set is wrongly classified but the size of the tree has minimized to 63 number of leaves and 105 size of the tree.

In conclusion, the classification model created with default parameter values and 10-fold cross validation has shown better classification accuracy than splitting into 70% and 30% of the dataset for training and testing of the model respectively. Therefore, from the different decision tree models created in the previous experimentations, the first model described in **Table 5.12,** with the default parameter values and 10-fold cross validation has been selected due to best individual and overall cluster classification accuracy. The J48 decision tree is displayed in **Appendix 5** and rules generated from this model are shown in **Appendix 6.**

5.2.5.2. Artificial Neural Network (ANN) Classification Model

A different J48 decision tree model experiment is conducted and the best model that has shown better overall classification accuracy is selected. The data is first normalized to the range [-1, 1] which is suitable for the neural network algorithm. ANN model learns very fast if the attribute values are normalized to the range [-1, 1]. The ANN model is built for categorical label classes, which are derived from the application clustering, prediction of unknown dataset. In order to normalize the values of the attributes, the researcher used WEKA's pre-processing facilities so that all the variables fall in the range [-1, 1].

Most of the values of the variables of the dataset of this research were categorical or nominal. Thus the values should be changed into numeric values for normalization. Hence the distinct values of the categorical attributes have been assigned numeric values as shown in the following **Table 5.14.**

Attributes	Represented as
Sex	M =1 and F=2
Occupation	Hired=1 and Private=2
Saving Account	YES=1 and NO=2
Current Account	YES=1 and NO=2
Marital Status	Married=1 and Unmarried=2

Monthly Available Balance	Low=1, Medium=2, High=3, and Very high=4
Age classification	Age_Group_One=1, Age_Group_Two=2, and Age_Group_Three=3
Year of Join	Joined_06_07=1, Joined_08_09=2, Joined_10_11=3
ATM	YES=1 and NO=2
POS	YES=1 and NO=2

Table 5. 14: Representing the nominal values of the attributes by numeric values

After mapping these nominal values into the above stated numerical values, the WEKA's explorer normalize pre-processing facility has been used to normalize all values to fall in the range [-1, 1]. The attributes used in the decision tree model are used in the neural network Multilayerperceptron algorithm as well. After all necessary pre-processing is done; the neural network model experimentation is tested by changing the values of the hidden layer and learning rate parameters with a 10-fold cross validation as shown below:

- hiddenLayers: is one of the Multilayerperceptron parameter and it is the hidden layers of the neural network; and this is a list of positive whole numbers. The default value of this parameter is 'a' which stores = (Number of attributes + classes)/2, which is in this case=> (9+6)/2= 8.This default value has been used and changed to see the ANN classification accuracy
- learningRate: is another Multilayerperceptron algorithm and is defined as the amount the weights are updated and its default value is 0.3.

The best classification accuracy on the 10-fold cross validation neural network classification is carried out by changing its parameter values. The result of the neural network model with default parameter values is shown in **Table 5.15.**

Records count in each cluster	Clusters						Total
	Cluster1	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**	**Cluster6**	
Records in the cluster (A)	1292	2142	1604	1192	2854	1916	11000
Number of records correctly classified (B)	1292	2142	1604	1191	2854	1914	10997
Number of records wrongly classified(C=(A-B))	0	0	0	1	0	2	3
Correctly classified in %(accuracy)(D=(B/A)*100)	**100%**	**100%**	**100%**	**99.92%**	**100%**	**99.90%**	**99.97 %**
Wrongly classified instances in % (E=(C/A)*100)	**0%**	**0%**	**0%**	**0.08%**	**0%**	**0.1%**	**0.03%**

Table 5. 15: ANNs Multilayerperceptron algorithm classification model with default learning rate =0.3 and hidden layer =8

The model showed in **Table 5.15** is built with a 0.3 and 8 default learning rate and hidden layer values respectively of the neural network parameters and 10-fold cross validation. This model has generated 99.97% of overall classification accuracy. That means, from the total 11000 datasets given, 10997 of them are correctly classified. But the remaining 0.03% (3) very small instances are incorrectly classified. From the total 1916 potential customers showed above in **Table 5.15**, the model has correctly classified 1914 (99.90%) instances in their correct cluster (i.e. cluster6), and about 2 (0.1%) of them are wrongly classified to other cluster labels.

As it is stated in **Table 5.6** above, customers who are grouped in cluster4 have low probability of being a potential customers' of EFT of POS service. Thus from these 1192 low level customers of POS service in this cluster, 1191 (99.92%) of them are correctly classified in their correct class, but 1 (0.08%) customer is wrongly classified in other cluster. Generally, the high value customers, which are in cluster6 has equivalent errors with cluster4, which contains low probabilistic POS customers.

The neural network model has also been traced with different parameter values. For example, **Table 5.16** shows experiment with learning rate 0.4 and hidden layer=8.

Records count in each cluster	**Clusters**						**Total**
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	
Records in the cluster (A)	1292	2142	1604	1192	2854	1916	11000
Number of records correctly classified (B)	1292	2142	1604	1191	2854	1914	10997
Number of records wrongly classified(C=(A-B))	0	0	0	1	0	2	3
Correctly classified in %(accuracy)(D=(B/A)*100)	**100%**	**100%**	**100%**	**99.92%**	**100%**	**99.90%**	**99.97 %**
Wrongly classified instances in % (E=(C/A)*100)	**0%**	**0%**	**0%**	**0.08%**	**0%**	**0.1%**	**0.03%**

Table 5. 16: ANNs Multilayerperceptron algorithm classification model with learning rate (0.4) and hidden layer (8)

The above result is achieved by changing the default values of the learning rate and hidden layer to 0.4 and 8 respectively. This experiment returned exactly the same output with the default parameter values. Furthermore, additional neural network values are tested with 0.5 and 8; 0.8 and 8; 0.3 and 9; 0.3 and 10, learning rate and hidden layer respectively also gave the same result with the default neural network Multilayerperceptron algorithm parameters except the one with 0.3 learning rate and 10 hidden layers.

When the learning rate is 0.3 with 10 hidden layer parameter the neural network model has less classification accuracy than the rest as depicted below in **Table 5.17.**

Records count in each cluster	Clusters						Total
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	
Records in the cluster (A)	1292	2142	1604	1192	2854	1916	11000
Number of records correctly classified (B)	1292	2141	1604	1191	2854	1914	10996
Number of records wrongly classified(C=(A-B))	0	1	0	1	0	2	4
Correctly classified in %(accuracy) (D=(B/A)*100)	**100%**	**99.95%**	**100%**	**99.92%**	**100%**	**99.90%**	**99.96%**
Wrongly classified instances in % (E=(C/A)*100)	**0%**	**0.05%**	**0%**	**0.08%**	**0%**	**0.1%**	**0.04%**

Table 5. 17: ANNs Multilayerperceptron algorithm classification model with learning rate (0.3) and hidden layer (10)

Table 5.17 shows that, the overall classification accuracy is 99.96%, which has very close accuracy with the above values. The only difference created in this is in cluster2, where from a total of 2142 instances of cluster2, about 2141 (99.95%) instances are correctly classified in their respective clusters, while the rest 1 (0.05%) of the them are wrongly classified. However, this model has the same accuracy in classifying high probabilistic and low probabilistic POS customers i.e. cluster6 and cluster4 respectively.

In general, all the neural network models stated above have encouraging and better classification accuracy in overall and individual cluster accuracies. Moreover, the neural network model built with a default parameter values and 10-fold cross-validation is considered for comparison with the decision tree model.

5.2.6. Decision Tree and Neural Network Models Comparison

One of the basic targets of data mining is to compare different models and to select the better classification accuracy accordingly. Therefore, detailed experimentation for different models has been conducted. Accordingly, the best classification algorithm which is appropriate for this problem domain has been selected.

The researchers in collaboration with the domain experts have formulated three criteria's to select the best model from these two. These are: the overall classification accuracy,

and the model accuracy in classifying both low and high probabilistic POS customers. The experts explained that more attention should be given to high level customers because high profit is gained from them. Consequently, the model which generates a minimum classification error in correctly assigning high value customers (i.e. cluster6) in their cluster is their choice. They also suggested that, the bank spends large amount of money in creating and controlling high level customers, it is better if the low level customers are not classified into high value customers

The results of the overall classification, low level customers, and high level customers for both classification models is described in detailed below:

- The 10-fold cross validation is used to compare these two models where default parameter is taken for both cases.
    - The decision tree model has an overall classification accuracy of 10951 (99.55%) which are correctly classified and 49 (0.45%) that are wrongly classified. Again 1312 (99.54%) of the high probabilistic customers (cluster6) are correctly classified, and 6(0.46%) are misclassified. Besides this, 1677 (99.23%) are correctly classified and 13(0.77%) are wrongly classified low probabilistic customers (cluster4).
    - On the other hand, the neural network model has an overall classification accuracy of 10997 (99.97%), and about 3 (0.03%) are wrongly classified. About 1914 (99.90%) high level customers (cluster6) are correctly classified, while 2 (0.1%) of them are misclassified. But in the low level customers (cluster4), 1191 (99.92%) are correctly classified and 1 (0.08%) is wrongly classified.

From the above results, the overall performance level of the decision tree model is 99.55%, which is less than the accuracy of the neural network (99.97%). The accuracy of the neural network is better and exceeds by 0.42% in the overall accuracy. Only 49 (0.45%) and 3 (0.03%) are wrongly classified in decision tree and neural network respectively. In addition to this, in the high level customers (cluster6), the decision tree in which its accuracy level is 99.54% is less accurate than the neural net that has an

accuracy of 99.90%. The neural net has small wrongly classified high value customers than the decision tree.

Moreover, the neural network model has a better correct classification of the low value customers than the decision tree model. The neural network model has an accuracy of 99.92% in correctly classifying low value customers in cluster4; while the decision tree model has 99.23% of classification accuracy.

The incorrect classifications of the models may be created due to the result of the replacements of the outlier's data by their Modal values in the cluster model building. In addition to this, some customers' category have very high values of attributes. For example most of the customers have saving account for their personal use and other such variables are common in the dataset, as a result, such values could have dominated the other values of the variables and may resulted misclassifications.

In conclusion, the neural network model has shown a better classification accuracy and performance than the decision tree model. Therefore, it is reasonable to suggest that the neural net model is the best classifier for electronic expansion in banking industries.

5.2.7. Evaluation

Data is the nucleus of any data mining task, and the level to which the model meets the objective of the business is assessed in this stage. However, usually data required for effective mining is not prepared in the format the data mining requires. As it is stored in heterogeneous databases and in different format. This data might contain outliers, missing values, inconsistent data types within a single field, that must be cleansed, transformed, and integrated into the format suitable for data mining.

The goal of the business is to come up with a model that could find the appropriate number of clusters of customers according to their possibility of response to the CRM and direct marketing operation and also to assign new customers to the appropriate cluster index. Hence, a data pre-processing task is performed to this end.

One of the core segmentation output evaluation is with regard to the probability of customers to give a positive reaction to the Electronic Fund Transfer (EFT) of the POS service and it is defined based on the different financial information for the EFT of the POS service customers. Therefore, classification and clustering models were developed to fulfil the basic business objective of the payment card system and marketing department of the bank.

The model building part includes a clustering model building using the k-means algorithm. This task which was repeatedly conducted resulted in different cluster models that segmented customers. Even though, the various values of k (i.e. k=6, 5, 4) have a reasonable good customer segmentation, the model with k=6 has a better segmentation than all because it comprised potential EFT of POS service customers.

Besides this, the classification models are created with WEKA 3-7-2 J48 decision tree and neural network Multilayerperceptron algorithm. With this various models of the J48 decision tree and neural network algorithm is experimented with a 10-fold cross validation and by splitting of the dataset into 70% training and 30% for testing. This was tested with different parameter values of both the decision tree and neural net model. As a result, the decision tree model with default parameter values gives a better description of segments and shows rules that have precious support to assign new customer record to one of the clusters and it generated a better overall classification accuracy and individual classification accuracy in classifying high and low EFT customers. At the same time the neural network classification model with default parameter values (i.e. default hidden layer and learning rate values) have shown a better overall and individual classification accuracy in classifying both low and high level EFT of POS service customers'. And generally, when the J48 decision tree and neural network classification models are compared, the neural net generated better individual and overall classification accuracy level.

In conclusion, the obtained result is very satisfactory because it shows the possibility of applying data mining techniques to solve the exiting problems of the bank (Dashen Bank S.C.). It is believed that, if further research is undertaken by any other experts, they

might obtain better result that could give them similar customer behaviour that can help the Dashen Bank S.C. to go one step forward in increasing EFT of the POS service customers.

5.2.8. Model Deployment

The Dashen Bank S.C. doesn't use target data to select potential customers, rather they use general information to expand the EFT of POS service customers. Therefore, the segmentation output obtained is very encouraging; it could be used for this modern banking strategy and other operational tasks of the bank.

The result of this research proved that behavioural segmentation is possible and is best to identify and create a long lasting relationship with customers. In order to apply this result, the generated rules must be converted to more specific rules, which are simple to understand and apply. Therefore, this model can be deployed with further modification and evaluation to identify and provide special services to the bank customers. This research can also help to identify potential market place to EFT of POS service after analyzing further segments investigated.

CHAPTER SIX
CONCLUSION AND RECOMMENDATIONS

6.1. Conclusion

Identifying the characteristics of EFT of POS users of bank customers helps in selectively expanding the EFT of POS service among different bank customers. The result of this data mining experimentation shows that different categories of customers have their own potential in using EFT of POS service.

This study assessed the application of data mining techniques, classification and clustering to support expansion of Electronic Fund Transfer (EFT) of POS service at Dashen Bank S.C. The service has been introduced in Ethiopia since 2006 G.C by Dashen bank. However, it has not expanded and used as expected.

The idea of this research is intended to analyze the EFT of POS service along with the customers' data so as to investigate different segments of customers. This in turn helps the bank to see which type of customers intensively uses the EFT of POS service, thereby find possible market so as to expand it.

CRISP-DM process model was used while undertaking the experimentation. CRISP-DM process model includes business understanding, data understanding, data preparation, modeling, evaluation, and model deployment phases. Both the clustering and classification models were created for the different experimentation purposes.

On the other hand, this research is designed to investigate very high, high, medium, and low value customers based on their EFT of POS service behaviour usage. The main goal of this research is to come up with good cluster segments of customers according to their EFT of POS service usage. And this model used to differentiate potential EFT of POS service customers who should be addressed in the electronic transaction expansion marketing using the available information in the appropriate variables. The identification of these different types of customers enables the payment card system marketing

department of Dashen Bank S.C. to create good customer relationship management while serving special services to potential customers.

Consequently, in order to fulfil the above stated objectives, goals of this study, different related works in the area of data mining , Card system, CRM, and customer segmentation have been carried out.

The data, cardholder customers' profile, used in this research is gathered from the payment card system section of the bank. The processing to change this data into a format suitable for the data mining tasks to be performed for this research has been carried out.

A model is built that segments customers in different groups according to the probability of their result to the EFT expansion and it has obtained encouraging output. The fundamental criteria used to evaluate the output of the segmentation were measured based on some variables that are indicated at the first experimentation. These variables were selected as basic attribute by the domain experts and GainRationAttribueEval from the WEKA 3-7-2 select attribute evaluator.

After data pre-processing and data preparation, different clustering models are built using the k-means algorithm, with values of K= 6, 5, and 4. These values of K are selected and experimented after a discussion with domain experts. The clustering model built at 4, 5, and 6 values of k is evaluated and interpreted. From all these models, the model at K=6 value shows a better segmentation of customers. It enabled the creation of dissimilar clusters of high, medium, and low value of customers. Generally, this model, model at K=6, has separated high and medium level customers. And hence accepted and approved by the domain experts.

On top of this, the classification models are experimented with the J48 decision tree and neural network Multilayerperceptron algorithms. Different models for each of these algorithms are built, and the best overall classifier model from the decision tree J48 and neural net Multilayerperceptron algorithm has been selected. The overall classification accuracy, accuracy in classifying potential customers, and accuracy in classifying low

value customers with minimum error parameters have been identified together with the domain experts for both algorithms. And thus, the J48 decision tree and neural net algorithms with default parameter values have shown better accuracy level. Finally, the researcher compared these two different classification models and it has obtained a neural network algorithm with default parameter values which has scored a better performance.

The output of this research is very encouraging as it identifies high, medium, and low value customers properly. The neural network classification model with default parameter values has resulted better overall classification accuracy of **99.97%**. Therefore, the researcher believed that the model built in this study will bring a better outcome in EFT of POS service expansion, as well as the result of this research can be obtained as large percentage as of the targeted customers to the service suggested and also minimize the marketing cost.

While undertaking this research, considerable time of this study has been spent on the data preparation as the bank doesn't have data warehouse. Besides this, it was time taking to consult and present on the usefulness of the models, which were used for clustering and classification tasks of data mining to the domain experts. However, different parameter values for both the neural network (ANNs) and decision tree have not been tried.

6.2. Recommendations

This research discovered the application of data mining techniques for modeling the reaction of customers in the EFT of POS service expansion though this is done for academic purpose. The researcher gives the following recommendations based on the investigations of the study.

Develop Company Data Warehouse

It is known that, data is the principal and heart of data mining task. The quality of data matters on the performance of the data mining algorithms in extracting previously unknown knowledge. Hence, the data must be cleansed, integrated and transformed in the format suitable for data mining tools. So, data pre-processing, and data preparation took much of the time given for this research, but had there been an established company data warehouse before hand, the time allotted for the data pre-processing/preparation could be minimized. The bank does not have a data warehouse that contains all the important information that could help the data mining tasks/processes and other statistical analysis such as OLAP.

Model Performance Improvement and Future Data mining Researches

Even though, this research has a better overall classification accuracy of **99.97**% with the neural network Multilayerperceptron algorithm, other classification model parameters like neural network and decision tree parameter values might give a better or nearest performance in predicting the class level accuracy. But this research has emphasized on clustering modeling with K-means algorithm, J48 decision tree and neural net Multilayerperceptron classifier only. Hence, different classification and clustering algorithms can be applied in EFT of POS service expansion. Therefore, additional research is suggested to compare the performance of different clustering and classification algorithms by using different techniques like time series, summarization etc with different parameter values.

Data mining needs very large amount of dataset for both training and testing. Thus, increasing the dataset size/records may give a better clustering and classification models.

Besides this, additional attributes especially attributes that contain the financial information of the cardholder customers' behaviour can bring a better clustering and classification modeling/performance. In addition to this, other clustering algorithms like HierarchalClusterer, FilteredClusterer, MakeDensityBasedClusterer, EM, and classification models like LMT, Id3, and RandomForest has to be integrated to show and compare different algorithms , and thus to come up with a reasonable and best system.

While data mining is an iterative and interactive process, it needs modification and update and therefore, the researcher suggested applying these different data mining techniques for classification and clustering models for a future research is appropriate.

Data Mining Awareness and Providing Support to Researchers

Individuals at different levels in the organization should be aware of the preconditions towards creating a long lasting relationship with customers. And data mining is at the heart of customer relationship management. Therefore, it should be integrated in an attempt to identify, manage, and create relationship with profitable customers. To utilize data mining capabilities, awareness on its qualities should be created among employees of the bank.

As this technology is crucial in managing customers, the employees especially the IT (payment card system) should have knowledge of data mining. The researcher strongly recommends the bank to have a research centre that will support for future researchers in giving necessary information for different fields of study as it is crucial for its future dream. If they provided this, the outcome of the researches in the bank will increase. The bank uses complex and huge amount of data in different operations, the research centre will help applying data mining technologies in this operations.

To the Bank

There are many cardholder customers, but they are not ATM and EFT of POS service users; hence the bank should make intensive advertisement and promotion to attract more customers on EFT of POS service. Besides this the POS terminal should be installed in different business areas which can be easily accessible for customers' in order to achieve

the dream of the bank in expanding the modern banking system. According to the experimentation made, male customers are high probable users of EFT service but an emphasis should be given for female customers for the betterment of the bank profit.

Other Organizations

Customer Relationship Management (CRM) contains different functions like customer segmentation, cross-selling and so on. Only customer segmentation has been conducted in this research. But, applicability of data mining in different functions of CRM can be further discovered and researched in the specified company, and other organizations which have a high interaction with customers and which have other competitors should use this technology.

REFERENCES

Askale, W. (2001)' Possible application of data mining technology in supporting loan disbursement activity at Dashen Bank S.C.', Unpublished Master's Thesis, Addis Ababa University, Addis Ababa.

Bekezela (2008)' Financial Standards Forum in Ethiopia', Ncube publishing, vol. 3, pp. 46-49.

Berry, M. and Linoff, G. (1997)' Data Mining Techniques for Marketing Sales and Customer Support', John Wiley & Sons publishing, New York, pp.34-45

Berry, M. and Linoff, G. (2004)' Data Mining Techniques for Marketing Sales and Customer relationship management', 2nd edn, Wiley publishing, Inc. Indianapolis, Indiana., vol. 2, pp. 12-33.

Bouckaert, R., Frank,E., Hall,M., Kirkby,R., Reutemann,P., Seewald,A., and Scuse,D.(2008)'Weka Manual', University of Wakiato, version 3-6-0.

Bounsaythip,C. and Rinta-Runsala,E.(2001)' Overview of data mining for customer behaviour modeling', version1, VTT information Technology.

Bull,C.(2003)'Strategic issues in customer relationship management implementation', business process management journal, Categorical values: Data Mining and Knowledge Discovery ,pp 283–304 .

Chen,R.,Wu,R. and Chen,J.Y.(2005)' Data Mining Application in Customer Relationship Management Of Credit Card Business', Proceedings of the 29th annual international computer software and applications conference (COMPSAC'05) 0730-3157/05, viewed 6 November 2008, Institute of information management: national chiao tung university,Taiwan.

Colet,E. (1997)' Clustering and Classification: Data mining approaches', vol. 3, pp.23-26.

Dani, R. , Radha,K. and Subramanian,V.(2001)' An Electronic Payment System Architecture for Composite Payment Transactions', Journal of payment card system, No.2.

David, L. and Yasmin, A. (2006)' Banks and Data mining in electronic commerce', Journal of statistical Science, vol. 21(2), pp. 234-246.

Deshpande, S. and Thakare, V.(2010)' Data Mining System and Applications', International Journal of Distributed and Parallel System (IJDPS), vol.1, pp. 32-44.

Donald, L. and Sparks, W. (2008)'International Review of Business Research Papers: Electronic payments in sub sahran Africa', Vol. 4, No.1, pp.325-336.

Dorian, P. (1999)' Data Preparation for Data Mining', Morgan Kaufmann Publishers, Inc.taniwam, pp. 34-36.

Edelstein, H. (2000)' Building profitable customer relationships with data mining', Journal of Electronic Commerce Research, VOL. 3, NO. 1.

Encarta dictionary: Microsoft Press (2009).

Fayyad,Usama,Piatetsky-shapiro,G. andv Smyth,P. (1996) 'From data mining to knowledge discovery in data bases', vol.3,pp. 34-56.

FTC (2006)' Facts For consumers',International journal of Electronic banking, pp. 6-9.

Gray , P. and Byun,J. (2001)' Customer Relationship Management',vol.2, No.3.

Grover, E. (2007)'The future of network development: what future for MasterCard and Visa', Lafferty International Cards & Payments Council Conference, Cape Town.

Han, J. and Kamber, M. (2006)'Data Mining: Concepts and Techniques', Second Edition, Morgan Kaufmann Publishers, San Francisco.

Han,J. and Kamber, M.(2001)' Data Mining: Concepts and Techniques', Morgan Kuffmann publishers, San Francisco.

Henok,W. (2002)' Application of data mining techniques to support Customer Relationship Management at Ethiopian air lines', Unpublished Master's Thesis, Addis Ababa University, Addis Ababa.

Huang, Z. (1998)'Extensions to the K-means algorithm for clustering large datasets with categorical values: Data Mining and Knowledge Discovery', pp. 283–304.

Isayas, M. (2007)'Dashen to Resume Issuing Visa Cards', All Africa Global Media Publisher, Visited on June 13, 2008.

James, G. (1996)' Electronic Fund Transfer System', United states patent, patent number 5,546,523, Date of patent:Aug 13.

Jeffrey, W. (2004)'Data Mining: An overview', vol.3, pp. 14-18.

Joseph, P. (2002)'Data mining with neural networks: Solving business problems for application development to decision support', New York:McGraw-Hill.

Kelly,S.(2002)' Mining data to discover customer segments' ,Henrystewart publications 1478-0844, Interactive marketing,Vol.4, No.3, pp. 235-242.

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.

Lori, A. (2006)'Data Mining for Information Professionals', vol.10, pp.120-128.

Maalouf,L.and Mansour,N.(2006)' Mining airline data for CRM strategies',welley publishing, vol.13, No.2.

Maeda, Y. and Yokoyama, T. (1983)'Simplifying Key Management in Electronic Fund Transfer Point of Sale Systems', Vol.19.

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

Mudimigh,A., Saleem,F., Ullah,Z. and Aboud,A.(2009)'Implementation of data mining engine on CRM-improve customer satisfaction', vol.4, No.4.

Nielsen, F. (2001)'Neural networks: algorithms and applications', Niels Brock Business College, Supervisor: Geert Rasmussen.

Onut, S. and Erdem,I.(2000)'Customer Relationship Management in banking sector and a model design for banking performance enhancement', No.5, pp. 25-29.

Osmar, R. and Zaiane, B. (1999)'Principles of Knowledge Discovery in Databases', vol.5, pp. 89-102.

Pete, C. (2000)'CRISP-DM 1.Step-by-step data mining guide', CRISP-DM consortium, vol.6, pp. 60-71.

Pete,C.,Julian,C.,Randy,K.,Thomas,K.,Thomas,R.,Colin,S. and Rudiger,W. (2008)'CRISP-DM step by step data mining guide', The CRISP-DM consortium, viewed 4.

Pritscher ,L. and Hans, F. (2008)'Data mining and strategic marketing in the airline industry' , Zurich-airport, Switzerland.

Rajanish, D.(2002)'Data Mining in Banking and Finance: A Note for Bankers', Indian Institute of Management, Ahmadabad.

Rene, T.(2010) 'Applying Data Mining to Banking', Business Management Articles, vol. 5, pp. 3-7.

Saarenvirta, G. (1998)'Mining Customer Data of Credit Card Business', No. 7, pp. 1-12.

Shegaw, A. (2002)'Application of data mining technology to support in preventing and controlling child mortality at Butajira', Unpublished Master's Thesis, Addis Ababa University, Addis Ababa.

Shiffaman,G. and Kanuk,L. (1991)'Consumer Behavior', 4th edition. Prentice Hall,Inc.

Shigeki, K.(2006)'An Interpretation Method for Classification Trees in Bio-data Mining', KES, Part II, LNAI 4252, pp. 620 – 627.

Singapore Institute of Management (SIM) (2002)'Data mining and customer relationship', vol.10, pp. 34-38.

Sofia and Seid (2008)' Change Strategy: The Case of Dashen Bank', Deliverable III, customer relationship management, November 2005, Spartan.

Spartan InfoTech Co.WLL (2008)' Effective Customer Relation Management', vol .1, No. 2, pp. 26-29.

SPSS INC. (1991)'Rapid Response in Government: Fight Crime and Improve Security with Data Mining', Seminar on February 25, 2003.

Srivastava, J. (1991)'Data mining for customer relationship management', viewed 15 July 2008, vol. 12, pp. 4-8.

Sumanjeet, S.(2009)'Emergence of payment systems in the age of electronic commerce: The state of art', Global Journal of International Business Research, Vol. 2, No. 2.

Thearling, K. (2004)' Data Mining and Customer Relationships', Vol.3, pp. 9-14.

Tilahun ,M. (2009)'Possible Application of Data mining Techniques to Target Potential Visa Card Users in Direct Marketing (The case of Dashen Bank S.C.)', Unpublished Master's Thesis, Addis Ababa University, Addis Ababa.

Two Crows Corporation (1999)' Introduction to data mining and knowledge discovery',3rd edn, pp. 3-18.

Two Crows Corporation (2005)'Introduction to data mining and knowledge discovery', 3rd edn, pp. 4-16.

Tzyh, H. and Michael, H. (2006)' Core capabilities for exploiting electronic banking', Journal of electronic commerce research, Vol.7, No.2.

Witten, I. and Frank, E. (2000)'Data mining: Practical Machine Learning Tools and Techniques with Java Implementations', Morgan Kaufmann publishers, San Francisco.

Xinying, L. and Peizhi, W. (2008)'Data Mining Technology and Its Application in Electronic Commerce', vol. 3, pp. 56-65.

Xiong.H., Wu,J. and Chen,J.(2006)'K-means clustering versus Validation Measures: A Data Distribution Perspective', vol. 6, pp. 34-50.

APPENDICES

Appendix 1: The Original Collected Sample Data

D	E	F	G	H	I	J	K	L	M	N	O
Date Of Captuer: Month,Day,Year	*Address*	Sex	*Date of Birth*	*Saving Account*	*Current_Account*	*Total_Number_of_Years_being_card_holder*	*Age*	*Age:Young,Middle-Aged,Old*	*Available Balance*	*ATM*	*POS*
7/10/2007	AA	M	2/10/1964	No	yes	4	47	Middle-Aged		YES	YES
5/12/2010	AA	M	2/1/1978	No	yes	1	33	Young		YES	YES
5/12/2010	AA	F	5/5/1960	No	yes	1	51	Old		YES	YES
5/12/2010	AA K-11 NHO-1186	M	11/28/1975	No	yes	1	36	Middle-Aged		YES	YES
5/12/2010	AA HNO-417	M	10/11/1976	No	yes	1	35	Young		YES	YES
5/12/2010	AA	M	6/3/1980	No	yes	1	31	Young		YES	YES
5/12/2010	AA	M	7/8/1970	No	yes	1	41	Middle-Aged		YES	YES
5/12/2010	AA	M	1/6/1966	No	yes	1	45	Middle-Aged		YES	YES
5/12/2010	BD	M	10/28/1971	No	yes	1	40	Middle-Aged		YES	YES
5/12/2010	AA	M	1/12/1979	No	yes	1	32	Young		YES	YES
7/10/2007	AA	M	8/8/1980	No	yes	4	31	Young		YES	YES
5/12/2010	AA	M	10/2/1982	No	yes	1	29	Young		YES	YES
5/12/2010	AA	M	11/30/1967	No	yes	1	44	Middle-Aged		YES	YES
5/12/2010	AA	M	8/3/1983	No	yes	1	28	Young		YES	YES
5/12/2010	AA	M	1/24/1984	No	yes	1	27	Young		YES	YES
5/12/2010	AA	M	10/12/1973	No	yes	1	38	Middle-Aged		YES	YES
5/12/2010	AA	M	8/9/1956	No	yes	1	55	Old		YES	YES
5/12/2010	AA	M	12/21/1978	No	yes	1	32	Young		YES	YES
5/12/2010	AA	M	1/2/1972	No	yes	1	39	Middle-Aged		YES	YES

Appendix 2: Partial View of the decision tree generated with 10-fold validation technique

J48 pruned tree

Occupation = Hired

| Saving_Account = YES

| | Year_Of_Join = Joined_2010_11

| | | Marital_Status = Married

| | | | Age_Classification = Age_Group_One

| | | | | ATM = YES

| | | | | | Sex = M

| | | | | | | POS = YES: cluster3 (7.0)

| | | | | | | POS = NO: cluster4 (2.0)

| | | | | | Sex = F: cluster5 (139.0)

| | | | | ATM = NO

| | | | | | Available_Balance = High: cluster2 (0.0)

| | | | | | Available_Balance = Low: cluster2 (17.0)

| | | | | | Available_Balance = Medium: cluster5 (2.0)

| | | | | | Available_Balance = Veryhigh: cluster2 (0.0)

| | | | Age_Classification = Age_Group_Two

| | | | | ATM = YES

| | | | | | Available_Balance = High: cluster1 (12.0)

| | | | | | Available_Balance = Low: cluster3 (55.0)

| | | | | | Available_Balance = Medium: cluster3 (199.0)

| | | | | | Available_Balance = Veryhigh: cluster3 (362.0)

| | | | | ATM = NO

| | | | | | Sex = M: cluster3 (81.0/1.0)

| | | | | | Sex = F

| | | | | | | Available_Balance = High: cluster2 (0.0)

| | | | | | | Available_Balance = Low: cluster2 (16.0)

| | | | | | | | Available_Balance = Medium: cluster3 (3.0)

| | | | | | | | Available_Balance = Veryhigh: cluster2 (72.0)

| | | | Age_Classification = Age_Group_Three

| | | | | Available_Balance = High: cluster4 (0.0)

| | | | | Available_Balance = Low

| | | | | | Sex = M: cluster4 (8.0)

| | | | | | Sex = F

| | | | | | | ATM = YES: cluster4 (2.0)

| | | | | | | ATM = NO: cluster2 (3.0)

| | | | | Available_Balance = Medium

| | | | | | Sex = M: cluster3 (4.0)

| | | | | | Sex = F: cluster5 (54.0)

| | | | | Available_Balance = Veryhigh: cluster4 (147.0)

| | | Marital_Status = Unmarried

| | | | Sex = M

| | | | | Available_Balance = High: cluster4 (29.0)

| | | | | Available_Balance = Low: cluster4 (89.0)

| | | | | Available_Balance = Medium

| | | | | | Age_Classification = Age_Group_One: cluster4 (211.0)

| | | | | | Age_Classification = Age_Group_Two: cluster3 (11.0)

| | | | | | Age_Classification = Age_Group_Three: cluster4 (0.0)

| | | | | Available_Balance = Veryhigh: cluster4 (776.0)

| | | | Sex = F

| | | | | Available_Balance = High: cluster2 (11.0)

| | | | | Available_Balance = Low

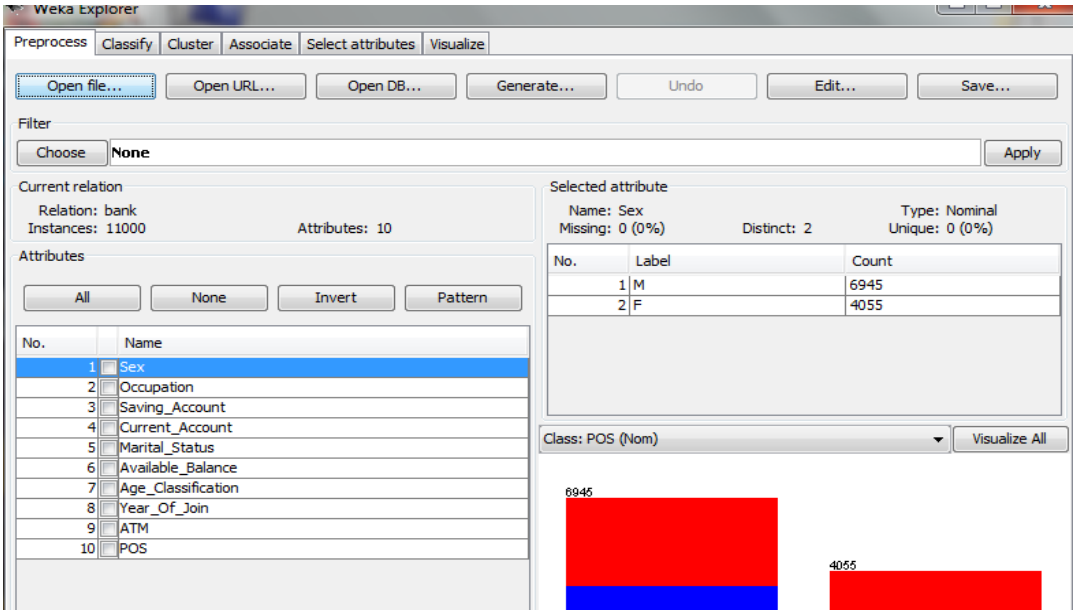
| | | | | | Age_Classification = Age_Group_One: cluster2 (8.0)

Appendix 3: Threshold values for Age_Classification and Available_Balance Attributes

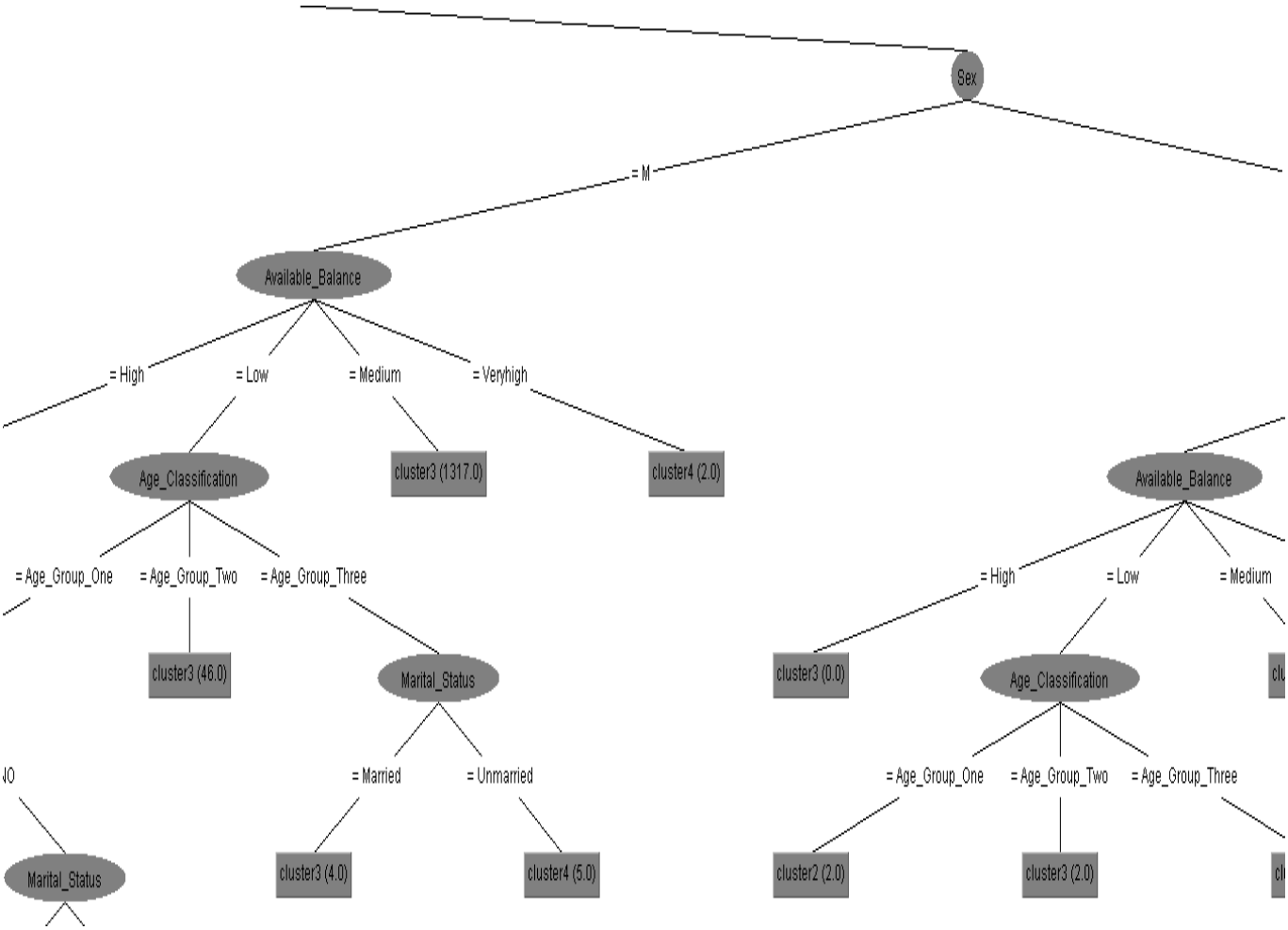
a. Age_Classification: This is the age of the card holder customer.
    i. If (Age>=18) and (Age<34) then Age is categorized as "Age_Group_One"
    ii. If (Age>=34) and (Age<50) then Age is categorized as "Age_Group_Two"
    iii. If (Age>=50) then Age is categorized as "Age_Group_Three"

b. Monthly_Available_Balance (MAB): This is the monthly balance of the card holder customers in the bank.
    i. If (MAB>=50) and (MAB<10,000)
        Then MAB is categorized as "Low"
    ii. If (MAB>=10,000) and (MAB <35,000)
        Then MAB is categorized as "Medium"
    iii. If (MAB>=35,000) and (MAB <60,000)
        Then MAB is categorized as "High"
    iv. If (MAB>=60,000) then MAB is categorized as "Veryhigh"

Appendix 4: The WEKA 3-7-2 Interface with the Dataset Opened to Start the First Clustering Run



Appendix 5: The Partial View of the Decision Tree J48 Model



Appendix 6: Sample Rules to Predict New Instances of Records Customers into Their Corresponding Cluster.

Rule 1: If occupation of the customer= “Hired”

Saving_Account=”Yes”

Marital_Status=”Married”

Age_Classification=”Age_Group_One”

ATM=”YES”

Sex=”M” and

POS=”YES”, then the customer is classified in **Cluster 3**

Rule 2: If occupation = “Hired”

Saving_Account=”YES”

Marital_Status=”Married”

Age_Classification=”Age_Group_Three”

Available_Balance=”Low” and

Sex=”M”, then the customer is categorized under **Cluster 4.**

Rule 3: If occupation = “Hired”

Saving_Account=”NO”

Sex=”M”

Available_Balance=”Low” and

Age_Classification= “Age_Group_One”

ATM=”NO” and

Marital_Status=”Married”, then the customer is categorized under **Cluster 3.**

Rule 4: If occupation = “Hired”

Saving_Account=”NO”

Sex=”M”

Available_Balance=”Low” and

Age_Classification= “Age_Group_Three”

Marital_Status=”Unmarried”, then the customer is categorized under **Cluster 4.**

Rule 5: If occupation=”Hired”

Saving_Account=”YES”

Age_Classification = “Age_Group_Two”

Sex=”M” and

Available_Balance=”High”, then the customer belongs to **Cluster 6.**

Rule 6: If occupation =”Private”

Age_Classification=”Age_Group_One”

Sex=”M”

Marital_Status=”Married”

Year_Of _Join=”Joned_2010_11” and

Saving_Account=”YES”, then the customer is classified in **cluster 1**

Rule 7: If occupation =”Private”

Age_Classification=”Age_Group_Two”

ATM=”YES”

Available_Balance=”Veryhigh” and

Marital_Status=”Married”, the customer belongs to **cluster 1.**

Rule 8: If occupation =”Private”

Age_Classification=”Age_Group_Two”

ATM=”NO”

Marital_Status=”Married” and

Sex= “F”, then the customer is classified in **Cluster 2**

Rule 9: If occupation =”Private”

Age_Classification=”Age_Group_One”

Sex=”F”

ATM=”YES” and

Marital_Status=”Married”, then the customer is classified in **Cluster 5**

Rule 10: If occupation =”Private”

Age_Classification=”Age_Group_Two”

ATM=”YES”

Available_Balance=”High”

Marital_Status=”Unmarried”

Sex=”M” and

Year_Of_Join=”Joined_2006_07”, then the customer is classified in **Cluster 6.**