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

APPLICATION OF DATA MINING TECHNIQUES TO SUPPORT CUSTOMER RELATIONSHIP MANAGEMENT (CRM) AT ETHIOPIAN TELECOMMUNICATION CORPORATION (ETC)

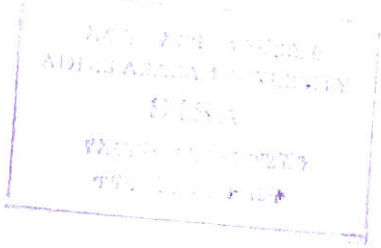
A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

BY

FEKADU MEKONNEN

JUNE 2004

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



APPLICATION OF DATA MINING TECHNIQUES TO SUPPORT CUSTOMER RELATIONSHIP MANAGEMENT (CRM) AT ETHIOPIAN TELECOMMUNICATION CORPORATION (ETC)

BY

FEKADU MEKONNEN

Name and Signature of Members of the Examining Board

Ato Nigussie Tadesse, Chairman, Examining Board _____

Dr. Nega G/Yesus, Advisor _____

Dr. Osei Nana Adjei, External Examiner _____

TABLE OF CONTENTS

ACKNOWLEDGEMENT

Above all, I must say few words of thank to GOD, the Almighty, who helped me in all respects during those challenging times.

I would like to thank very much my advisor Dr Nega G/Yesus for his valuable contributions during the preparation of the thesis. Moreover, his unmatched effort to support all students in those hard times really deserves deep appreciation.

I also would like to extend by thanks to Angoss Software Corporation to provide me the data mining tool which was very useful for the data mining tasks performed.

Special thanks go to Ato Tefera Tolosa, who has always been there for me when I needed his help in all matters.

Ato Tesfaye Birru, General Manager of ETC, and all pertinent staff were helpful to provide me access to the data and share experience of the telecom industry which was critical to complete my research.

I also would like take this opportunity to appreciate members of this year graduating class for the spirit of cooperation among ourselves during our stay.

Last, but not least, I would like to extend my heartfelt thank to all members of my family for their unreserved support whenever I am in need of it.

ABSTRACT

The value of relevant and reliable information in this globalized, competitive and dynamic business environment is too high as it allows optimal decision making in all respects. Especially, in customer-oriented businesses like in the telecom industry, adequate information regarding customers is vital so that appropriate customer relationships that maximize mutual benefit can be established.

Behavioral based customer segmentation, which is one of the core applications of Customer Relationship Management (CRM), provides useful insights for designing and implementing an appropriate CRM strategies and programs that result in success. To this end, data mining has strong potential in exploring natural segmentation schemes that lies within customers' data.

This study is an attempt to explore the customers' data to find underlying customer segments that may be useful for marketing decision making. The research followed the CRISP data mining process model.

First the business problem was analyzed and a corresponding data mining tool, techniques and algorithms were selected. Besides, relevant data was collected, analyzed and prepared.

Then, an automatic cluster detection models were built to choose the best attributes for the final segmentation. Next, the final automatic cluster detection model was generated, analyzed and evaluated together with domain experts with different parameters. Finally, a decision tree model, setting cluster index as dependent variable, was built. Corresponding rules to the decision tree were also generated.

The results obtained were encouraging and can be further refined for possible deployment.

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND

Customer Relationship Management (CRM) has become one of the hot topics towards business success in the new millennium. Bose (2002) defines the terms as an integration of technologies and business processes used to satisfy the needs of a customer during any given interaction move, specifically, it involves acquisition, analysis and use of knowledge about customers in order to sell more goods or services and to do it more efficiently.

Such an application of technology for the ultimate value creation to businesses has resulted in success both in the short run and long run. According to McKinsey Marketing Solutions (2002), 10-20 percent performance improvement was achieved to many businesses in less than a year. However, more importantly, its long term strategic effect has been accepted by many companies in order to enable them capture and retain profitable customers. It is pointed out further that the traditional CRM was less favored because of the following four reasons. First, efforts often fail to focus on the most impact full elements of the business. Existing data as well is often underestimated in creating new solutions. Another reason is that functions beyond marketing and sales are overlooked. Even results of the CRM plan are not properly tied to financial and budgeting process with targets and key matrices.

Fortunately, current technology has enabled businesses to reduce the impact of the above problems to a greater extent. Kotler (1998) states that, CRM principally revolves around marketing and begins with a deep analysis of consumer behavior. It uses information technology to gather data, which can then be used to develop information required to create a more personal interaction with the customer.

Marketing is the process of planning and executing the conception, pricing, promotion and distribution of ideas, goods and services to create exchanges that satisfy individual and organizational objective. Nowadays, customers, that have real value to a company, are the center of marketing strategies. Accordingly, businesses have found it essential to acquire new customer as well as to retain those that have high value. One of the major focuses in marketing is customer segmentation (Schiffaman and Kanuk, 1991).

Customer segmentation is the process of dividing customers into homogenous groups on the basis of common attributes and is at the heart of CRM. Segmentation describes the characteristics of customer groups (called segments /clusters) with the data. By determining similar classes of customers, more targeted communication is possible and marketing return on investment can be enhanced since marketing messages are accurately reaching those customers most likely to respond. Further more, different marketing strategies can be developed that are more appealing to members of the specified group. Segmentation requires the collection, organization and analysis of customer data (Bounsaythip, 2001; Schiffaman et al., 1991).

Database marketing techniques have been used to identify customer groups with high revenue potential, select criteria for mailing list and improve customer retention rate.

The knowledge derived from these segments will enable one to focus on more targeted promotion. Furthermore, knowing customer needs better and treating them accordingly can increase their lifetime value.

One of the major component/input for market analysis is the need for customer data. Such data/database should contain relevant attributes (fields) about virtually each customer. Witten and Frank (2001) write how such data has become extremely valuable only recently. Surprisingly, gathering raw data useful for marketing for the purpose of 'selling' it to marketers is becoming a recent business. Even government bodies (in U.S) are also realizing to sell personal data of citizens to these vendors to supplement their budget expenditure. E-mail service providers as well provide/give access to their customer database and even store messages in the electronic boxes that target individual based/direct marketing (Forcht & Cochran, 1999).

For many reasons, business database has grown dramatically but not by same rate with technologies to analyze and extract information from these voluminous data. However, recent break through, especially in the area of data mining and artificial intelligence, are very promising in handling and extracting valuable information for effective planning and decision-making purpose (Witten et al., 2001).

Data mining is defined by Two Crows Company (2003) as information extracting activity whose goal is to discover hidden facts contained in data bases using a combination of machine learning, statistical analysis, modeling techniques and database technology in the areas such as decision support, predication, forecasting and estimating. It is a process of uncovering of hidden information from the data that is useful.

The application of data mining technologies has been growing especially in CRM and particularly in the field of marketing. Generally, it is very useful to plan and implement effective customer relationship strategies, which basically are targeted for the maximization of the value of the business in the long run.

To effectively exploit the potential of data mining, databases (random data in some cases) should be first organized so as to further the mining process. This is one of the compelling factors to implement data warehousing strategies and application of supporting technologies so that a competitive business advantage could be safeguarded using the mining process. Hence, it is easy to infer that data warehousing and data mining are to be considered together (Forcht et al., 1999).

Saarenvirta (2002) describes that several data mining techniques are applied to support CRM. The most used methods for targeted marketing (market segmentation) are clustering and classification. They are used to segment the market into meaningful sets (classes) upon which specific marketing strategies are planned and implemented.

Classification techniques /rules are used to partition the database into predefined classes. Before the grouping process, class descriptions are done together with users (as per their decision making problem).It is a kind of supervised /directed learning by which each member/ case of the database is predicted to a class where it can fulfill class definitions. This is done using different kind of rules set for the purpose of classifying to appropriate classes (DSS Research, 2001).

Clustering is similar to classification but classes are not predefined. The model /system take into consideration the similarity between members and form clusters according to some matrix. The underlying principle of clustering is that members of a

cluster should be similar to each other and, as the same time, each cluster must be different from the other (Saarenvirta, 2002).

1.2 STATEMENT OF THE PROBLEM

Data mining technologies have proved themselves to be of high value in extracting information from customer data/database for the purpose of supporting decision-making. Especially in the presence of good collection of customers' data /database, businesses will undoubtedly obtain competitive advantage over their competitors by using these technologies and are very useful in creating value to the business in the long run. Companies, which involve in providing service like banks, insurance, telecommunication companies and supermarkets, are potential users of data mining techniques for their overall customer relationship management. *This research focuses the application of data mining techniques for Ethiopian Telecommunication Corporation (ETC) in supporting its efforts in customer relationship management.*

ETC is currently a fully government owned telecommunication company which provides solely various telecom services in Ethiopia. It is a multi-million capital organization providing telephone (fixed and mobile), Internet and various related services (ETC Online Report).

On one hand, the organization has rich customer related data on which the application of data mining techniques, especially classification and clustering methods, could result in valuable information for marketing decision-making. On the other hand, there is no such an integrated system or model being applied to segment customers and hence no clear-cut group/class based marketing orientation. Moreover, though most of the time, demand exceeds the supply, in certain type of services like in

different application of mobile telephone, there exists unsold capacity on which the organization invested millions of Birr. Moreover, the company is suffering from huge amount of uncollected revenue earned from past years.

The Ethiopian government also passed a directive stating any further expansion of the telecommunication infrastructure to be made from the proceeds /revenue generated from the sale of telecommunication services across the country.

As a matter of fact, it is quite visible that the demand for telecommunication service in the country will grow significantly as private businesses are expected to be established and/or expand. Together with the increase in population, the organization has to revisit its customer relationship management strategies and should support its undertakings with recent technologies so as to generate the required fund for its expansion from the fast growing demand.

There are also indications from the government side to sell part of the shares of the corporation. As the government needs to sell at a higher price, the organization's performance in customer service is one major aspect that can attract local/ foreign investment.

Since the time is when effect of globalization has been expanding through out the world and as the international finance organization (like the World Bank and IMF) are insisting for the permission of operation to foreign companies in Ethiopia, the company should expect the possibility that other foreign telecom companies may be allowed to operate in the country in the short run. To this end, the company should be ready in advance to have a competitive advantage and exploit its goodwill through

planning and implementing of an integrated CRM initiative/program. Otherwise it could easily lose in the global/ local market.

This study attempts to address questions in the area of market segmentation. The basic research question will focus whether there exists a natural and meaningful grouping of customers underlying/ hidden in the customer data that is useful for designing an appropriate CRM strategies for the purpose of maximizing revenues/reduce costs. In other words, the study explores which customers (group) are profitable and which are not.

1.3 JUSTIFICATION

The traditional way of customer CRM is being replaced with the emergence of IT-based customer centered one. Businesses are now planning and implementing various marketing strategies to gain competitive advantage over their rivals and create long-term value to their investment. Such application proved to yield better results. Besides, the environment is being rapidly changing because of information and communication technologies. Hence, adoption of this recent trend is becoming mandatory. However, it is important to note the fact that employing such sophisticated system wouldn't guarantee success and be sought from different perspective, most importantly, its long term cost benefit implication.

Despite these global developments, no critical efforts are under way to change the traditional customer relationship management style especially in the utilization of data mining technologies for customer segmentation/marketing and hence for supporting planning and implementing effective marketing strategies at ETC.

As the primary focus of this research is to explore the potential applicability of data mining techniques in the telecom industry, it is regarded as a timely effort to respond to the local and international market conditions and demand. Future trends as well suggest the need for an advance preparation for designing appropriate customer relationship management strategies in the principle of long-term value creation and integration to various functions of the organization as a whole. Hence, in view of the organization's problems and opportunities, conducting such kind of study is undoubtedly relevant.

1.4 OBJECTIVES

1.4.1 General Objectives

The general objective of this research is to help the organization maintain an appropriate Customer Relationship Management (CRM) through the application of data mining techniques (clustering and/or classification) for the purposes of transforming customer data into meaningful segments of customers based on their underlying/defined similarities.

1.4.2 Specific Objectives

The specific objectives are:

1. To identify the type of customer data residing in the operational database.
2. To select the data mining tool and algorithm to be used based on the type of data mining function to be performed.

3. To prepare the data for analysis, this involves extracting the data and transforming it to the format required for the data mining algorithm (data selection, data cleansing, organizing the training data sets).
4. To build and train the data mining model.
5. To evaluate (test) the model.

1.5 RESEARCH METHODS

The general approach of the study is basically quantitative as the core process is to collect and analyze customer data. However, as the resulting clusters (market segments) should be meaningful /relevant for marketing decision making, a close relationship with the domain experts and understanding of the business operation is mandatory. In this regards it could have qualitative aspect as well.

1.5.1 Literature Review

Relevant literature (books, journals, magazines and the Internet) pertaining to the subject matter of data mining, customer relationship management and customer segmentation has been compiled.

1.5.2 Business Understanding

Interviews, observations and document review were made to assess the need of users, analyze the business problems, and have good background knowledge in interpreting results of the data mining process.

1.5.3 Data Mining Methodology

To solve the business problems and meet objectives that have already been identified, the researcher followed the following steps in order to develop a data mining model and employ data mining techniques.

a) Identifying Available Data Sources

The data source was a data containing customer records of Ethiopian Telecommunication Corporation. The data/database contains data pertaining to customers who use the mobile service provided by the organization on credit basis i.e., the target population was customers who are post paid mobile telephone users. The data set contains customer types and their respective attributes, and selection was made in respect to the organization needs and problems.

b) Data Collection and Preparation for Analysis

Before subjecting the raw customer data to analysis, it must be converted into a form suitable for analysis. Data understanding and preparation activities, such as analysis, editing, coding, cleaning, integration and transformation were conducted.

c) Build and Train the Data mining Model

After the data has been cleaned, formatted and transformed, the data was used to build clustering and classification models. A training, testing and validation data sets were used to generate clusters and an explanation (rules) of the dependent (target) variable in terms of the independent (input) variables.

d) Evaluating (Testing) the Model

In order to check the output of the model's performance, the process of the modeling and the results there of were done and counterchecked together with users and

domain experts of the organization so that the segmentation process provides useful information for making optimal customer related decisions.

1.6 SCOPE AND LIMITATION

The scope of this research is focused on the post-paid mobile telephone customers of the Ethiopian Telecommunication Corporation and their calling behaviour for a month. The study is restricted on building a data mining model for segmenting the customers, interpreting the resulting segments, and developing a classification rules for each segment. The development of a classification prototype was not feasible mainly because of time constraint.

1.7 APPLICATION OF RESULTS

The results from this research will support the routine and strategic decisions made by the Ethiopian Telecommunication Corporation. By planning and implementing an appropriate marketing strategy and provide an attractive offer through the right channel and at the right time, each customer contact is more likely to achieve its goal. And by the result, customers are provided improved services and the profitability of each customer will increase. It is considered to contribute a lot for the company in addressing inefficiencies of the existing customer relationship in the area of post paid (credit based) mobile telephone service. The research can be easily replicated for the pre paid mobile customers and fixed telephone customers with slight modification as the attributes that are used to describe calling behavior of a customer remain almost the same.

It is also believed that it could initiate researchers in the area as it is an initial attempt for exploiting the potential of data mining techniques in the telecom industry in the area of customer relationship management.

In general, the results may support marketing to develop/improve new/existing products and services based on the needs of customers. Moreover, marketing efforts target selected promotions that increase the average purchases made by a given customer (segment), thereby lowering marketing costs and as the same time improve customer satisfaction.

1.8 THESIS ORGANIZATION

This research contains seven chapters. The first chapter deals with a general background, statement of the problem, justification, objectives, research methods, scope and limitation, and the possible application of the research work. The second chapter deals with different aspects of data mining. The third chapter covers customer segmentation within the framework of customer relationship management and in relation to data mining. Chapter four presents data mining techniques and algorithms used for the segmentation process. Chapter five briefly introduces Ethiopian Telecommunication Company and its post-paid mobile telephone service. Chapter six, the most relevant chapter, discusses the different stages of the experimentation towards building the data mining model and interpretation of the results. The final chapter, chapter seven, deals with the conclusions and recommendations based on the investigations of the study.

CHAPTER TWO

DATA MINING

2.1 INTRODUCTION

The contemporary challenge of information explosion creates both challenge and opportunities to both organizations and individuals. Though not in the same pace, the advancement of ICT in general enables users of information to increase their capacity of utilizing information for various decision-making purposes. Recent fields are emerging by developing new theories and application frameworks in a move to manage the challenge and exploit the opportunities (Two Crows Corporation, 1999).

The biggest challenge lies in changing huge amount of data, most of the time stored in digital form, to relevant information and knowledge that enables decision makers make optimal decisions. For this purpose, various data processing, storage and communication systems are developed and being utilized by users significantly. These systems (software) allow users have relevant and reliable information on time, which is critical, especially in the business world filled with stiff competition (Witten et al., 2001).

One of the disciplines, which significantly contribute in changing data to valuable information, is data mining. Data mining is the process of exploration and analysis, by automatic or semi automatic means, of large quantities of data in order to discover meaningful patterns and rules. Data mining is also defined as a process of extracting nontrivial, implicit, previously unknown and potentially useful information (Han et al., 2001).

2.2 THE CORE OF DATA MINING

To exploit properly the potential of data mining, adequate knowledge and hands-on experience is very essential in the core areas of competency, which are data mining techniques, the data itself and the modeling process (Berry et al., 2000).

2.2.1 Data Mining Techniques

Data mining techniques refer to the conceptual approach to extract information from the data. Basically, there are three popular data mining techniques namely automatic cluster detection, decision trees and neural networks. Each technique is most appropriate for special kind of data mining task.

2.2.2 Data

On of the core components of data mining is the management (preprocessing) of the data itself. Data refer to facts that rarely are relevant for decision making. Because of information and communication technology (ICT), the production and dissemination of data increased alarmingly however the exploitation of the same to support decision making is not growing in the same pace. Data mining is one of the disciplines that try to fill this wide gap.

To make the data ready for data mining, a number of tasks should be performed. These include choosing of the right data, cleaning the data, understanding the data structure and integrating and transforming data. In most of the time, these preprocessing stages take significant share of the total data mining effort. Data warehousing solves considerable effort of this data processing and is an ideal place to start the data mining process (Two Crows Corporation, 1999).

Business understanding

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

Data understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

Data preparation

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

Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements

on the form of data. Therefore, stepping back to the data preparation phase is often needed.

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 that it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment

Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort, it is important for the customer to understand what actions are needed to be carried out in order to actually make use of the created models.

However, as the term 'data mining' is widely used to refer to the whole knowledge discovery process, the same use and understanding of the word is adopted through out this research and hence, the two terms mean the same thing here afterwards for convenience.

2.4 DATA MINING AND RELATED FIELDS

Data mining has a strong relationship and considerable overlap with other fields as it is a recently emerged but fast growing discipline. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization.

2.4.1 Data Mining and Data Warehousing

Data warehousing involves the integration, cleaning and transformation of data from various sources and from different formats. Such undertaking creates fertile ground for data mining as required data will be available in the data warehouse and considerable part of the data preparation work is completed. Since a consideration is given for data mining while designing a data warehouse, other real benefits could be exploited. However, a data warehouse is not a prerequisite 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 (Forcht et al., 1999).

2.4.2 Data Mining and Database Management

Data base management systems (DBMS) also provide good ground to data mining as the broader purpose of data mining is to discover knowledge from large databases. DBMS offer essential capabilities to data mining as it contains a consistent data model and advanced high level query languages that help users to get the required information for their need. The data about data (meta data) found in the databases is

also valuable for data mining as the nature of the attributes can be easily understood (Connolly and Begg, 2000).

These greatly facilitate the data mining process as data is collected, cleaned and integrated for modeling from different databases. In such environment, the data mining task could be done with less cost and time but with relevant output.

2.4.3 Data Mining and OLAP

There exists a kind of confusion about the distinction of the two terms. OLAP (On Line Analytical Processing) greatly utilizes the application of complex queries to analyze multidimensional data mostly from the data warehouse query languages. The user exploits OLAP capabilities to check his generalization and relationships in the data by forming the necessary query. That is, the user should first hypothesize and use OLAP and use the data for support. However in data mining, generalizations and relationships are generated from the data itself. It requires in depth analysis to induce new information from the data using relatively complex techniques (Connolly et al., 2000).

2.4.4 Data Mining, Artificial Intelligence (AI) and Statistics

Data mining was first introduced when AI, and statistical techniques were applied to common business problems. Well, the scope of data mining now widens from business problems. Its scientific application, for example, has dramatically increased for variety of purposes. The basic capability of data mining provided by these fields is the capacity of pattern recognition from data using complex and powerful techniques and tools like neural nets and decision trees. The advancement of Information and

Communication Technology (ICT) has also accelerated the growth of data mining both in terms of techniques and tools as well as scope of application (Mitchell, 1997).

2.5 DATA MINING PROBLEM TYPES

Usually, the data mining process involves a combination of different problem types, which together solve the business problem (CRISP-DM, 2000).

2.5.1 Data Description and Summarization

Data description and summarization aims at the concise description of characteristics of the data, typically in elementary and aggregated form. This gives the user an overview of the structure of the data. Sometimes, data description and summarization alone can be an objective of a data mining project. This kind of problem would be at the lower end of the scale of data mining problems.

2.5.2 Segmentation

This data mining problem type aims at the separation of the data into interesting and meaningful subgroups or classes. All members of a subgroup share common characteristics. Segmentation can be performed manually or (semi-) automatically. The analyst can hypothesize certain subgroups as relevant for the business question based on prior knowledge or based on the outcome of data description and summarization. However, there are also automatic clustering techniques that can detect previously unsuspected and hidden structures in data that allow segmentation.

2.5.3 Concept Descriptions

Concept description aims at an understandable description of concepts or classes. The purpose is not to develop complete models with high prediction accuracy, but to gain insights.

Concept description has a close connection to both segmentation and classification. Typically, there is segmentation before concept description is performed. The important distinction is that classification aims to be complete in some sense. The classification model needs to apply to all cases in the selected population. On the other hand, concept descriptions need not be complete. It is sufficient if they describe important parts of the concepts or classes.

2.5.4 Classification

Classification assumes that there is a set of objects – characterized by some attributes or features which belong to different classes. The class label is a discrete (symbolic) value and is known for each object. The objective is to build classification models (sometimes called classifiers), which assign the correct class label to previously unseen and unlabeled objects. Classification models are mostly used for predictive modeling. The class labels can be given in advance, for instance defined by the user or derived from segmentation.

Classification is one of the most important data mining problem types that occur in a wide range of various applications.

Data mining can serve as a research tool, during research and development phases. Especially the pharmaceuticals industry, as it involves testing of various combinations of chemicals to yield a drug. To choose which chemical have a potential for drug also is a problem that can be addressed by data mining which save million of dollars from expenditure on non-promising one. It can also be used for process improvement as most of the manufacturing processes are controlled with statistical methods based computer programs. The process itself has many inputs to produce a given output creating fertile ground for the application data mining and hence enabling to save input resources and get quality output. Data mining has resulted success is the marketing and customer relationship management. It provides useful insights about customers so that appropriate marketing campaign is implemented to only target customers (group) who most likely respond and become valuable. This will greatly help the firm to increase marketing efficiency and effectiveness. It serves for the identification and anticipation of customer needs and wants for the purpose of maximizing mutual benefit through provision of quality goods and services. It is playing leading role in employing CRM at every point where the organization has contact with the customer (Berry et al., 2000).

CHAPTER THREE

CUSTOMER RELATIONSHIP MANAGEMENT, CUSTOMER SEGMENTATION AND DATA MINING

3.1 CUSTOMER RELATIONSHIP MANAGEMENT

3.1.1 Overview

Many companies recently are adopting customer centric strategies programs, tools, and technology for efficient and effective CRM. This is because of the understanding that for making optimal decisions an integrated and detail information, which is reliable and relevant, is mandatory.

CRM has become number one focus in today's competitive market. More than ever, the ability to understand and mange close relationship with customers has become a prerequisite to achieve business goals. Past and present trends imply the relevance of customer information and knowledge to build strong relationships with their customers over longer period of time through providing customer satisfaction, as the same time, earn business value (Kim, Suh, & Hwang, 2003).

The nature and scope of CRM is not yet clear and many researchers in the area of marketing are undergoing for the development of conceptual foundations of managing relationships with customers. Researchers from computer and information science disciplines are also searching for methodologies, techniques and software tools to assist decisions in the management of relationships with customers.

Many argue that customers' centric orientation (one-one relationship) is a subset/extension of the marketing orientation (group based relationship). Others disagree to this argument and describe the trend as a fundamental shift from managing market to managing a specific customer. In the fist case (marketing orientation) the firm has control over the marketing mix, however, in the second case, the firm is directed by customer tastes and preferences. CRM principally revolves around marketing and begins with the deep analysis of customer behavior. CRM is based on the customer centric orientation to deal with different behavior of individual customer to obtain and maintain a share of each customer rather than a share of the entire market with the help of appropriate ICT (Xu, Yen, Lin, & Chou, 2002).

CRM is an information industry term for methodologies, software and usually internet capability that help enterprise mange customer relations in an organized way. It is also defined as an all-embracing approach integrating sales, customer service, marketing and other functions that touch customers so that by integrating strategy, people, process and technologies, relationships with customers, distributors, and suppliers are maximized. Basically, CRM is a notion regarding how well an

organization can keep the most profitable customers at the some time reduces costs, increase value of interaction and hence maximize profit (Bose, 2002).

CRM system can be viewed as information system aimed at enabling organizations to realize customers' focus. One view of CRM is utilization of customer knowledge to deliver relevant products and services to the customer. Another view focuses in the database technology specially data warehousing and data mining which are crucial for the functionality and effectiveness of CRM system. CRM normally involves business process change and introduction of new technology (Yang & Padmanabhan, 2004).

The purpose of CRM is to improve marketing productivity. Productivity is measured in terms of efficiency, effectiveness and economy. This can be achieved through creating cooperative and collaborative processes that help reduce transaction costs, increase revenue and finally create value to business during the lifetime of a customer. It is an integrated effort to identify, maintain and build up a network for the mutual benefit of both sides through interactive, individualized and value added contacts over a long period of time. In some cases, CRM is regarded as a dominant /core paradigm of marketing. It is regarded as a shift of marketing role from manipulating the customer to genuine involvement with customer through appropriate communication and sharing of knowledge (Parvatlyar et al., 2002).

CRM can also be seen as business strategy aimed at gaining a long-term competitive advantage of optionally delivering customer value and extracting business value simultaneously. For the vast majority of businesses, the ability to acquire, retain and enhance customer relationship is the last place left to find an advantage (Bull, 2003).

As indicted above, varied views (definitions) are forwarded by many authors for the term 'customer relationship management'. The definitions can be categorized into three as technology centric, customer lifecycle centric and strategic centric (Kellen, 2002). According to Kellen, the first category of definitions reflect the position of vendors who are involved in supplying CRM technology and they regard the term almost synonym to technology. The second category views the term as a four phased customer lifecycle namely attracting, transacting, service and supporting, and enhancing. This perspective is that of practitioners to describe flow of activities during a life time of a customer rather than focusing on product life cycle. The last category defines the term as a technique to compete well in the market and create value to the business.

3.1.2 Principles and Tasks of CRM

CRM principles

There are three basic principles of CRM implementation namely personalization, loyally and lifetime value. The first one deals about treating customers individually so that products and services are designed and offered based on the preferences and behavior of the customer. Loyalty refers to the company's retention capacity through continuous contacts / relationships so that the customer gets satisfied and less likely to swift to other companies. The last one focuses on the selection of good customers from 'bad' through analysis of their respective behaviors. Decision would be taken to drop bad ones and keep the good ones as the motive of the organization is to maximize its profits in the long-run and has limited resources to spend for customer

care. From economical perspective, it is less costly to retain a customer than to find a new one. The major goals of CRM are increasing revenue growth through customer satisfaction and reduce cost of sales, distribution and minimize customer support costs (Parvatlyar & Sheth, 2001, Gray & Byun, 2001).

Key CRM Tasks

The basic questions CRM tries to answer are basically two: to know the status of customers to find out their profitability level and to focus on strategies and ways the customers can grow to derive maximize profit to the business and as the same time keep customers satisfied and remain loyal.

The basic tasks to address these questions can be categorized into four tasks/processes: customer identification, customer differentiation, customer interaction and customization (Gray et al., 2001).

1. *Customer identification*

It is the first step and refers to the attraction and or knowing of customers through marketing channels, transactions and interactions over time for the purpose of growing to a profitable one.

2. *Customer differentiation*

This refers to segmenting of customers into different perspective from the company's point of view as each customer has their own lifetime value.

3. *Customer interaction*

As the customer is to be exploited for maximum long term value, analysis of his/her behavior over time is mandatory to know and offer the right goods and services at the right time.

4. *Customization (Personalization)*

The final goal of CRM is to treat each customer uniquely so that each customer long term value and/or loyally increases.

The above tasks are greatly facilitated with the use of IT.

3.1.3 CRM and IT

CRM technology facilitates communication and management of customers through automating the information channels like face to face, mail, phone, fax, web, and e-mail. Moreover, the technology enables the companies improve performance almost in every functional areas and business processes including sections that have closer contact with the customer like marketing, sales, field service, contact center etc. Data warehousing and data mining tools and other related technologies help to analyze and extract hidden knowledge from customer data and have become the backbone of CRM systems (Bose, 2002).

Though potentially all firms can benefit from CRM technology, for certain companies, it is best suited. These companies include those which accumulate lot of data on each customer on the course of their business like financial and telecommunication companies (IBM, 2001).

3.1.4 Nature of Customers

As customer are not equally profitable to the business, appropriate segments should be made to select those who are legible for a given marketing program or to determine what kind of marketing effort is necessary for each segment of customers to finally maximize the profitability of the firm.

There are three distinct types of relationship customers: the top, middle and lower groups. The tops are those with higher productivity and loyalty. These customer groups need a good treatment to keep them for longer period through offering the possible best offer, products and services. The middle ones are those who have the potential to be grown to the top group. Here, a lot to be done to make this group profitable and loyal. The lower group is marginally profitable. The task of targeting this group should be made in care, as the cost of such activity may be greater than the benefit derived. Treating this group of customers would affect the treatment of the other better groups, as the resources in any organization are limited to perform CRM projects. The identification of these three groups is one of the major tasks of CRM so that appropriate CRM strategies are designed and implemented for each group as CRM primarily involved in targeting the right offer to the right customer at the right time for the right price (Bull, 2003).

3.2 CUSTOMER SEGMENTATION

3.2.1 Overview

Customer segmentation is an expensive engagement. The success of shareholders investment is greatly dependent on how well appropriate marketing strategies and programs are designed and implemented. This design and implementation in turn is based on the quality of segmentation made, as each segment requires unique marketing strategy and program. Once a company adopts a specific segmentation scheme, all resources, skills and processes are directed towards it which ultimately determines the values and cultures of the organization (Ulwich & Elsenhauer, 2002).

Traditional segmentation, which is based on industry classification, business size, geography or some statistical classification, rarely meets the criteria for ideal segmentation scheme. Companies should strive to find natural segmentation scheme that groups customers with common outcomes and characteristics.

The challenge is to uncover natural segmentation scheme, as the way of discovery is not obvious. However, the benefit of such natural segmentation is far-reaching as the firm can discover a world of new markets of opportunities which other competitors cannot see until it is created. Such perspective helps companies to align resources, skills and values around these opportunities, resulting in higher return on capital and flexible for possible changes in the market. However many companies still built their values and cultures in a lesser than optimal segmentation scheme and should revisit their status and uncover possible natural segmentation schemes to guarantee their profitability in the long run (Ziberna & Zabkar, 2003).

On the other hand, higher income and education contributed a lot to have customers with complex tastes and intelligent decision making ability. The availability of information on a wide variety of products and services greatly facilitated buying decision making. Moreover, the number of suppliers has been increasing at faster rate than the growth in demand (Vrines, 2001).

These conditions have changed the market very hard to survive to sellers. To survive and become profitable in the market, it is the order of the day, for sellers, to think thoroughly in identifying every opportunity, prioritize their marketing strategies, and design and implement an appropriate program in order to gain competitive advantage for securing a better market share.

Segmentation, which basically is a marketing orientation approach, is no more provides a competitive advantage by itself and now often considered as a minimum requirement for doing business. The trend is to adopt customer centric approach by which every customer is treated individually and uniquely. Emphasis is also given not for profit from a given transaction but on customers' lifetime value (Ahola, & Rinta-Runsala, 2001).

Marketing data generally needs some processing to be accessible for effective and efficient management decision making. It is very critical to have relevant and reliable information about the customer needs, wants and behavioral characteristics to be able to adjust and direct marketing activities for the benefit of both parties: the seller and the buyer. Customer segmentation is a process of changing the raw data of marketing nature into groups to yield an understanding of customers in general. The bases of

segmentation may be based on key strategic variables to make the results of segmentation meaningful and relevant for making informed decisions. The segmentation process involves the following steps: decision of the data source, selection and determination of bases of segmentation, selecting variables to describe segments, sample selection, data gathering, formation of segments, description of segments (profile formation), and usage of segmentation results for problem solving (Ulwich et al., 2002).

3.2.2 Bases and Variables of Segmentation

One of the challenging tasks during segmentation is to choose and derive appropriate segmentation bases and variables on which the segmentation process is performed and resulting segments are interpreted. This greatly depends on the type of segmentation problem. Variables can be categorized as geographical, demographical psychological and behavioral (Basgoze & Gokturk, 2003).

Behavioral segmentation clusters end-users based on the requirements of the decision makers or in respect to the problem to be addressed. In many instances, behavioral based segmentation is recommended if the prime motive of the firm is to maximize benefit (profit) by identification of marketing opportunities. This is so because of the fact that customer behavior analysis enable to know the firm what are the needs and preferences of customers so that an appropriate marketing programs are designed and implemented accordingly that result in increased long term profitability through maximizing revenue and cutting of costs. Moreover, this kind of segmentation creates relatively homogeneous solution according to selected characteristics. As the

resources, skills and values are directed to maximize return, such kind of segmentation has got popularity these days (Bounsaythip et al., 2001).

However, problem of accurately and usefully describing segments is cumbersome. There is also a possibility of using combined variables taken from different categories but doesn't guarantee better segmentation results.

3.2.3 Criteria for Successful Segmentation

There are six basic criteria for a given segmentation to be relevant and lead to profitable strategies and programs namely *identifiability, substantiality, accessibility, stability, responsiveness and actionability*. Identifiability refers to how well the segments are separated into clusters that are dissimilar to each other but contain similar members within them. Substantiality refers to the size of a given segment whether it is legible for a design of a particular marketing strategy/program. Accessibility refers to the extent at which customers are reached through different channels including direct while stability is about the continuity of segments with their attributes in the future. Responsiveness refers to the level of initiation of customers in a particular segment to a given message over a specified period uniquely from other segments. The final criteria, actionability, is about the extent segments react and show marketing efforts in line with the organizations expectations and competencies (Vriens, 2001).

3.2.4 Segmentation in the Telecommunication Industry

Behavioral segmentation usually can be best utilized in the telecoms industry to manage customer relationships. The segmentation can be performed on reliable

transaction data as every call transaction is captured, stored and assessed in a form of Call Detail Record (IBM, 2002).

Before segmentation of this kind, first, the term 'customer behavior' should be clearly defined in context. In the telecom industry, customer behavior can be defined as the combination of calling behavior and service utilization behavior. The first refers to the patterns of call during a period of time and possibly may be represented by the duration of calls, number of calls made and total revenue generated. The second element of behavior, service utilization, could be represented by different ways that can measure the total number of services that each customer uses. Demographic data could be also used to describe clusters after segmentation (IBM, 2001).

3.3 APPLICATION OF DATA MINING IN CRM AND MARKET SEGMENTATION

The lifecycle of customers can be seen as four stages: prospects, new customer, established customer and former customer (Berry et al., 2000).

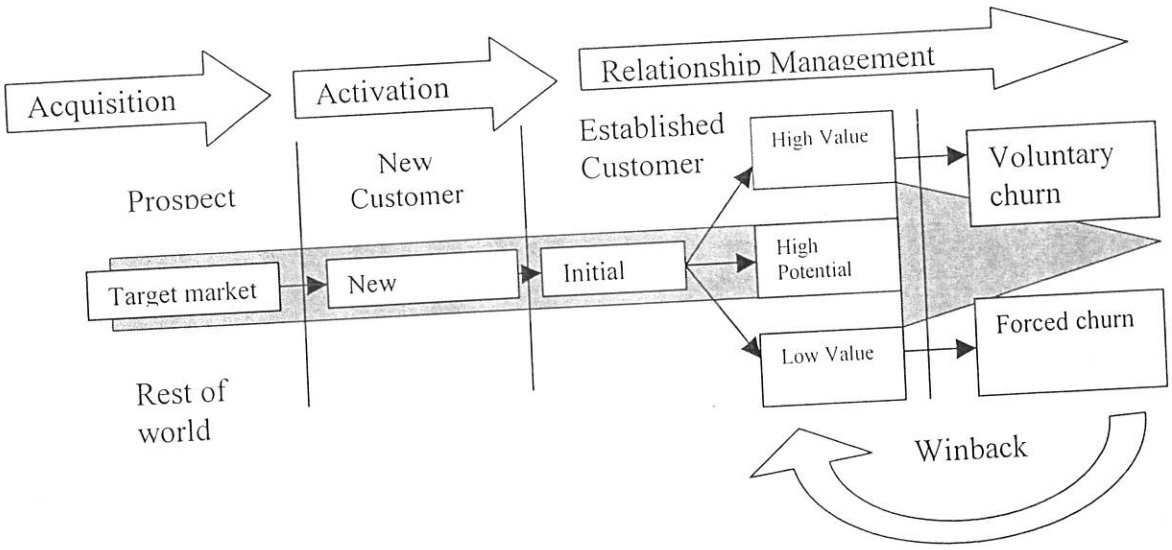


Figure 1 : Different stages in customer life cycle

The first stage is at a time where the potential customer has to be targeted using different marketing campaigns. Finding potential customers segments and profiling (using data mining) will help to send the right message to the most probable responder through an appropriate marketing channel.

At the second stage, potential customers responded to the marketing messages should be treated further so that they become established customers. Predictive data mining could help to determine which responders will become real customers and hence appropriate treatment can be forwarded.

At the third stage, customers established relationships and data related to their behavior is available. During this period, the customer should be stimulated and be grown to the maximum possible. To this end customer behavior modeling can be done using both descriptive and predictive data mining. Customers with high value

should be rewarded and retained. Those with high potential should be grown to the next level. However, those with a very low potential should be abandoned provided that the cost of carrying exceeds the benefit derived.

At the final stage, customers who established business relationships may churn due to many reasons. Data mining can be employed to identify those customers who are likely to churn. Even after some customers are lost, a marketing strategy should be designed to win back those customers who are churned to other competitors.

Different application of data mining call for different, and in many cases combination of, data mining techniques and methods. In the next chapter, data mining methods that are useful for customer segmentation and employed in this research are discussed.

CHAPTER FOUR

DATA MINING METHODS FOR CUSTOMER SEGMENTATION

4.1 OVERVIEW

The task of customer segmentation is done mostly using clustering and classification techniques in data mining. Clustering techniques are used for searching natural segments in the data based on the bases and variables selected. Various clustering techniques and algorithms exist and each is most suited for particular situations. On the other hand, classification techniques are used for finding rules (profiles) for each cluster so that a new customer can be assigned in a particular segment that matches best his/her characteristics. More over, the segment rules (profiles) serve as a basis for interpretation so that decision makers can understand the nature of each segment, and design and implement appropriate marketing strategies (Two Crows Company, 1999).

4.2 CLUSTERING TECHNIQUES AND ALGORITHMS

Clustering in data mining is a discovery process that groups a set of data such that the intra-cluster similarity is maximized and the inter-cluster similarity is minimized. The modeling process is unsupervised that is no prior knowledge is available to exactly guide the process. The data mining process is done together with domain experts so that the segmentation is in line with their problem/opportunity at hand and the results should provide new information/insight about customers and this should help to make

optimum decisions at different lifecycles of a customer on which discovery data mining is potentially useful (Bounsaythip & Rinsa-Runsala, 2001).

Clustering algorithms can be broadly divided into hierarchical and non-hierarchical ones. Hierarchical algorithms form a tree-like structure either in a bottom-up or top-down ways are called agglomerative and divisive approaches respectively. In the first case, agglomerative approach, each object (customer) is regarded as a separate (initial) cluster and at every consecutive step each cluster, with its closest cluster, forms another high level cluster until all clusters are under a highest cluster or a termination criterion is fulfilled. In the second case, however, all objects are placed first in one cluster, and by consecutive steps, the initial cluster is divided into smaller clusters until each object is treated as cluster or a termination criterion is fulfilled. Unlike in hierarchical clustering, non-hierarchical clustering procedures do not construct a tree like structure. Rather, cluster centers are initially selected (formed) and each object is assigned in a cluster based on its proximity to the cluster center (Ahola et al., 2001).

The K-means clustering algorithm is a non-hierarchical one and is very popular in data mining. Another neural network implementation algorithm, Self Organizing Map (SOM), is as well very popular clustering algorithm in data mining. The k-means clustering algorithm, which is used in this research, is used for customer segmentation by various researches even in the telecom industry (Bounsaythip et al., 2001).

4.2.1 The K-Means Method

The K-means algorithm is one of the most applied in practice and is relatively simple to understand and implement. It is primarily suitable to clustering a data set

containing variables with numeric (continuous) values. It is also possible to adjust the algorithm to fit to a data set with categorical attributes (Bishop, 1998; Angoss, 2002). The algorithm works based on the concept of distance and partitions the data set into predefined number of clusters (K). It initially assigns points randomly for each cluster and calculates cluster centroid for each cluster. At this point, the cluster centroids are the same as the values of the randomly selected/formed vector for each cluster. Then, a point (object, case, record) in the data set is taken and assigned to a cluster having the closest centroid to the object and cluster centroids are updated for the change (iteration) based on the distance of the point from cluster centroids. This process continues until all data points are assigned to given cluster. At the segment formation (partitioning of the data set into groups) is completed and, the analysis and interpretation of each cluster can be done.

According to Bishop (1998), the K-means algorithm involves a simple re-estimation procedure. Assuming there are N data points x^n in total, and the purpose is to find a set of K representative vectors μ_j where $j = 1 \ldots K$, the algorithm seeks to partition the data points $\{x^n\}$ into K disjoint subsets S_j containing N_j data points. This would minimize the sum-of-squares clustering function given by

$$J = \Sigma_{j=1}^{k} \; \Sigma_{n \varepsilon Sj} \; [x^n - \mu_j]^2 \qquad [1]$$

Where μ_j is the mean of the data points in set S_j and is given by

$$\mu_j = 1/N_j \; \Sigma_{n \varepsilon Sj} \; x^n. \qquad [2]$$

The calculation of the means can be formulated as a stochastic on-line process. In this case, the initial centers are randomly chosen from the data points, and as each data point x^n is presented, the nearest μ_j is updated using

$$\Delta\mu_j = \eta(x^n - \mu_j) \quad [3]$$

, where η is the learning rate parameter. Once the centers of the basis functions have been found in this way, the covariance matrices of the basis functions can be set to the co-variances of the points assigned to the corresponding clusters (Bishop, 1998). One of the challenging tasks in this process is the determination of the 'optimal' value of clusters (K) so that a natural segmentation scheme is discovered upon which the organization direct its resources. This value may also depend on the organization's capacity to manage how many customer groups at a time, as this require a design and implementation of particular marketing programs for each segment. The optimal value of K can then be arrived by segmenting the data set into acceptable values and comparing the results together with the business analysts (Saarenvirta, 1998).

4.2.2 Cluster Interpretation

Once the clusters have been created using clustering algorithms, they need to be interpreted. Though there are several approaches to understanding clusters, according to Berry et al. (2000), the three that are commonly used are:

1. 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.
2. Using visualization to see how the clusters are affected by changes in the input variables.

3. Examining the differences in the distributions of variables from cluster to cluster, one variable at a time.

4.3 DECISION TREES

Data Mining uses decision trees to classify objects to values of the dependent variable based on the values of independent variables. There are two main types of decision trees. Decision trees, which are used to predict categorical variables, are called classification trees because they place instances in categories or classes. Decision trees used to predict continuous variables are called regression trees. Classification trees label records and assign them to the proper class. Classification trees can also provide the confidence that the classification is correct. In this case, the classification tree reports the class probability, which is the confidence that a record is in a given class. Regression trees, on the other hand, estimate the value of a target variable that takes on numeric values.

A decision tree is a tree like structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes. The top most node of a tree is the root node and the lowest nodes are leaves. Each branch will lead either to another decision node or to the bottom of the tree, leaf node (Han et al., 2001)

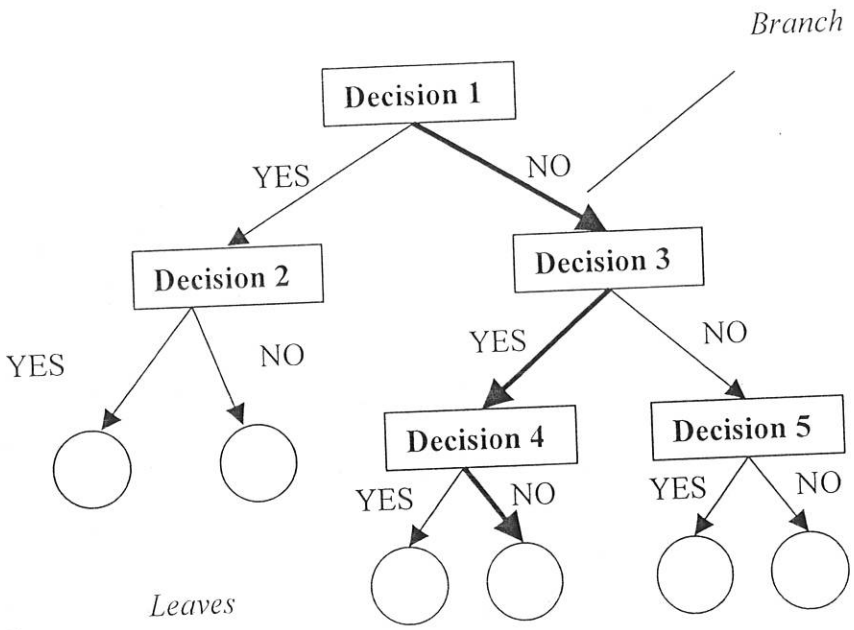


Figure 2 : A decision tree

4.3.1 Decision Tree Building

The basic algorithm for decision tree induction is a 'greedy' algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. The algorithm selects an attribute from the rest of attributes with a strategy of searching a local optimum solution (at each node) that leads to a global optimum solution. However this is not always true. For searching a local optimum solution, various methods can be employed to calculate the attribute selection measure. The iterative divide and conquer process executes until no further split is required (Witten et al., 2001).

4.3.2 Decision Tree Pruning

While constructing the decision tree, a stopping criterion is usually used to limit the depth of the tree and the minimum number of members required in a given node for further splitting. Alternatives from using a stopping criterion is to let the tree finish growing but use pruning methods to reduce the tree to a smaller size possible from its full size but without compromising accuracy (Han et al., 2001).

CHAPTER FIVE

A SURVEY OF CRM AT ETHIOPIAN TELECOMMUNICATION CORPORATION

5.1 OVERVIEW

The introduction of telecommunication in Ethiopia dates back to 1894. In order to achieve its objectives, the organization had undergone through series of development programs. The formation of current telecom bodies in the industry was made as follows (ETC Online Report).

- Ethiopian Telecommunication Board (1953)
- Ethiopian Telecommunications Authority (1974)
- Ethiopian Telecommunications Corporation (1996)

The purposes of ETC are:

- To engage, in accordance with development policies and priorities of the government, in the construction, operation, maintenance and expansion of telecommunication services;
- To provide domestic and international telephone, tele-fax and other communication services;
- To provide communication services using integrated information technology, including rebroadcast of television broadcasts;

- To repair, assemble and manufacture telecommunications equipment;
- To render training services to telecommunication personnel.

5.2 CUSTOMER RELATIONSHIP MANAGEMENT AT ETC

5.2.1 Service Delivery at ETC

The corporation has a strong commitment for realizing customer satisfaction, as it is evident in its vision, mission and objectives. At strategic level, the environment is conducive to design and implement various CRM policies and projects.

However in reality, a strong criticism is repeatedly forwarded from many customers and the private press that the telecom services available now are not to the standard, as the current dynamic environment requires.

It is not difficult to see that the reason for such strong criticism is the fact that the strategic policies and objectives are not implemented well through integration of people, process, technology and other resources. Employees are required to efficiently and effectively discharge their duties and responsibilities; appropriate business processes should be set up and be in place in view of customer satisfaction, and advanced technologies should be exploited to facilitate the overall operation of the company. Of course, these requirements are applicable to all business organizations that operate in this dynamic, competitive and customer oriented environment.

Taking this fact into account seriously, ETC has been designing and implementing various programs that enhance its operational efficiency and effectiveness. One of the core projects in place in this regard is the business process re-engineering using

advanced ICT focusing on appropriate customer relationship and care. Processes regarding service delivery are being studied and appropriate system analysis and design being finalized. Implementations are also undergoing at various organs of the corporation.

New communication channels like the web are being used to inform the public about the organization's general status and performance. The company's website, besides hosting other companies' websites, is being used to create good customer relationship and facilitate sale of telecom services. Use of other media (TV, Radio, and Press) has been also utilized for the same purpose.

5.2.2 The Mobile Telephone Service

The wireless telecom industry has been changing very fast providing state of the art technology especially in the mobile telephone industry. Mobile cell phones are widely used through out the world with diverse applications facilitating the exchange of data/information in different forms (text, audio, and video).

The ETC introduced mobile telephone service in 1999.Currently, Ethio Mobile Division is managing the services, which is one of the divisions in the organizational chart. The division has three sections; Mobile Technical, Mobile customer service, and Mobile resources.

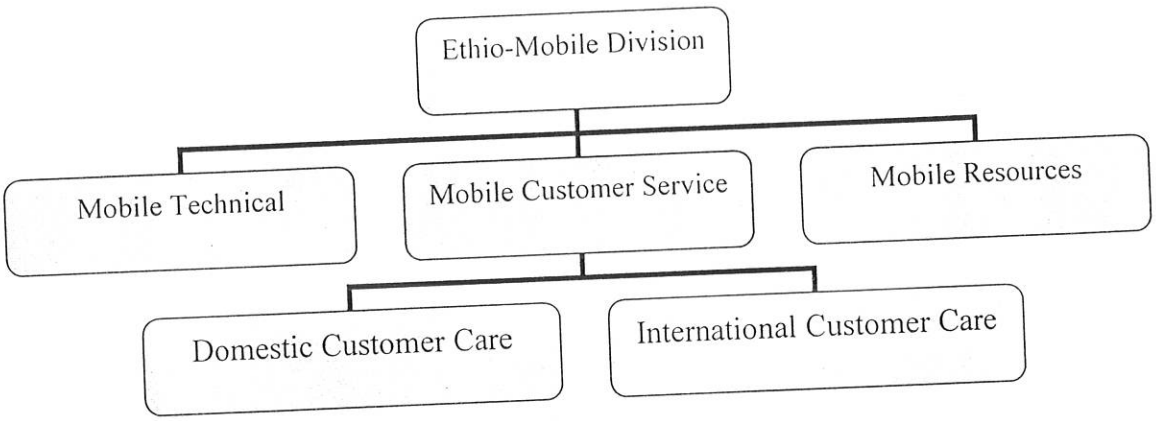


Figure 3 : Organization chart of Ethio Mobile division

The mobile technical section is responsible for the proper functioning of the switch and other related technical matters. The call detail record (CDR) is generated at this section and provided to mobile resources for bill processing. The mobile resources section is in charge of securing resources from concerned organs of the organization like personnel and logistics. It is also responsible for bill preparation. The mobile customer service section which has two subsections (Domestic and International) takes care of all customer relation ship beginning from sale of a mobile phone line. It also offers after sale services like providing security for loss of mobile phones, processing title transfer and treating customer complaints. More over, the complex task of revenue collection is another major responsibility of this section. The international customer care section is concerned with the relationship to other telecom companies worldwide and manages problems related to international services (like Roaming) and standards (GMS).

The sale of the service was credit based (post paid system) when it is first started. Later, the prepaid mobile system was introduced. At the moment, there are about one hundred thousand mobile telephone subscribers, about fifty thousand in each group (prepaid and post paid). Currently, mobile telephone services include additional services like call waiting and call forwarding with special arrangement with the corporation.

The current mobile services are available in the pre-paid scheme only. Sale of services in the post paid scheme was abandoned for the time being as the management and control of operations is relatively complex. The fact that revenues are collected in advance is one major reason the company chose the prepaid scheme. According to the finance department, about eleven million birr is not yet collected from post paid mobile customers. The amount is very significant as the company is in real need of funds to build its infrastructure and expand its services.

The demand for the post paid arrangement is increasing as it doesn't require the availability of prepaid cards. Besides, the demand for prepaid one is almost met especially in the capital. Because of this fact, the company is intending to re-introduce the post paid system.

This research is conducted on the post paid mobile customers' data of one month as the data was readily available relatively than the data for the pre-paid and fixed ones. The application of the research can be easily projected for the later two as it is concerned with basic attributes common to all telephone service users.

5.2.3 Customer Segmentation at ETC

Customer segmentation is one of the core components of CRM applications. In the case of ETC, customers are categorized as individuals, businesses, government, NGO and international organizations, and employees. The classification is primarily being used to set priority and requirements for each group to be eligible to purchase telecom services, as the demand for these services exceeds the supply at the moment. This kind of segmentation do not provide insights about customers and is less favored these days as it does not provide a basis to understand customers and manage relationships accordingly for the purpose of creating value to the business. It is the behavioral segmentation that provides considerable valuable information to design and implement marketing strategies and programs. There fore, the researcher believes that the output of this research would be of great help in this respect.

5.2.4 Relevant data sources in the organization

Call Detail Record (CDR)

CDR is transaction data generated by the switch for each call made by a customer. It contains a wealth of data on which data mining application could be successful. However it contains low level data and hence needs considerable data preparation effort. This data is a source of information to prepare monthly bills.

Data bases related to customers

There are two data bases maintained regarding mobile customers; billing data base and customer data base. The first data base which contains attributes taken from the CDR is used for the monthly preparation of bills. Attributes like monthly duration,

tax, amount of revenue are the major ones. The second data base is about the personal details of customers mainly containing demographic attributes like address, category, etc.

CHAPTER 6

EXPERIMENTATION

6.1 *OVERVIEW*

This chapter comprises the core component of the research. It deals with the description of the data mining process undertaken based on the CRISP data mining reference guide (model) which is discussed previously in the literature. This model has six phases: business understanding, data understanding, data preparation, model building, evaluation and deployment.

As noted before, the general objective of the research was to segment customers based on their behavior, specifically, their calling behavior with a purpose of grouping them according to their value (long-term profitability) to the business.

6.2 *BUSINESS UNDERSTANDING*

Based on the evaluation made to understand the general background, processes employed and the business problem to be addressed, value (long run profitability) of a customer can be evaluated based three fundamental variables: total number of calls made, total duration used and the amount of revenue generated. Based on these variables, other related and derived variables can be generated to see whether these variables may yield better data mining result.

Hence, high value customers can be defined as those with relative lower total number of calls and lower amount of duration but with higher revenue. Consequently, low value customers are those with relative higher number of calls, and higher amount of

duration but with lower revenue. There are intermediary ones (variants) between these kinds of customers. For example, customers with high number of calls, high amount of duration and with high revenue are also high value customers next to the one mentioned above as they create more traffic congestion relatively. On the other hand, customers with lower number of calls, lower amount of duration and revenue are lower value customers but better than the low value customers mentioned above for the same reason of traffic congestion. Possible groups between these two can be categorized as medium value customers or could be categorized further if needed. The fact that the corporation suffers from high call traffic congestion during peak hours is one major reason to evaluate the value of customers.

Other attributes showing location and time of calls may provide hidden patterns that could be useful in the formation of meaning full segments.

6.2.1 Data Mining Tool Selection

Finding free data mining software with strong capabilities that performs the required tasks of data mining i.e. clustering and classification, was one of the challenging tasks the researcher faced. This was so because of scarcity of funds for such purpose and the high cost of data mining tools even for research purposes.

The researcher had to first set criteria for tool selection and search for the same. The criteria used to select one tool from the other were the following:

- The performance of the tool in terms of speed and quality.
- The time allowed to use the tool.

- The application area the tool proved its performance (Telecommunications, Customer Segmentation).
- Application of the tool for the selected data mining tasks (clustering and classification)
- The clustering and classification algorithms supported (K-means or SOM, and decision trees)
- Compatibility of the tool to the operating system available at hand (MS windows).
- The input file format the tool supports (MS-Excel, MS-Access, and MS-Word)
- The number of records (rows) and attributes (column) the tool can handle.
- User friendliness.

After considerable search and contact with tool providers, and review of tools available, three data mining packages were obtained freely which were WEKA version 3.2.3, Knowledge Studio Version 4.1.1 and Ghost Miner version 2.0. After detail evaluation of the tools as per the criteria, Knowledge Studio Version 4.1.1 was selected with good score. GhostMiner version 2.1.1 wasn't selected among other things primarily because of the time allowed to use was only 30 days and WEKA Version 3.2.3 wasn't selected as it was relatively less user friendly. The evaluation was conducted by giving weights to the above parameters and then, ranking the total scores of the tools accordingly.

The effort to obtain Clementine and Intelligent Miner data mining tools, which are very popular for such kind of data mining task, was failed because of financial limitation and poor internet connection respectively.

The researcher was satisfied by the capabilities of the selected software as its overall performance was very good during the course of the data mining process.

6.3 DATA UNDERSTANDING

Next to understanding of the problem to be addressed clearly and selection of an appropriate tool, the succeeding task is to analyze and understand the content and structure of the data available. To fulfill this requirement, raw data was initially collected regarding customer behavior from the Call Detail Record (CDR) and careful analysis of the data and its structure is done together with the business (domain) experts by evaluating the relationships of the data with the problem at hand and the particular data mining tasks to be performed. The following sections describe collected the nature of the data and its structure.

6.3.1 Initial Data Collection

As indicated above, the major data source was the CDR. The CDR is a rich data source since for every call made by a customer history of the call in respect to many variables like date, time, duration and destination is recorded in detail. This monthly transaction data was found to have a very big size. This record contains other numerous attributes and comprises of low level data that needs considerable

processing. The collected data pertaining to the data mining task to be performed looks like the following.

Caller	Callee	Call detail
009253657	00253738	040301121741000158
009253657	0002110074	040301131851000030
009253657	00252779	040301144023000118
009253657	00253738	040301162016000113
009253657	002519253140	040301170948000147
009253657	002519253140	040301171246000114
009253657	002519253140	040301171443000101
.....	.....	.....

Table 1 : Raw data taken from the CDR

Hence, the relevant attributes to be analyzed for the research were caller telephone number, the callee telephone number, and the call detail column. The last column contains date of the call, the time the call made/started, and the amount of duration the call stayed together.

The relevance of the attributes is checked with domain experts that make strategic marketing decisions.

To make it manageable and perform sampling, the data needed to be divided into smaller size. Further, this text data required transformation into an excel format to allow some processing and be fed to another software that could further process it to a format ready for data mining. Generation of additional derived attributes was also required.

6.3.2 Description of the Data Collected

The data collected for analysis, as indicated above, was of one month transaction data (March, 2004).It had a size of about 900 MB containing call detail records of all

mobile telephone users (prepaid as well as post paid).The total number of customers was about 100,000 and the number if records was about 15,000,000.The number of relevant columns was only three.

Description of each column follows.

Caller telephone number

This refers to the ID of the customer. It serves to distinguish each customer. All numbers are eight digit numbers and start with '09' to show that the telephone is a mobile one. The next six digits fall between the ranges of 200000 – 250000 for post paid subscribers.

Callee telephone number

The callee telephone number contains information regarding the destination of call i.e. whether the call is a local one or international. The local call can also be analyzed into calls made to fixed or mobile telephones. Calls made to fixed telephone can be further sub divided by Regions.

Call detail

This column contains three attributes together. Each six digit refers to a single attribute. The first one, from the left, indicates date of the call allocating the first two digits for year (YY), the next two digits for month (MM) and the last two digits for day (DD). The second six digit refers to the time the call was made/started similarly allocating two digits for hour (HH), minute (MM) and second (SS) respectively. The last six digits refer to the amount of time the call stayed by allocating two digits, for hour (HH), minute (MM) and second (SS) likewise.

The date of the call can be used to determine whether the call is made during weekdays or weekends. It could be useful to determine whether the call is made at the beginning, middle or end of the month. The time of the call (HH/MM/SS) can be extended to know whether the call is made during the day or night. The last attribute, the amount of duration the call stayed (HH/MM/SS) is useful to compute the amount of duration.

6.3.3 Data Exploration

The collected data provide information indirectly. The total frequency of the caller telephone number along records show the total number of calls made in that month. This figure is one of the basic attribute values that can represent a given customer. When this total number of calls is further divided by call location (local or international) and callee telephone type (fixed or mobile), it further describes well the calling behavior of a customer.

The amount of duration of each call can be summed up to give the total amount of duration a customer used. This figure as well is one of the basic attribute values to represent a customer. The total duration can be also divided as done to total number of calls by call location and callee telephone type to reveal the behavioral calling pattern of a customer in terms of time.

The other basic variable to represent a given customer is the total revenue generated from each customer. This figure can be calculated based on the pricing scheme in place which is based on many factors like the amount of duration, destination of call, time of the day and time of the week.

Other derived attributes which can be relevant for the data mining task can be derived from the above attributes. For example, average duration per call (total duration/total number of calls), average revenue per call (total revenue/total number of calls) and average revenue per duration (total revenue/total duration) are the important ones in respect to the data mining objective at hand.

6.3.4 Data Quality Verification

The collected data contains missing, incomplete and irrelevant data. In many of these cases, data in the call detail record column is missing or incomplete. Hence, a call detail column less than 18 digit should be removed.

The reliability of data and completeness of records are relatively good as the data is produced electronically.

6.4 DATA PREPARATION

Data preparation involves a series of steps to provide the final data set for modeling. It includes data selection, cleaning, construction, integration, and formatting.

6.4.1 Data Selection

The list of attributes selected for the data preparation process, as noted earlier, includes caller telephone number, callee telephone number, and call detail records columns. Another column (category) from customer data base was also important and integrated as it may contain valuable information.

The number of rows (records) selected was based on stratified sampling. From the total of around 50,000 mobile post paid customers, a sample of 11,117 was taken.

This was taken from each ten thousand range of numbers proportionally. This is done because of the domain experts' opinion that each ten thousand group of customers show different characteristics to one another as a group. The size of the sample is relatively high. This was done because of the fact that the data mining task to be performed (clustering) needs relatively higher number of records as the purpose is to explore and generalize about the population. This task would have been better if it was done on the entire population but this was constrained by shortage of preprocessing time. Hence the researcher was convinced based on the domain experts' opinion to take as much samples as possible proportionally from each group.

To select the data, the researcher used splitting software (The Splitter) to divide the whole data (900 MB) into 30 parts each with 30 MB size. Then, data of 30 MB each was taken randomly from each range of numbers proportionately.

6.4.2 Data Cleaning

The data was cleaned by removing the records that had incomplete (invalid) data and/or missing values under each column. Removing of such records was done as the records with this nature are few and their removal does not affect the entire data set.

To clean the data, first the selected 30 MB sized data was further divided into a size of 3MB, a size MS-Excel can handle at a time. Then, all sample data that had been changed into MS-Excel and was entered to and cleaned by Visual Fox Pro data base management system.

6.4.3 Data Construction

Almost all attributes were derived from the Call Detail Record (CDR).This step of data preparation took the researcher considerable time and effort. This kind of construction had to be done as the CDR contains raw data, which is not in the way appropriate for the business goal to be addressed, and the corresponding data mining tasks to be performed.

The attributes were derived from the CDR for each customer (telephone number) using MS-Excel and Visual Fox Pro data base management system. These attributes are listed in the final data set developed after the data preparation phase.

6.4.4 Data Integration

Only one attribute, customers' category, is integrated from the customer data base as other relevant attributes like age, position and sex are not maintained in the data base.

6.4.5 Data Formatting

This step involves changing of the data into a format suitable for the data mining tool (algorithm). The tool (algorithm) selected doesn't require preliminary formatting except for the outliers in the data. Outliers, which may mislead the k-means algorithm, are replaced with a maximum value that is set together with domain expert for each attribute. The final processed data set including the attributes and their description after the data preparation phase looks as follows.

Field Name	Data Type	Description
Caller	Number	The telephone number that identify each customer.
Category	Number	The classification code given to different type of customers
TNumber	Number	The number of calls a customer made per month.
NFixedAA	Number	The number of calls made to Addis Ababa
NfixedRegions	Number	The number of calls made to Regional States(in Ethiopia)
NInternational	Number	The number of calls made to foreign countries.
NMobile	Number	The number of calls made to Mobile phones
NDay	Number	The number of calls made during the day
NNight	Number	The number of calls during the night
NWeekday	Number	The number of calls made from Monday to Saturday
NWeekend	Number	The number of calls made on Sunday
NBOM	Number	The number of calls made at the beginning of the month
NMOM	Number	The number of calls made at the middle of the month
NEOM	Number	The number of calls made at the end of the month
TDuration	Number	The total amount of duration a customer used.
DFixedAA	Number	The duration of calls made to fixed telephone in Addis Ababa
DFixedRegions	Number	The duration of calls made to fixed telephone in Regional States(in Ethiopia)
DInternational	Number	The duration of calls made to foreign countries.
DMobile	Number	The duration of calls made to Mobile phones
DDay	Number	The duration of calls made during the day
DNight	Number	The duration of calls during the night

DWeekday	Number	The duration of calls made from Monday to Saturday
DWeekend	Number	The duration of calls made on Sunday
DBOM	Number	The duration of calls made at the beginning of the month
DMOM	Number	The duration of calls made at the middle of the month
DEOM	Number	The duration of calls made at the end of the month
TRevenue	Number	The total revenue generated from a customer
DPerCall	Number	Average duration per call
RPerCall	Number	Average revenue per call
RPerDuration	Number	Average revenue per duration

Table 2 : List of all attributes with their type and description

6.5 MODELING

6.5.1 Selection of Modeling Technique

The data mining techniques selected for customer segmentation are automatic cluster detection and decision trees. The selection was made because of the fact that these techniques are widely applied for segmentation problems. Moreover, the techniques are implemented well in the selected data mining tool and hence there was an opportunity to choose among related algorithms.

For clustering purpose, the available algorithms in the tool are K-Means and Expectation Maximization (EM). The K-Means algorithm is selected as it is a very good general-purpose clustering algorithm and is recommended for most situations. More over it is good in handling discrete and numeric attributes. However, the EM algorithm is recommended when there is a large amount of missing data (Angoss, 2002).

Regarding decision trees, the algorithms found in the tool are KnowledgeSEEKER and

HeatSEEKER. The first one was selected as it is a powerful, flexible algorithm that is especially good for exploration purposes. It can handle a large amount of variables with either a continuous or discrete dependent variable. However, the later works for few attributes and does not have the strengths the former offers. There are four classification measures used by the algorithms to build the decision tree namely: Unadjusted - Raw P-value Measure, Adjusted - P-value Measure, Entropy Variance - Non P-value, and Gini Variance - Non P-value Measure. The second one was selected because of the fact that it is the most statistically sound approach since it reduces the possibility of producing chance results and puts all variables on a common statistical footing. Finally, from the available pruning methods (Error-Complexity Pruning, Critical Value Pruning and Reduced-Error Pruning), the last one is chosen since it performs well on relatively higher test data and to trees with a discrete dependent variable (Angoss, 2002; Esposito & Semeraro, 1997).

6.5.2. Test Design

Before starting modeling, a plan should be first set to guide the training, testing and evaluating process and put a test for the quality and validity of the model.

All of the sample size (100%) was used to train the clustering model whereas for the decision tree model, 60% was used for training the model, 30% served as a test data and the remaining 10% were employed as a validation set.

The process of segmentation and the interpretation of resulting segments were examined together with domain experts so that the final output provides good basis

for possible design and implementation of an appropriate CRM strategies and programs.

6.5.3. Model Building

The model building phase is divided into three subsections namely attribute selection models, automatic cluster detection models and decision tree models. The first section involves the identification of best attributes to be used for the next modeling, automatic cluster detection. The later involves automatic formation of clusters using selected attributes. Brief analysis and interpretation of clusters is also made in this section. This section ends by choosing the best classifiers for the decision tree model. The last subsection deals with building of decision tree and develops rules by taking the final clusters as dependent variables.

Attribute Selection

The construction of both clustering and classification models is very important so that best attributes are selected from the resulting decision tree built.

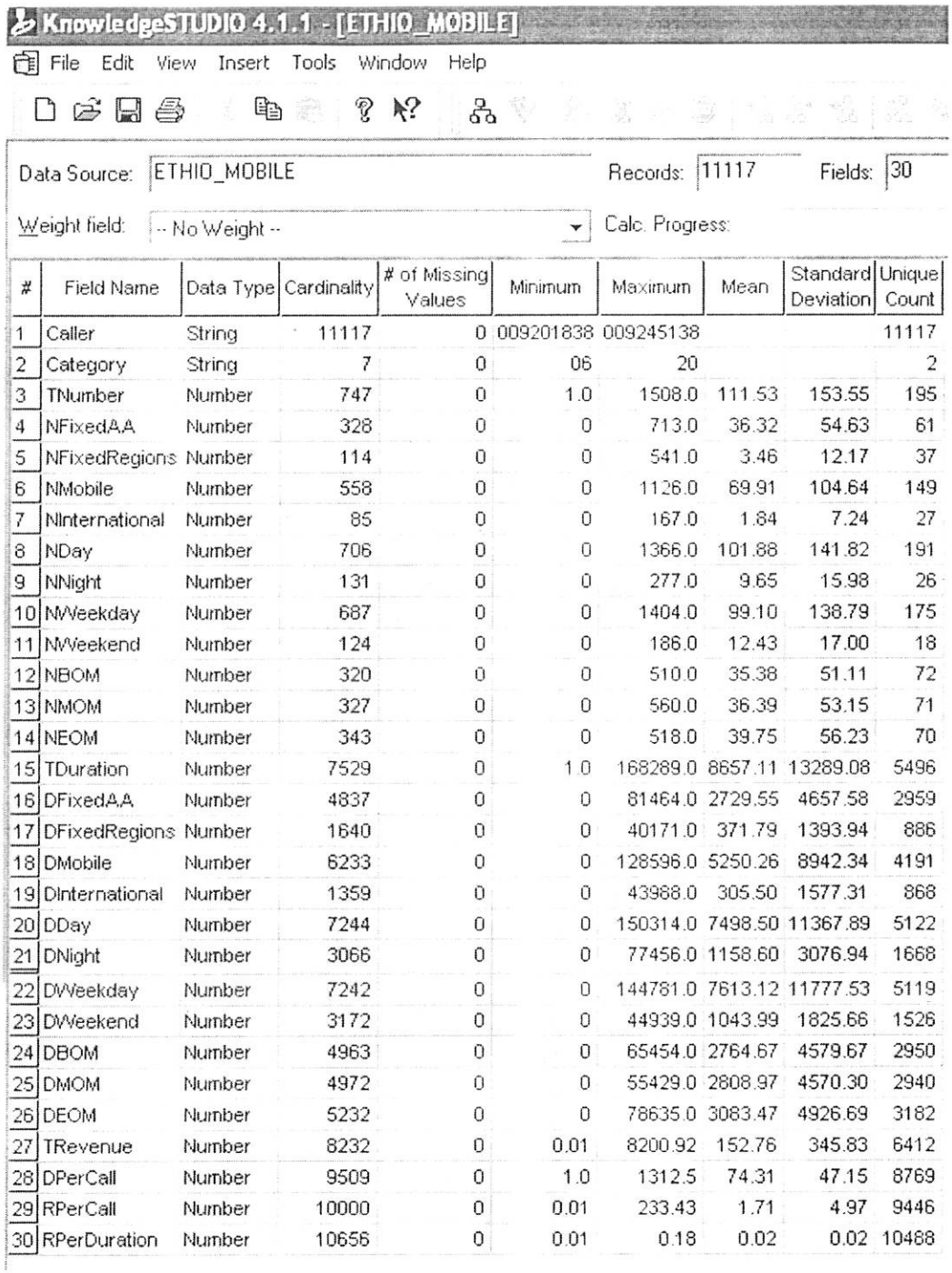
Six different models (three automatic cluster detection and three decision tree models) were built at different values of 'k' to distinguish those attributes which have higher information content so that the next clustering will be made based on these attributes that would result a better model to understand and easier segments to interpret. The attributes used for this clustering includes all those, except caller telephone number, listed in the data set created after completion of the data preparation phase.

The basic parameters that can take different values in the automatic cluster detection models are the number of clusters to be created (k) and the number of iterations

required (i). The number of customer segments (k), which can range from 4 to 10 in most of the time, is also dependent on the capacity of the firm to properly manage different clusters. After discussion with domain experts, the value of k was set to be from four to six. The other parameter, the number of iteration (i) refers to the maximum number of times the algorithm reads the data to form clusters. The number runs from 10-10,000 and as it increases, the longer the algorithm runs and the more accurate the result will be. The algorithm will stop running if it reaches this limit.

As the purpose of this experiment involves in the identification of best attributes, all attributes are used for the initial model building at the same level of maximum value of i (10,000), but different values of k (4, 5 and 6).

The attributes together with their details as calculated and displayed in the Knowledge Studio are presented as follows.


KnowledgeSTUDIO 4.1.1 - [ETHIO_MOBILE]

File Edit View Insert Tools Window Help

Data Source: ETHIO_MOBILE Records: 11117 Fields: 30

Weight field: -- No Weight -- Calc. Progress:

#	Field Name	Data Type	Cardinality	# of Missing Values	Minimum	Maximum	Mean	Standard Deviation	Unique Count
1	Caller	String	11117	0	009201838	009245138			11117
2	Category	String	7	0	06	20			2
3	TNumber	Number	747	0	1.0	1508.0	111.53	153.55	195
4	NFixedAA	Number	328	0	0	713.0	36.32	54.63	61
5	NFixedRegions	Number	114	0	0	541.0	3.46	12.17	37
6	NMobile	Number	558	0	0	1126.0	69.91	104.64	149
7	NInternational	Number	85	0	0	167.0	1.84	7.24	27
8	NDay	Number	706	0	0	1366.0	101.88	141.82	191
9	NNight	Number	131	0	0	277.0	9.65	15.98	26
10	NWeekday	Number	687	0	0	1404.0	99.10	138.79	175
11	NWeekend	Number	124	0	0	186.0	12.43	17.00	18
12	NBOM	Number	320	0	0	510.0	35.38	51.11	72
13	NMOM	Number	327	0	0	560.0	36.39	53.15	71
14	NEOM	Number	343	0	0	518.0	39.75	56.23	70
15	TDuration	Number	7529	0	1.0	168289.0	8657.11	13289.08	5496
16	DFixedAA	Number	4837	0	0	81464.0	2729.55	4657.58	2959
17	DFixedRegions	Number	1640	0	0	40171.0	371.79	1393.94	886
18	DMobile	Number	6233	0	0	128596.0	5250.26	8942.34	4191
19	DInternational	Number	1359	0	0	43988.0	305.50	1577.31	868
20	DDay	Number	7244	0	0	150314.0	7498.50	11367.89	5122
21	DNight	Number	3066	0	0	77456.0	1158.60	3076.94	1668
22	DWeekday	Number	7242	0	0	144781.0	7613.12	11777.53	5119
23	DWeekend	Number	3172	0	0	44939.0	1043.99	1825.66	1526
24	DBOM	Number	4963	0	0	65454.0	2764.67	4579.67	2950
25	DMOM	Number	4972	0	0	55429.0	2808.97	4570.30	2940
26	DEOM	Number	5232	0	0	78635.0	3083.47	4926.69	3182
27	TRevenue	Number	8232	0	0.01	8200.92	152.76	345.83	6412
28	DPerCall	Number	9509	0	1.0	1312.5	74.31	47.15	8769
29	RPerCall	Number	10000	0	0.01	233.43	1.71	4.97	9446
30	RPerDuration	Number	10656	0	0.01	0.18	0.02	0.02	10488

Table 3 : List of all attributes with their calculated fields

Here follows an illustration of the important steps performed to build both models for each value of *k*, by taking the case where k equals to 5.

The first two reports show attributes used for and the corresponding training process of the automatic cluster detection.

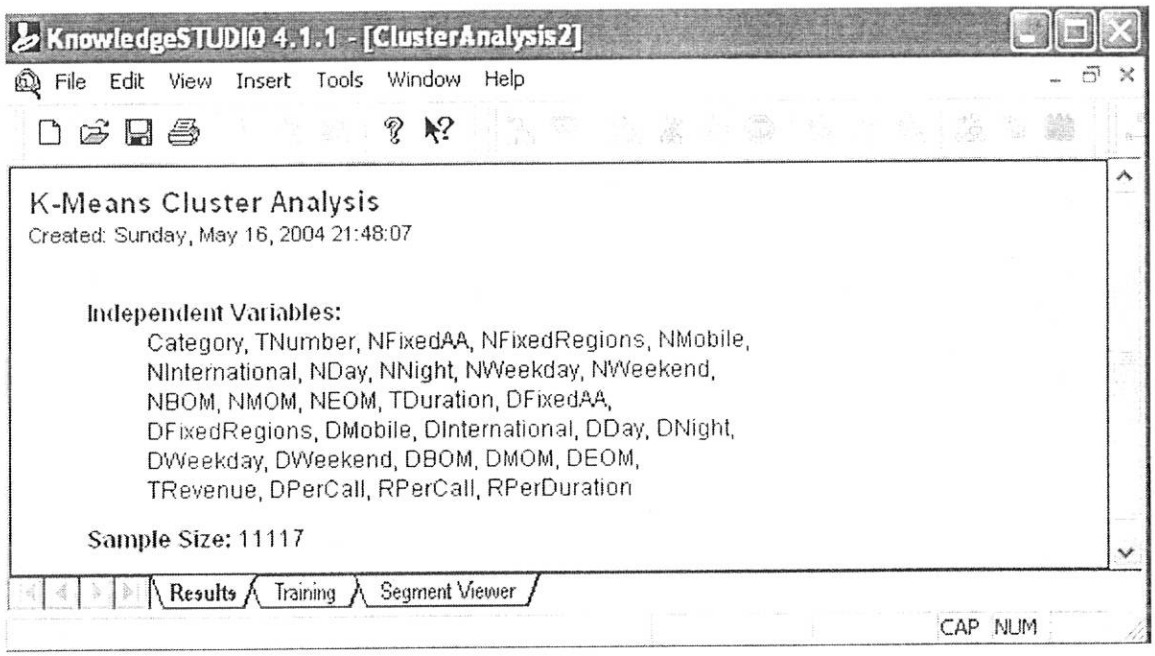


Figure 4 : Variables used in the first automatic cluster detection

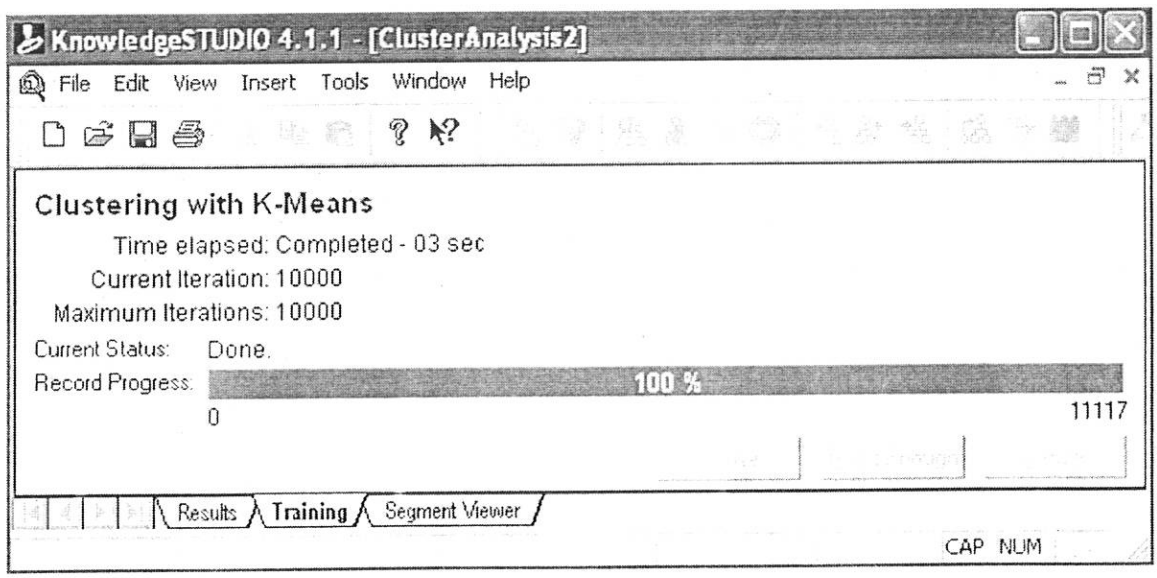


Figure 5 : The training process where k=5 and i=10000

After the training of clustering process had completed, segments were created. The following report shows the resulting segments and their respective sizes from the above clustering experiment.

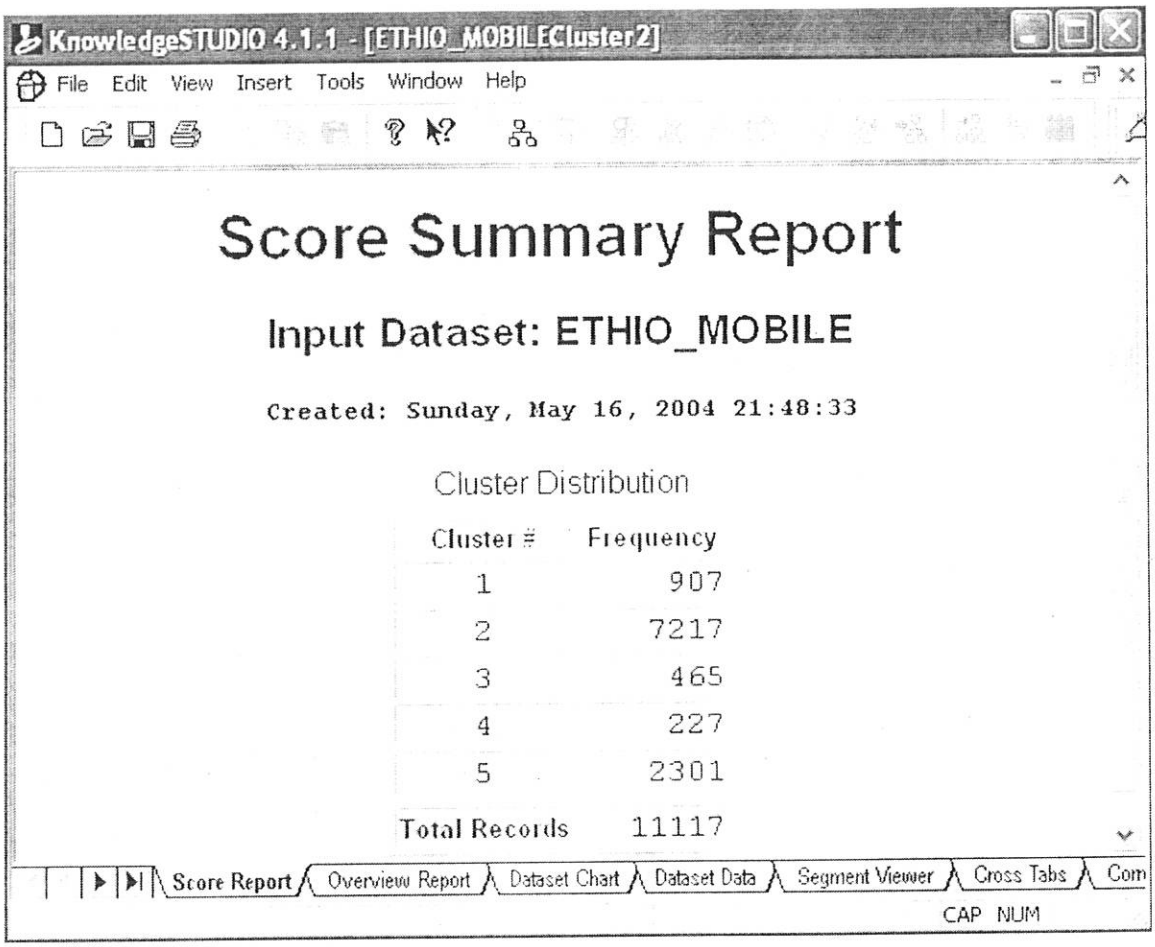


Table 4 : Cluster distribution where k=5

The next step after each record falls into one of the segments is to construct a decision tree model to see which attributes best serve to classify records into clusters by setting the cluster index as dependent variable. The following view shows part of the decision tree constructed to identify the important attributes for the next cluster run.

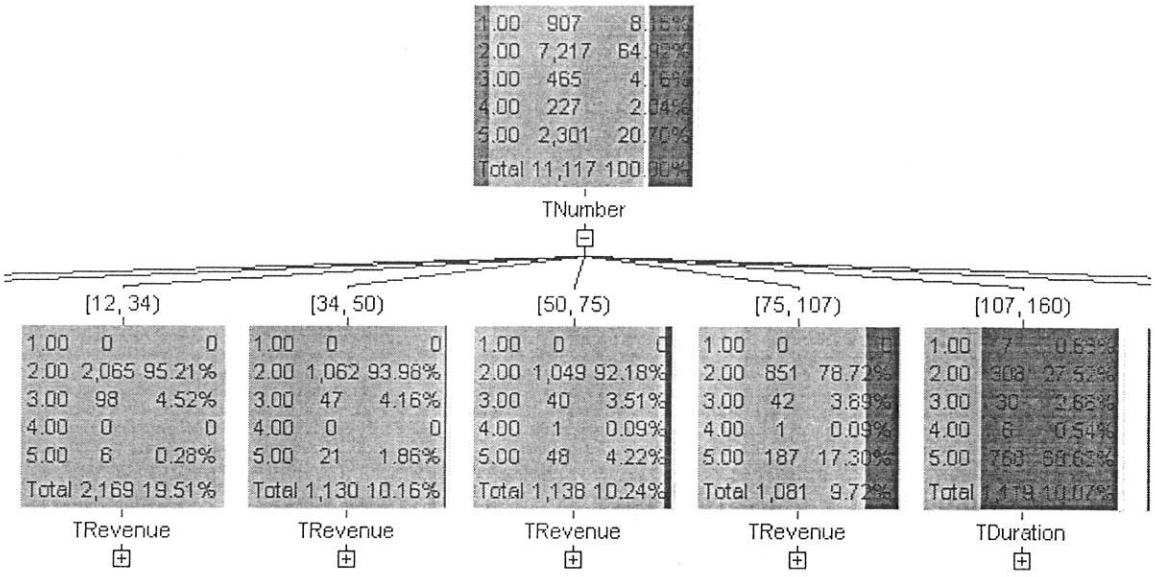


Figure 6 : Partial view of the decision tree built where k=5, and i=10000

Similar models of automatic cluster detection and decision tree models were also developed for other two values of *k* (four and six). After comparing the three decision trees built for three values of *k*, the relevant attributes were selected. The selection is made considering the relative positions of attributes, the amount of records classified at a time and the frequency the attribute appeared in top levels of the trees.

The attributes selected for the next step of automatic cluster detection were: *total number of calls made (TNumber), total amount of duration used (TDuration), total amount of international calls duration (DInternational), total amount of revenue generated (TRevenue) and average revenue generated from a call (RPerCall).*

Automatic Cluster Detection

This step is meant to yield the final segmentation plan that serve to develop the decision tree model.

This clustering model was done using the selected attributes where the values for k were four, five and six. Hence, a total of nine automatic cluster models were built at three different values of *k* (four, five and six) as wells as *i (1000, 5,000, &10,000).*

Here follows the steps involved in the automatic cluster detection, similar to the one in the first clustering experiment, when the value of *k* was five but *i* was five thousand.

The first view shows the attributes used in the clustering.

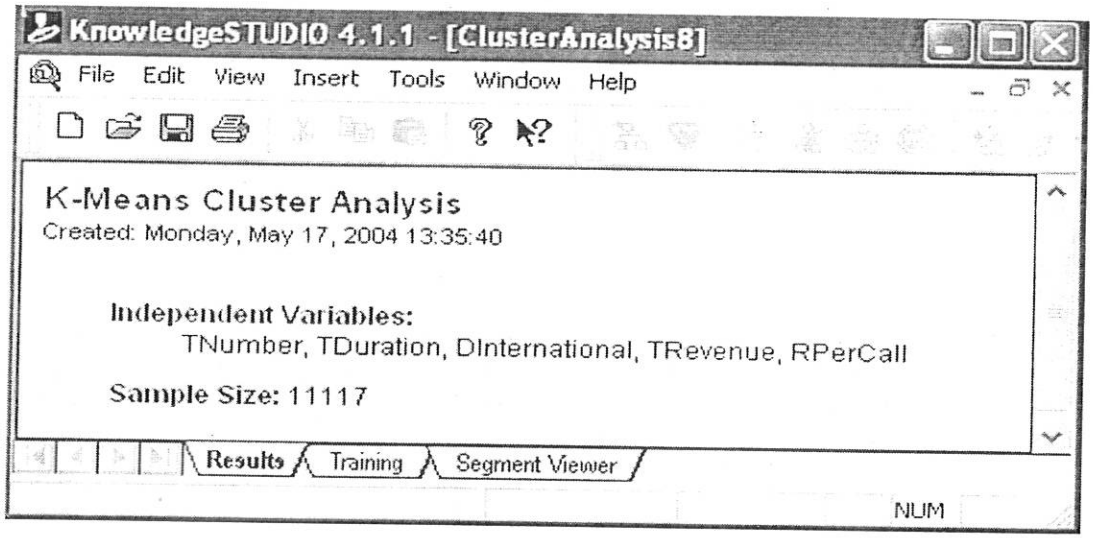


Figure 7 : Final attributes used for automatic cluster detection where k=5

The following view shows the training process.

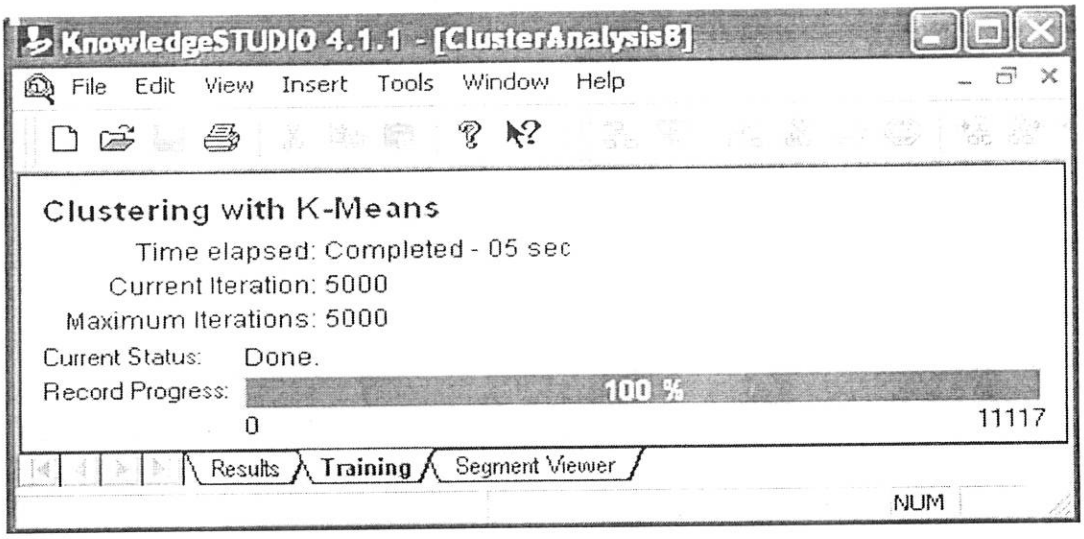


Figure 8 : The training process of automatic cluster detection where k=5 and i=5000

The final one indicates the resulting segments from this particular automatic cluster detection.

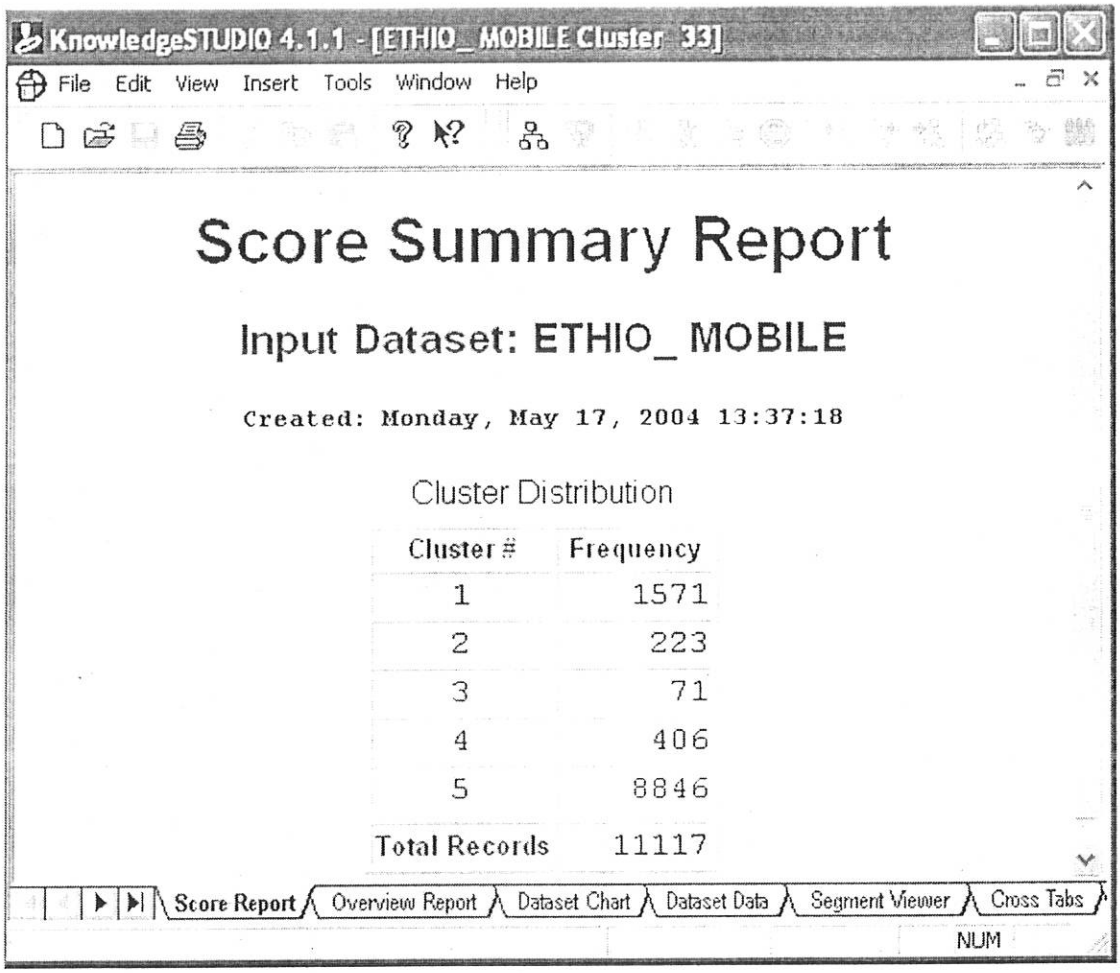


Table 5 : Cluster distribution where k=5 and i=5,000

Assessment of the automatic cluster detection model

The nine different models were analyzed. The different values of *i* did not bring any change in the distribution of the segments. Moreover, the average values of the corresponding clusters are the same. Hence, there are only three models to compare at the three different values of *k* that is four, five and six.

The comparison of these models was made in such a way that the averages of attributes of each cluster in a model are compared in respect to the other model. Such comparison is made after mapping the particular average into five discrete values as very low, low, medium, high and very high. The setting of the range was made

together with domain experts and considering the economic situation of the environment.

Here follows the analysis tables for the three models. The following abbreviations are used for convenience.

ATN- Average Total Number of calls

ATD-Average Total Duration used

ADI-Average Duration of International calls

ATR-Average Total Revenue

ARPC-Average Revenue Per Call.

First Segmentation (k=4)							
Cluster #	**Size**	**Share**	**ATN**	**ATD**	**ADI**	**ATR**	**ARPC**
1	1398	13%	394.77	31715.17	344.18	442.00	1.18
2	9158	82%	62.47	4348.91	50.53	59.51	1.00
3	292	3%	26.87	3174.25	1554.82	298.2	13.84
4	269	2%	272.44	29707.38	5146.07	1217.13	9.51
Total	**11117**	**100%**					
Cluster #	**Size**	**Share**					
1	1398	13%	HIGH	HIGH	VERY LOW	LOW	VERY LOW
2	9158	82%	LOW	VERY LOW	VERY LOW	VERY LOW	VERY LOW
3	292	3%	VERY LOW	VERY LOW	LOW	LOW	HIGH
4	269	2%	MEDIUM	MEDIUM	VERY HIGH	VERY HIGH	MEDIUM
Total	**11117**	**100%**					

Table 6 : Cluster description based on average values of attributes for k=4

The above table contains the average values of the attributes for each segment/cluster formed in the model where the number of segments (k) was set to four. The lower part of the table shows the relative values of attribute averages compared among segments. As indicated above, these five discrete values to each attribute was allocated based on the mapping of the value of each attribute against the discrete values to allow comparison.

Cluster	Description	Possible rank
1	High number of call, high amount of duration , and low amount of revenue	3rd
2	Low number of call, very low amount of duration , and very low amount of revenue	4th
3	Very low number of call, very low amount of duration, and low amount of revenue	2nd
4	Medium number of call, medium amount of duration, and very high amount of revenue	1st

Table 7 : Cluster summary and corresponding ranks based on basic attributes for k=4

Based on the above summary, it is possible to give rank (the third column) to each segment as the values of the attributes provide information as to the degree of profitability (value) of each segment.

For instance, cluster four is composed of high value customers since high amount of revenue is generated with relatively medium traffic congestion. Traffic congestion increases as the number of calls and amount of duration are high. On the other hand cluster two is composed of low value customers as the revenue generated is very low. Cluster one is also composed of low value customers but better than cluster two. This is so because of the fact that it yields better revenue though it causes high traffic congestion. Cluster three is also composed of customers better than the cluster one and two since the traffic congestion is relatively low. But still, it is not composed of profitable customers.

This segmentation (where k=4) didn't provide sufficient information as it can't separate medium customers from low value customers.

Second Segmentation (k=5)							
Cluster #	**Size**	**share**	**ATN**	**ATD**	**ADI**	**ATR**	**ARPC**
1	618	6%	513.11	43074.68	567.85	628.00	1.27
2	8263	74%	43.16	2870.47	33.44	39.59	0.98
3	293	3%	22.69	2821.48	1461.57	278.55	14.27
4	244	2%	242.66	27290.92	5341.00	1216.73	10.45
5	1699	15%	223.23	16766.13	201.73	229.79	1.16
Total	11117	100%					
Cluster #	**Size**	**share**					
1	618	6%	VERY HIGH	VERY HIGH	VERY LOW	MEDIUM	VERY LOW
2	8263	74%	VERY LOW	VERY LOW	VERY LOW	VERY LOW	VERY LOW
3	293	3%	VERY LOW	VERY LOW	LOW	LOW	HIGH
4	244	2%	MEDIUM	MEDIUM	VERY HIGH	VERY HIGH	HIGH
5	1699	15%	MEDIUM	LOW	VERY LOW	LOW	VERY LOW
Total	11117	100%					

Table 8 : Cluster description based on average values of attributes for k=5

The above table is similar in nature to the previous table (table 6) except that the number of clusters is five. It is also summarized in the following table.

Cluster	Description	Possible rank
1	Very high number of call, very high amount of duration , and medium amount of revenue	2nd
2.	Very low number of call, very low amount of duration , and very low amount of revenue	5th
3	Very low number of call, very low amount of duration , and low amount of revenue	3rd
4	Medium number of call, medium amount of duration , and very high of revenue	1st
5	Medium number of call, low amount of duration , and low amount of revenue	4th

Table 9 : cluster summary and corresponding ranks based on basic attributes for k=5

This segmentation, where the value of k is five, is better than the previous one as it carries more information.

Cluster four contains high value customers as the revenue generated is very high as well as the traffic congestion caused is medium. Cluster one is next as revenue generated is medium. Cluster three is the third since there is low traffic congestion than cluster five with which it has same revenue level. Cluster two is the last as it generates very low revenue.

Third Segmentation(K=6)							
Cluster #	**Size**	**share**	**ATN**	**ATD**	**ADI**	**ATR**	**ARPC**
1	1707	15%	216.27	16246.59	192.29	222.58	1.15
2	8196	74%	42.24	2795.86	31.13	38.31	0.97
3	289	3%	24.14	2799.32	1316.96	254.22	13.32
4	102	1%	85.44	14356.14	5424.34	1148.99	18.06
5	626	6%	505.72	41981.87	375.54	578.27	1.18
6	197	2%	390.49	38097.87	4510.08	1174.34	4.51
Total	**11117**	**100%**					
Cluster #	**Size**	**share**					
1	1707	15%	MEDIUM	LOW	VERY LOW	LOW	VERY LOW
2	8196	74%	VERY LOW	VERY LOW	VERY LOW	VERY LOW	VERY LOW
3	289	3%	VERY LOW	VERY LOW	LOW	LOW	HIGH
4	102	1%	LOW	LOW	VERY HIGH	VERY HIGH	VERY HIGH
5	626	6%	VERY HIGH	VERY HIGH	VERY LOW	MEDIUM	VERY LOW
6	197	2%	HIGH	HIGH	VERY HIGH	VERY HIGH	LOW
Total	**11117**	**100%**					

Table 10 : Cluster description based on average values of attributes for k=6

This table also has similar nature with table 6 and 8 above. The interpretation follows after the summary next page.

Cluster	Description	Possible rank
1	Medium number of call, low amount of duration, and low amount of revenue	5th
2	Very low number of call, very low amount of duration, and very low amount of revenue	6th
3	Very low number of call, Very low amount of duration, and low amount of revenue	4th
4	Low number of call, low amount of duration, and very high amount of revenue	1st
5	Very high number of call, very high amount of duration, and medium amount of revenue	3rd
6	High number of call, high amount of duration, and very high amount of revenue	2nd

Table 11 : Cluster summary and corresponding rank based on basic attributes fork=6

This segmentation divides the customers into relatively distinct groups with unique features. Here, cluster four contains the top high value customers as the revenue generated is very high and the traffic congestion caused is very low. Cluster six is next only because of higher traffic congestion than cluster four. Cluster five is third because it provides the only medium level revenue. Cluster three and one are the fourth and fifth respectively. This is so because of the fact that the first one causes lesser traffic congestion than the later. Cluster two is the last as it offers very low revenue level.

Choosing the best Segmentation plan

One of the major tasks in reaching to useful segmentation is the determination of the number of segments i.e. the value of k. To determine this value, the support of

domain experts is very important as the understanding of each segmentation needs business and marketing know how.

Based on the feedback from the domain experts and the above interpretation, segmentations where the values of k are five and six were good. Especially, in the later segmentation (k=6), more dissimilar clusters were formed.

Decision tree model building

The last step of the modeling phase is to build a decision tree model setting the cluster index as dependent variable for the final segmentation plan adopted, in this case, segmentation where the values of k is five and six.

The total records (11117) were divided into three partitions. The first partition contains 60% and was used for training. The next 30 % was used for testing and the remaining 10% was used for validation.

The model finally produced rules that enable to assign a new record to one of the clusters.

The attributes used for this purpose were those that had been identified as the best classifiers.

The following view show the attributes used, the accuracy of the model, the pruning status etc.

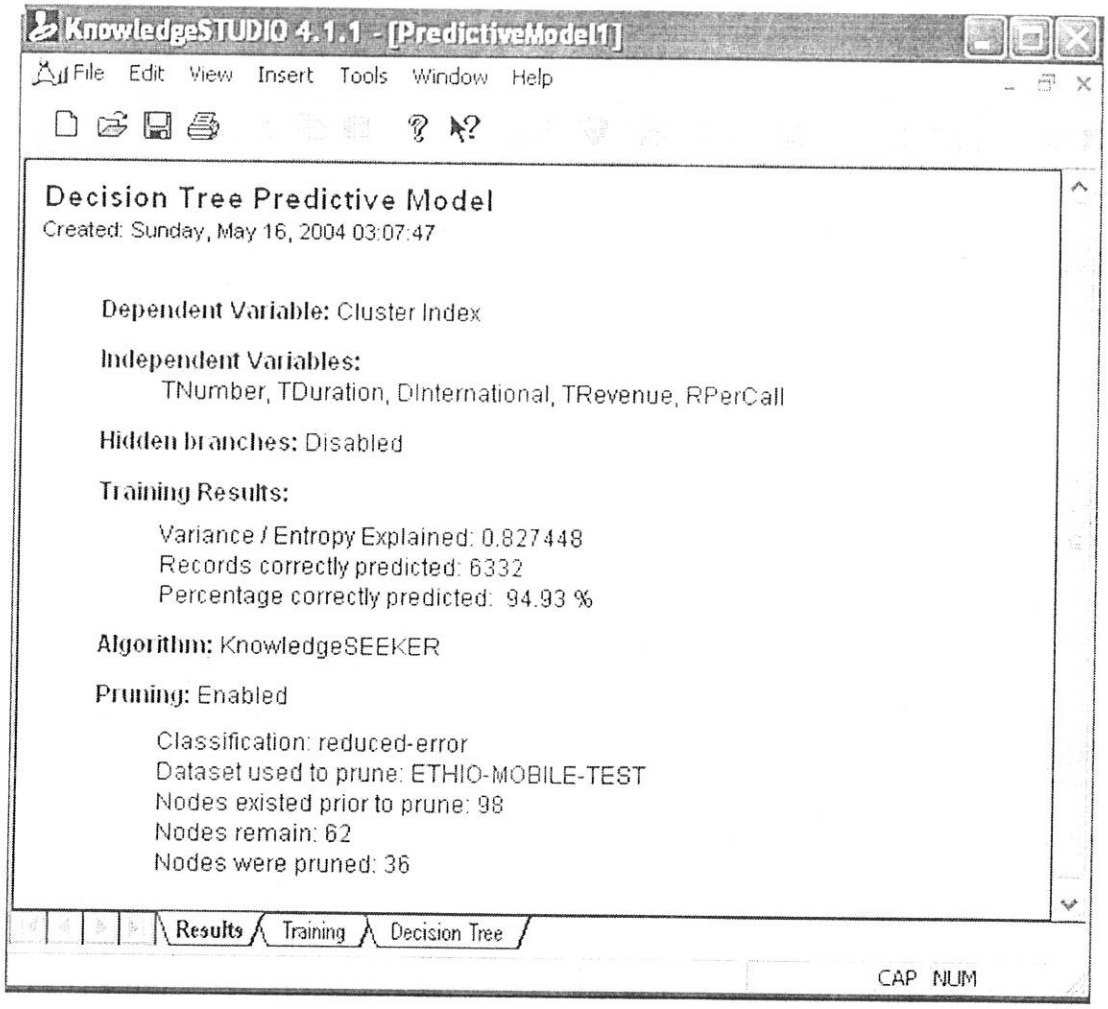


Figure 9 : Results of the decision tree built for k=6

The partial view of one of the trees built using the algorithm, measure of classification, and pruning method specified in tool selection is as follows

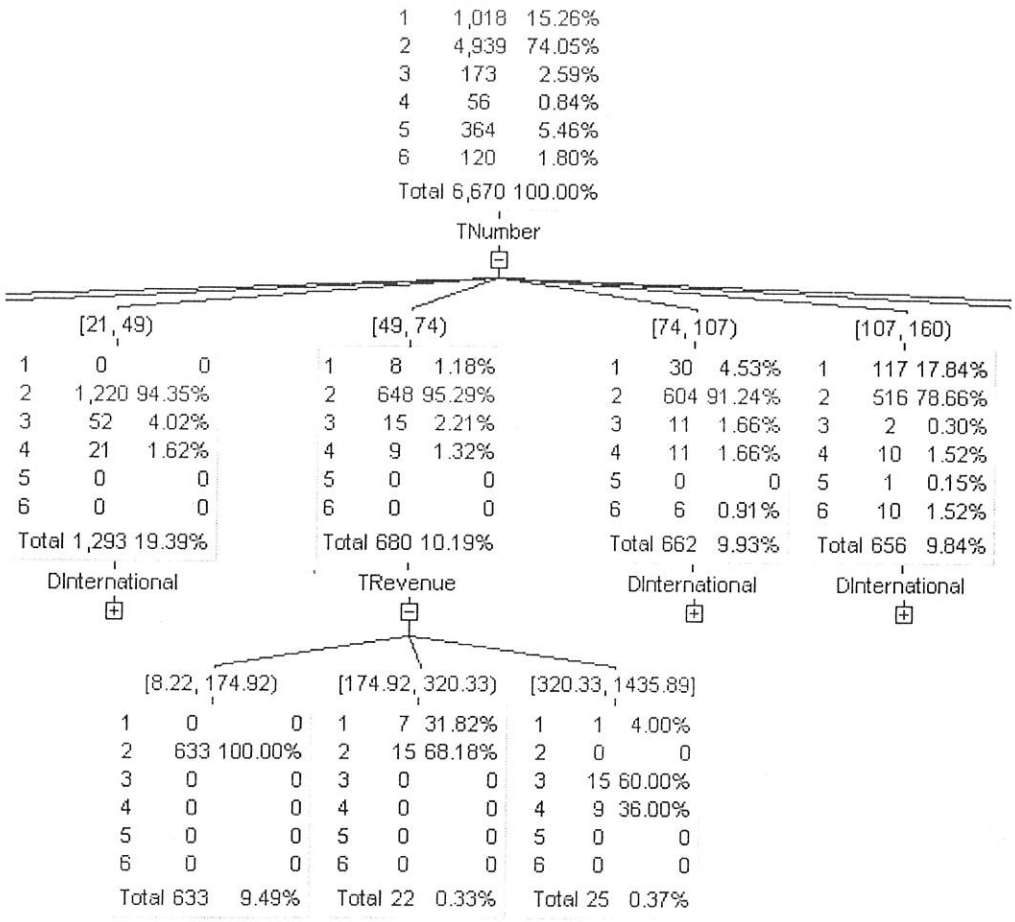


Figure 10 : Partial view of the decision tree for k=6

This partial view shows part of the decision tree built using the training set of 60 %(6670 out of 11117). It was also tested using the test data of 30% as indicated earlier. This decision tree model is the basis to generate the rules required to classify customers (new) to their respective segments (cluster indices). As can be seen, the tree starts with a root node containing 6670 customers and ends with leaf nodes with various numbers of classified customers with corresponding cluster indices (1-6) and respective accuracy level of prediction in that cluster.

Validation

This step is useful to evaluate the performance of the decision trees constructed. The accuracy level (94.93%) that appears in the above window is obtained when the model is validated with the training data set. However, most of the time, a model is evaluated by a data which is not used in training. In this case, the data set retained for independent test of the model is the validation dataset.

The validation set is a data which is not used in the process of the decision tree building. The confusion matrix reports based on the validation data for the two decision trees constructed were as follows.

Validation Summary Report

Input Dataset: ETHIO-MOBILE-VAL

Created: Sunday, May 16, 2004 03:08:08

Confusion Matrix - Cluster Index

		Predicted				
		1	2	3	4	5
Actual	1	62	0	0	7	2
	2	0	822	2	0	2
	3	0	7	16	6	1
	4	0	2	0	24	0
	5	13	16	0	0	130

Statistics

Total records	1112
Correctly predicted	1054
Percentage	94.78%
Valid records	1112

Figure 11 : Confusion matrix developed based on the validation dataset for k=5

The above confusion matrix evaluates the degree of accuracy of the decision tree where k=5 with a validation (unknown data to the model) data set which was kept previously for this purpose. It has a total of 1112 i.e. 10% of the total records (11117). The accuracy level 94.78 was computed when the number of correctly predicted customers (1054) is divided to the total customers predicted. It shows how well the model predicts customers for each cluster index.

Validation Summary Report

Input Dataset: ETHIO-MOBILE1-VAL

Created: Sunday, May 16, 2004 03:11:17

Confusion Matrix - Cluster Index

		Predicted					
		1	2	3	4	5	6
Actual	1	134	14	0	0	10	0
	2	6	811	2	0	0	0
	3	1	6	14	6	0	0
	4	0	0	3	13	0	1
	5	5	0	0	0	65	3
	6	0	0	0	1	0	17

Statistics

Total records	1112
Correctly predicted	1054
Percentage	94.78%
Valid records	1112

Figure 12 : Confusion matrix developed based on validation dataset for k=6

This matrix is similar to that of the previous one except it is generated for the decision tree where k=6.The interpretation is similar to the previous one.

From the generated two confusion matrices, it is possible to see that both predictive models result the same accuracy level (94.78). Moreover, for almost all clusters, the accuracy of prediction is fairly distributed for both cases which is a very good feature together with the accuracy level. With the discussion to the domain expert, the segmentation scheme where the value of k is six was selected as it contains more information (clusters) as the same time has equal level of prediction accuracy.

Rule generation

The rules applicable for the selected decision tree as generated by the tool were very detail and many. Sample of these rules is annexed as appendix 3.

6.6 EVALUATION

Once an optimum model is built, critical assessment of the same against the business goal to be achieved (business problem to be addressed) is very important.

The business goal to be met was to segment customers into meaning full groups so that an appropriate CRM strategies and programs can be designed and implemented. The basis of segmentation was according to customers' value to the business. And, customer value was defined based on three major attributes: total number of calls made, total amount of duration used and total revenue generated from the customer.

The data mining process having a task of clustering and classification was performed in order to achieve the business objectives set.

The results of the data mining process were encouraging and at least provide a way of reaching data mining solutions for market segmentation for the organization, if not yield a possible solution. Customers having similar calling patterns were grouped in

the same group whereas the groups formed were different from each other. This is the underlying criterion of segmentation (clustering). Moreover, the decision tree model provided a description of the segments and rules for assigning new records to segments.

The researcher believes that with further analysis of the results by marketing experts and IT specialists, the data mining process could be revisited to produce an optimal segmentation scheme so that relevant customer related information for informed decision making is acquired that help the organization manage its scarce resources efficiently, effectively and economically.

6.7 DEPLOYMENT OF RESULTS

Application of a segmentation scheme involves the consumption of considerable resources since people, business processes and technology together are integrated and directed based on the information obtained from the segmentation plan. Therefore, the results of this study can be deployed for marketing decision making after a thorough evaluation and integrating the necessary adjustments by group of domain experts.

CHAPTER SEVEN

CONCLUSION AND RECOMMENDATIONS

7.1 CONCLUSION

In general, the study focused in the application of data mining techniques in the area of CRM and more specifically for the purpose of customer centric market segmentation at Ethiopian Telecommunication Corporation. To this effect, related literature on data mining techniques, CRM and market segmentation was reviewed. In the experimentation part, the CRISP data mining process model was followed to complete the data mining task.

The business goal to be achieved was to group post paid mobile customers into similar groups based on their calling behavior during a given period. The criterion for segmentation was the degree of customers' value (long term profitability) to the business. This broad criterion was measured with three basic attributes and their derivatives. These basic attributes were total number of calls made, total amount of duration used and the amount total revenue generated. For this purpose, two data mining tasks (clustering and classification) were employed and a complete data mining process phases (business understanding, data understanding, data preparation, model building, evaluation and deployment) were completed.

From the result, it was possible to group customers according to their calling behavior and hence as per their long-term value to the business.

The results of the research were encouraging as the domain experts in the industry accepted it. The company can derive benefit through an appropriate utilization of the

research to improve its customer relationship management that is one of the hottest issues to be addressed in today's customer oriented, dynamic, and competitive market environment.

7.2 RECOMMENDATIONS

The researcher makes the following recommendations based on the findings of the study.

Need to build a data warehouse

This study was a victim of a very lengthy data preparation process that consumes considerable time, effort and other resources. It is very important to build a data warehouse not only for data mining purposes but also other important data analysis tasks as optimal decision making is only feasible if it is based on reliable and relevant information. Customers' data at various contact points should be collected and integrated as such data in these days becomes much valuable.

Undertaking further data mining researches

There is always a room for improvement in data mining as it is an interactive and iterative process. It needs refinement, update and enhancement as business problems always become diverse and complex. In these context, this research can be further refined through changing the techniques and algorithms, the data, and the modeling parameters used in the study. More specifically,

* The employment of the neural network for clustering and classification, which is very popular data mining technique, may yield better results.

Need for strong commitment for research and change

Application of customer relationship management and related technologies like data mining require continuous and flexible approach, and analysis of the dynamic nature of customers. Such researches should be initiated and conducted in house, and resulting outputs should be utilized to design, implement, and continually improve CRM strategies and programs at all levels of the organization.

REFERENCES

Ahola, J.& Rinta-Runsala, E., Data Mining Case Studies in Customer Profiling, LOUHI-project, 2001.

www.vtt.fi/tte/datamining/ publications/dm_case_studies.pdf

Angoss Software Corporation, Knowledge studio user manual.2003.

www.angoss.com

Basgoze,A.& Gokturk,M, Building Customer profiles using data Mining Techniques, Turkish symposium on Artificial Intelligence and Neural Networks, 2003.

Berry J.& Linoff,S., Mastering Data Mining, John Wiley & Sons Inc, 2000.

Bishop,M., Neural Networks for pattern recognition, Oxford University Press., 1998.

Bose, R., Customer Relationship Management: Key components for IT success, Industrial Management and Data System 102/2 , 89-97,2002.

www.emeraldinsight.com

Bounsaythip, C.& Rinta-Runsala,E., Overview of Data Mining for Customer behavior Modeling, LOUHI, 2001 .http://www.vtt.fi/tte/

Bull,C., Strategic Issues in Customer Relationship Management implementation, Christopher, Business process Management Journal, volume 9, No 5, 2003.

Connely M.T. & Begg E.C, Data Base Systems, A Practical approach to design, implement and management, 3rd Ed, 2000.

CRISP-DM., CRISP-DM 1.0: Step-by-step data mining guide. 2000. http://www.crisp-dm.org

DSS Research, Understanding Market Segmentation,2001.

http://www.dssresearch.com/liabrary/segment/understanding.asp

ETC Online Report.http://www.telecom.net.et

Esposito, F., Malerba,D.& Semeraro, G.,IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.19, No.5, May 1997.

http://www.di.uniba.it/~malerba/publications/PAMI97.pdf

Forcht,K.A & Cochran,K.,Using Data Mining and Data warehousing Techniques, Industrial Management and Data Systems,pp189-196,1999.

www.emeraldinsight.com

Gray,P.and Byun,J., Customer Relationship Management, March 2001 .

Han, F.and Kamber, M., Data Mining, Concepts and Techniques, Morgan Kaufmann publishers, academic press, 2001.

IBM Corporation, Enhance Your Business Applications: Simple Integration of Advanced Data Mining Functions Advanced data Mining Functions, 2002.

http://www.redbooks.ibm.com/redbooks/pdfs/sg246879.pdf

IBM Corporation, Mining Your Own Business in Telecoms Using DBA Intelligent Miner for data, 2001.

http://www.redbooks.ibm.com/redbooks/pdfs/sg246273.pdf

Kellen, V., Customer Relationship Management Measurement Frameworks, 2002.

Kim,J., Suh, E.& Hwang, H., A Model for Evaluating The Effectiveness of Customer Relationship Management Using the Balanced Scorecard, Journal of Interacting Marketing, Volume 17, No 2, 2003.

http://mis.postech.ac.kr/research/Full/A%20model%20for%20evaluating%20the%20effectiveness%20of%20CRM%20using%20the%20balanced%20scorecard.pdf

Kotler, P., Marketing Management: Analysis, Planning, Implementation and Control, 9th Edition, New Delhi, Prentice Hall of India, 1998.

McKinsey Marketing Solutions (MKMS), Tactical CRM: Three Steps to Mining Profits, Not Data.

http://www.mckinsey.com

Mckinsey Marketing Solutions, The New Era of Customer Loyalty Management.

http://www.mckinsey.com

Mckinsey Marketing Solutions, Unlock your Hidden Potential your Customer Relationship Management Investments.

http://www.mckinsey.com

Mitchell, M., Machine Learning, The MCGraw-Hill Companies, Inc, 1997.

Parvatlyar, A and Sheth,N.J ,Customer Relationship Management Emerging Practice, Process and Discipline.

Pritscher L.& Feyen, H., Data Mining and Strategic Marketing in the Airline Industry.

http://www.luc.ac.be/iteo/articles/pritscher1.pdf

Saarenvirta,G., Mining Customer data, A step by step look at a powerful clustering and segmentation methodology,1998.

http://www.db2mag.com/db_area/archives/1998/q3/98fsaar.html

Schiffaman, G.L and Kanuk, L.L, Consumer Behavior, Prentice Hall, Inc, 4th Edition, 1991.

Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery, 3rd Ed, 1999.

http://www.twocrows.com

Ulwich,T.& Jan A.Elsenhauer, A.J,The Natural Order of Segmentation: Aligning Company culture with its customers,

http://www.diwings.ch/e/druck/Market_Segmentation.pdf

Vriens,M., Market segmentation, Analytical developments and applications Guidelines, Technical Review series, March 2001.

http://www.intelliquest.com/resources/technical/MarketSegmentationOverview_MBIQ_June24.pdf

Witten, H.I and Frank, E, Data Mining, Practical machine Learning Tools and Techniques with JAVA implementation, 2000.

Xu, Y., Yen, C.D., Lin, B., & Chou, C.D. Adopting Customer Relationship Management Technology, Industrial Management and Data Systems 102/8 ,pp 442-552,2002.

www.emeraldinsight.com/

Yang Y.& Padmanabhan, B., Data Mining for customer Segmentation, A behavioral pattern-based Approach, Operations and Information Management department, University of Pennsylvania, Jan 2004.

Ziberna, A.& Zabkar, V. Application of End user Segmentation Using Statistical Methods, 2003.

http://mrvar.fdv.uni-lj.si/pub/mz/mz19/zabkar.pdf

APPEDICES

APPENDIX 1: CODES WRITTEN DURING THE DATA PREPARATION PHASE (Microsoft Visual FoxPro)

1. Cleaning missing and incomplete records

1.1. ***To delete records which are not 18 digits long and don't start with "0403" under the call records detail column***

```
close tables all

select 0
use sample exclusive
DELETE ALL FOR len(alltrim(sample.datetime)) <> 18 or left(alltrim(sample.datetime),4)<> "0403"
  pack
```

1.2. To bring the callee telephone number under the same format for the same types of telephones and locations.

```
close tables all
select 0
use sample
go top

do while !eof()
    do case
          case len(alltrim(sample.callee)) = 8
                replace sample.callee with "002519"+right(alltrim(sample.callee),6)
          case len(alltrim(sample.callee))  = 10
```

```
                replace sample.callee with "00251"+right(alltrim(sample.callee),7)
        case len(alltrim(sample.callee))  = 13
                if substr(alltrim(sample.callee),3,4) = "2510"
                        replace sample.callee with "00251"+substr(alltrim(sample.callee),6,7)
                endif
        case len(alltrim(sample.callee))  > 13
                replace sample.callee with right(alltrim(sample.callee),12)
    endcase

    select sample
    skip
enddo
```

1.3. To delete those irrelevant callee telephone numbers which doesn't represent real telephones

```
close tables all
select 0
use sample excl
DELETE ALL FOR len(alltrim(sample.callee)) < 12
PACK
```

2. Rearranging

```
close tables all
set multilocks on
select 0
use tele
cursorsetprop("buffering",5,"tele")

select 0
use sample
go top
do while !eof()
    select tele
    append blank
    replace tele.caller with sample.caller
    replace tele.callee with sample.callee
    replace tele.date with ctod(substr(alltrim(sample.datetime),5,2)+"/"+substr(alltrim(sample.datetime),3,2)+
        "/"+substr(alltrim(sample.datetime),1,2))
    replace tele.hh with val(substr(alltrim(sample.datetime),7,2))
    replace tele.mm with val(substr(alltrim(sample.datetime),9,2))
    replace tele.ss with val(substr(alltrim(sample.datetime),11,2))
    replace tele.duration1 with val(substr(alltrim(sample.datetime),13,6))
    =tableupdate(.t.,.t.)

    select sample
    skip
enddo

flush
```

3. To convert the duration(hh:mm:ss)into seconds

```
close tables all
SELECT 0
use tele
go top
do while !eof()
    select tele
    replace tele.duration with
        (VAL(SUBSTR(tele.duration1,1,2))*3600 +
        (VAL(SUBSTR(tele.duration1,3,2))*60 +
        (VAL(SUBSTR(tele.duration1,5,2))
    select tele
    SKIP
END DO
```

4. Main Pre Processing

```
*** ASSUMPTION : tele.dbf is sorted by Caller
    close tables all
    set multilocks on
    select 0
    SELECT 0
    use final
    cursorsetprop("buffering",5,"final")

    select 0
    use tele
```

```
** INITALIZATION
STORE 0.00 TO _TNumber, _NFixedAA,_NFixedRegions,_NMobile, _NInternational
STORE 0.00 TO _DFixedAA,_DFixedRegions,_DMobile, _DInternational
STORE 0.00 TO _NDay, _NNight, _DDay, _DNight
STORE 0.00 TO _NWeekday, _NWeekend, _DWeekday, _DWeekend
STORE 0.00 TO _NBOM,_NMOM,_NEOM,_DBOM,_DMOM,_DEOM
STORE 0.00 TO _TDuration, _TRevenue

select tele
set order to caller
 go top
 _caller = tele.caller

** Number of Calls
_allcalls = _allcalls + 1                          && Total Calls --> Local + International

if left(alltrim(tele.callee),5) == "00251"         && LOCAL calls
      DO CASE
            CASE substr(alltrim(tele.callee),6,1) == "1"          && Fixed Addis Ababa
                  _ NFixedAA = _ NFixedAA + 1
                  _DFixedAA = _ DFixedAA  + tele.duration
            CASE substr(alltrim(tele.callee),6,1) == "9"          && Mobile
                  _NMobile = _ NMobile + 1
                  _ DMobile = _ DMobile + tele.duration
            OTHERWISE && Fixed Region
                  _NFixedRegions = _ NFixedRegions + 1
                  _ DFixedRegions = _ DFixedRegions + tele.duration
      ENDCASE
 else                                              && International Calls
      _NInternational = _ NInternational + 1
```

```
        _ DInternational = _ DInternational + tele.duration
endif
_TDuration = _TDuration + tele.duration                    && Total duration --> Local + International

 if tele.hh >= 8 and tele.hh <= 20                         && Day time Calls & duration
        _NDay = _ NDay + 1
        _ DDay = _ DDay + tele.duration
else                                                       && Night Calls & duration
        _NNight = _ NNight + 1
        _ DNight = _ DNight + tele.duration
endif

if tele.date = {^2004-03-02} or cdow(tele.date) = "Sunday"  && Sunday & Holiday Calls & duration
        _NWeekend = _NWeekend + 1
        _DWeekend = _DWeekend + tele.duration
else                                                       && Week Days Calls & duration
        _NWeekday = _NWeekday + 1
        _DWeekday = _DWeekday + tele.duration
endif

do case
        case day(tele.date) <= 10                          && Begining of Month Calls and Duration
                _NBOM = _NBOM + 1
                _DBOM = _DBOM + tele.duration
        case day(tele.date) > 10 and day(tele.date) <= 20  && Mid of Month Calls and Duration
                _NMOM = _NMOM + 1
                _DMOM = _DMOM + tele.duration
        case day(tele.date) > 20                           && End of Month Calls and Duration
                _NEOM = _NEOM + 1
                _DEOM = _DEOM + tele.duration
 endcase
```

```
            replace final.DWeekday with _ DWeekday
            replace final.NWeekend with _ NWeekend
            replace final.DWeekend with _ DWeekend
            replace final.NBOM with _ NBOM
            replace final. DBOM with _ DBOM
            replace final. NMOM with _ NMOM
            replace final. DMOM with _ DMOM
            replace final. NEOM with _ NEOM
            replace final. DEOM with _ DEOM
            replace final.TDuration with _TDuration
            replace final.TRevenue with _TRevenue
            =tableupdate(.t.,.t.)

            _caller = tele.caller
            STORE 0.00 TO _TNumber, _NFixedAA,_NFixedRegions,_NMobile, _NInternational
            STORE 0.00 TO _DFixedAA,_DFixedRegions,_DMobile, _DInternational
            STORE 0.00 TO _NDay, _NNight, _DDay, _DNight
            STORE 0.00 TO _NWeekday, _NWeekend, _DWeekday, _DWeekend
            STORE 0.00 TO _NBOM,_NMOM,_NEOM,_DBOM,_DMOM,_DEOM
            STORE 0.00 TO _TDuration, _TRevenue

endif

_allcalls = _allcalls + 1

if left(alltrim(tele.callee),5) == "00251"
        DO CASE
                CASE substr(alltrim(tele.callee),6,1) == "1"
                        _NFixedAA = _ NFixedAA + 1
                        _DFixedAA = _ DFixedAA + tele.duration
                CASE substr(alltrim(tele.callee),6,1) == "9"
```

```
        case day(tele.date) > 10 and day(tele.date) <= 20
                _ NMOM = _ NMOM + 1
                _ DBOM = _ DBOM + tele.duration
        case day(tele.date) > 20
                _ NEOM = _ NEOM + 1
                _ DEOM = _ DEOM + tele.duration
endcase


****** TOTAL REVENUE
 if left(alltrim(tele.callee),5) == "00251"
        do case
                CASE substr(alltrim(tele.callee),6,1) == "1"
                        if tele.date = {^2004-03-02} or cdow(tele.date) = "Sunday" or tele.hh < 8 or tele.hh > 20
                                _TRevenue = _ TRevenue + ROUND((tele.duration * 0.33) / 60,2)
                        else
                                _ TRevenue = _ TRevenue + ROUND((tele.duration * 0.75) / 60,2)
                        endif
                CASE substr(alltrim(tele.callee),6,1) == "2"
                        if tele.date = {^2004-03-02} or cdow(tele.date) = "Sunday" or tele.hh < 8 or tele.hh > 20
                                _ TRevenue = _ TRevenue + ROUND((tele.duration * 0.70) / 60,2)
                        else
                                _ TRevenue = _ TRevenue + ROUND((tele.duration * 1.32) / 60,2)
                        endif
                CASE substr(alltrim(tele.callee),6,1) == "3"
                        if tele.date = {^2004-03-02} or cdow(tele.date) = "Sunday" or tele.hh < 8 or tele.hh > 20
                                _ TRevenue = _ TRevenue + ROUND((tele.duration * 1.10) / 60,2)
                        else
                                _ TRevenue = _ TRevenue + ROUND((tele.duration * 1.92) / 60,2)
                        endif
                CASE substr(alltrim(tele.callee),6,1) == "9"
```

```
                        if tele.date = {^2004-03-02} or cdow(tele.date) = "Sunday" or tele.hh < 8 or tele.hh > 20
                                _ TRevenue = _ TRevenue + ROUND((tele.duration * 0.30) / 60,2)
                        else
                                _ TRevenue = _ TRevenue + ROUND((tele.duration * 0.72) / 60,2)
                        endif
                OTHERWISE
                        if tele.date = {^2004-03-02} or cdow(tele.date) = "Sunday" or tele.hh < 8 or tele.hh > 20
                                _ TRevenue = _ TRevenue + ROUND((tele.duration * 1.63) / 60,2)
                        else
                                _ TRevenue = _ TRevenue + ROUND((tele.duration * 2.72) / 60,2)
                        endif
        endcase
    else
        _ TRevenue = _ TRevenue + ROUND((tele.duration * 10.72) / 60,2)
    endif

    select tele
    SKIP
enddo

select final
append blank
replace final.caller with _caller
replace final.caller with _caller
        replace final.TNumber with _TNumber
        replace final.NFixedAA with _ NFixedAA
        replace final.NFixedRegions with _ NFixedRegions
        replace final.NMobile with _ NMobile
        replace final.NInternational with _ NInternational
        replace final.DFixedAA with _ DFixedAA
        replace final.DFixedRegions with _ DFixedRegions
```

```
replace final.DMobile with _ DMobile
replace final.DInternational with _ DInternational
replace final.NDay with _ NDay
replace final.DDay with _ DDay
replace final.NNight with _ NNight
replace final.DNight with _ DNight
replace final.NWeekday with _ NWeekday
replace final.DWeekday with _ DWeekday
replace final.NWeekend with _ NWeekend
replace final.DWeekend with _ DWeekend
replace final.NBOM with _ NBOM
replace final. DBOM with _ DBOM
replace final. NMOM with _ NMOM
replace final. DMOM with _ DMOM
replace final. NEOM with _ NEOM
replace final. DEOM with _ DEOM
replace final.TDuration with _TDuration
replace final.TRevenue with _TRevenue
=tableupdate(.t.,.t.)

flush
```

APPENDIX 2: ATTRIBUTES AND THEIR SAMPLE VALUES IN THE FINAL DATASET

Caller	Category	TNumber(TN)	NFixedAA	NFixedRegions	NMobile
009201838	20	**164**	16	0	148
009201839	20	**556**	234	22	292
009201842	18	**481**	134	3	336
009201843	08	**217**	56	3	158
009201844	20	**397**	219	9	169

NMobile	NInternational	NDay	NNight	NWeekday	NWeekend	NBOM
148	0	143	21	129	35	62
292	8	516	40	493	63	149
336	8	469	12	434	47	111
158	0	206	11	199	18	65
169	0	390	7	361	36	115

NMOM	NEOM	Tduration(TD)	DFixedAA	DFixedRegions	DMobile
40	62	**8949**	710	0	8239
201	206	**64257**	30304	3432	29690
160	210	**26727**	6822	202	19214
76	76	**6586**	1185	64	5337
130	152	**18282**	9184	667	8431

DInternational	DDay	DNight	DWeekday	DWeekend	DBOM	DMOM
0	7319	1630	6865	2084	3456	2017
831	56619	7638	54298	9959	20670	17821
489	25699	1028	24057	2670	6049	8130
0	6402	184	6100	486	1581	2255
0	18117	165	16203	2079	5414	5637

DEOM	TRevenue(TR)	DPerCall	RPerCall	RPerDuration
3476	**85.76**	*54.56707*	*0.5229268*	*0.009583194*
25766	**931.39**	*115.5701*	*1.6751619*	*0.014494763*
12548	**384.23**	*55.56549*	*0.798815*	*0.014376099*
2750	**76.89**	*30.35023*	*0.3543318*	*0.011674765*
7231	**230.37**	*46.05038*	*0.5802771*	*0.012600919*

APPENDIX 3: SAMPLE RULES TO ASSIGN RECORDS TO APPROPRIATE CLUSTERS

Rule # 1:

If TNumber is greater than or equal to 160 and is less than 262 and DInternational is greater than or equal to 0 and is less than 322 and TRevenue is greater than or equal to 198.4 and is less than 366.43 and RPerCall is greater than or equal to 1.45722222222 and is less than 2.39551587302 then Cluster Index will be 1,

Rule # 2:

If TNumber is greater than or equal to 160 and is less than 262 and DInternational is greater than or equal to 0 and is less than 322 and TRevenue is greater than or equal to 366.43 and is less than or equal to 1435.89 and TDuration is greater than or equal to 13259 and is less than 23279 then Cluster Index will be 1.

Rule # 3:

If TNumber is greater than or equal to 107 and is less than 160 and DInternational is greater than or equal to 61 and is less than 322 and TDuration is greater than or equal to 3615 and is less than 13259 then Cluster Index will be 2,

Rule # 4:

If TNumber is greater than or equal to 107 and is less than 160 and DInternational is greater than or equal to 61 and is less than 322 and TDuration is greater than or equal to 3615 and is less than 13259 and RPerCall is greater than or equal to 0.408857142857 and is less than 1.12710743802 then Cluster Index will be 2,

Rule # 5:

If TNumber is greater than or equal to 21 and is less than 50 and DInternational is greater than or equal to 2283 and is less than or equal to 6093 and TDuration is greater than or equal to 2337 and is less than 3615 Cluster Index will be 3,

Rule # 6:

If TNumber is greater than or equal to 50 and is less than 75 and DInternational is greater than or equal to 961 and is less than 2283 then Cluster Index will be 3,

Rule # 7:

If TNumber is greater than or equal to 21 and is less than 50 and DInternational is greater than or equal to 2283 and is less than or equal to 6093 and TDuration is greater than or equal to 3615 and is less than 13259 then Cluster Index will be 4.

Rule # 8:

If TNumber is greater than or equal to 50 and is less than 75 and DInternational is greater than or equal to 2283 and is less than or equal to 6093 then Cluster Index will be 4.

Rule # 9:

If TNumber is greater than or equal to 262 and is less than or equal to 714 and DInternational is greater than or equal to 0 and is less than 201 and TRevenue is greater than or equal to 366.43 and is less than or equal to 1435.89 then Cluster Index will be 5.

Rule # 10:

If TNumber is greater than or equal to 262 and is less than or equal to 714 and DInternational is greater than or equal to 0 and is less than 201 and TRevenue is greater than or equal to 198.4 and is less than 366.43 and RPerCall is greater than or equal to 0.01 and is less than 0.53 and TDuration is greater than or equal to 23279 and is less than or equal to 63261 then Cluster Index will be 5.

Rule # 11:

If TNumber is greater than or equal to 107 and is less than 160 and DInternational is greater than or equal to 2283 and is less than or equal to 6093 then Cluster Index will be 6.

Rule # 12:

If TNumber is greater than or equal to 160 and is less than 262 and DInternational is greater than or equal to 2283 and is less than or equal to 6093 then Cluster Index will be 6.

DECLARATION

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

Fekadu Mekonnen

June, 2004

The thesis has been submitted for examination with our approval as university advisors.

Dr Nega Gebreyesus

June, 2004