ADDIS ABABA UNIVERSITY
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

Data Mining Application in Supporting Fraud Detection

On Ethio-Mobile Services

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

BY

GEBREMESKAL GIRMA

MARCH 2006

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
Faculty of Informatics
Department of Information Science

DATA MINING APPLICATION IN SUPPORTING FRAUD DETECTION ON ETHIO-MOBILE SERVICES

BY

GEBREMESKAL GIRMA

Name and Signature of Member of the Examining Board

Prof. B.R.K Rao, Chairman, Examining Board	
Dr B.L. Desai, Advisor	
Dr. Kumudha Raimond, External Examiner	

Chairman, Faculty	Signature	Date 01/08/06
Chairman, Graduate Council	Signature	Date

ACKNOWLEDGEMENTS

It seems to be a law of nature that however much time you have for an assignment, a panic situation always appears when the deadline approaches. Even though the work of writing this thesis has been stressful and felt hopeless sometimes, there has always been some progress and light at the end of the tunnel.

A great deal of people have helped me on my research journey, and made work and life more enjoyable meanwhile. There is always a risk in writing an acknowledgement; someone may be forgotten, and some people may not feel appreciated enough. However, more people may become angry if I do not write an acknowledgement at all since they deserve the appreciation, so I give it a try anyway.

First, I want to thank my advisor Dr.B.L.Desai for much help and encouragement. He helped me structure my work and my papers.

I want to thank the management and staffs of Ethiopian Telecommunication Corporation for nice and well functioning cooperation, especially Ato Mulugeta, Ato Dereje, Ato Ayele, W/t Sergut, W/o Leteberehan and also of course I want to thank all the people at the Mobile Service Division.

I want to thank all the people at the department for helping me with many of the practicalities in the life as Msc. student. I am also grateful to my friends, special thanks to Bekele, Dawed, Endale, Meskerem, Tibe, Woubeshet and Yihun for their unlimited cooperation whenever I need help.

Great thanks to my beloved parents and sister and brother who are always supported and encouraged me to continue my studies.

LIST OF TABLES and FIGURES

LIST OF TABLES and FIGURES

Abstract

The problem of Mobile frauds has been getting more and more serious for many years, and is even getting more and more worse not only in western countries but also in some developing countries. Fraud is the most significant threat to the communications business, eroding margins, consuming network capacity and jeopardizing customer relationships. Detection, Analysis and prevention mechanisms are emerging both from telecommunications operators and academia. In this paper, the possible application of data mining in supporting fraud detection on Ethio-Mobile Services has been tested by the use of neural network technique.

The methodology used for this research had three basic steps. These were: data collection, data preparation, and model building and testing. The required data was collected from Ethiopian Telecommunication Corporation which is called Call Detail Record. This record shows the behavior of each mobile phone users. Next, data preparation tasks (such as data cleaning, feature selection, data transformation etc) were undertaken. Several Neural network models were built and tested for their classification accuracy; and the model with encouraging results was taken.

TABLE OF CONTENTS

INTRODUCTION

1.1. Background

The amount of data in the world, in our lives, seems to continue and there is no end in sight. The accumulation of data has taken place at an explosive rate. The increase in computer hardware technology in the past has led to large supplies of powerful and affordable computers, data collection equipment and storage devices. This technology results the processing and accumulation of excess amount of data without being analyzed and used to discover important knowledge from it. Traditional online transaction processing systems (OLTPs) are good at putting data into databases quickly, safely and efficiently but are not good at delivering meaningful analysis in return. Analyzing data can provide further knowledge about a business by going beyond the data explicitly stored to derive knowledge about the business.

Moreover, with these advances in data acquisition and storage technologies, the problem of how to turn raw data into useful information becomes a significant one. Having reached sizes that defy (disobey) even partial examination by humans, modern databases and collections of data sets are literally swamping users. This is where data mining or knowledge discovery in databases (KDD) comes into picture.

Data mining is the discovery of knowledge from data, and uses a variety of tools ranging from classical statistical methods to neural networks and other new techniques originating from machine learning and artificial intelligence. Recently, data mining has been used with substantial results in enabling and improving data base marketing and process optimization.

According to Han and Kamber (2001) the major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of data and the immediate need for turning such data into useful information and knowledge. Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research.

Shapiro (2000) defined Data mining or knowledge discovery in database (KDD) as "a nontrivial extraction of implicit, previously unknown, and potentially useful information from data." This definition encompasses a number of different technical approaches, such as clustering, data summarization, learning classification, finding dependencies networks, analyzing changes, and detecting anomalies

Generally, data mining technology has become a new paradigm for decision making, with applications ranging from database marketing and electronic commerce to fraud detection, credit scoring, warranty management, even auditing data before storing it in a database (Levin and Zahavi, 1999)

The telecommunication industry is the one of the various areas with huge amount of electronic data. This wealth of information is kept in company and master databases, but remains unrecognized in terms of its information content and unused in terms of company management. Therefore, the application of data mining is needed to discover patterns and relationships in the data that may be used to make valid predictions (Moxon, 1996).

The rapidly expanding and highly competitive nature of the telecommunication industry creates a great demand for data mining in order to help understand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of service (Han and Kamber,2001;Bolton and Hand,2002).

Hence, the demand for data mining in discovering patterns and relationships of telecommunication fraud in the huge amount of electronic data is crucial (Moxon, 1996).Telecommunication fraud has been identified as the single biggest cause of revenue loss for telecommunication. Current statistics point to a global loss of USD 55 billion per year, making telecommunication fraud a bigger business than international drug trafficking.

The International Communication Union estimates that more than 2000 variants of telecom fraud exist and that the number is growing with the advent of new service. Telecom fraud attacks are becoming increasingly sophisticated and are tapping into the arrival of these new telecommunication technologies.

Current legislation offers minimal protection against this type of criminal activity and many companies choose not to report fraud for fear of undermining customer confidence for the security of their own services. Fraud is often swept under the carpet as 'bad debt'. Typically, amongst European telecom as much as 30 to 35 percent of bad debt can be written off as a direct result of fraud. For other continents this percent can be greater (Jacobs, 2003).

Jacobs (2003) also pointed out that telecom carriers' write off USD 700million annually to fraud and this figure is expected to increase with the growth of wireless services. As stated by the International Communication Union, more than 30 million people in Africa have access to mobile phones in 2002, compared to only 2 million in 1998. The success of Africa's mobile industry, like that of any continent, has become a target for criminals. Therefore mobile fraud is the most significant threat to the telecommunication business, eroding margins, consuming network capacity and jeopardizing customer relationships.

Due to this fact, a number of researches have been done to show the possible applications of data mining techniques in supporting telecommunication industry in different parts of the world. For example, Bolton and Hand (2001) made research on statistical fraud detection by simple rule-based detection systems and by experts or by application of supervised learning methods to known fraud/non-fraud cases. Pinkal and Rosset in Andreascu and Zilliacus (2001) distinguish between profiling at the level of individual calls, daily call patterns and overall call patterns, and describe what outlier detection methods for detecting anomalous behavior are effective.

Moreover, Gashaw (2004) made a research on possible application of data mining in supporting Ethiopian Telecommunication Corporation. He found out the patterns of potentially solvent and insolvent customers of postpaid mobile phone users. The data he used in his study was customer profile, usage of offered services, and financial transaction.

Jember(2004) also made a research on possible application of data mining in supporting Ethiopian Telecommunication Corporation; and a successful result was obtained in detecting fraud of mobile phone technology (post paid).

Hence, doing this research might fill the gap or make the previous research complete since it includes prepaid mobile, identify other pattern of fraudster behavior (since fraudsters always change their behavior), etc based on the data from Call Record Detail.

Ethiopian Telecommunication Corporation Profile

The Ethiopian telecommunication started with a humble beginning more than a hundred years ago by establishing a telephone link between the capital city and some major provincial cities. Today, telecommunication has extended to the interior of the country and uses technologies such as micro-wave, satellite and even fibber optics.

Starting from those years, the organization had undergone through series of development programs. The major objectives of the corporation are to support the free market economy and investment ventures, to satisfy the demands of the private sector for telecom services and to fully participate in and help the integrated rural development program of the country and to generate profit in order to secure funds for further improving its network.

Ethiopian Telecommunication's development programs are not only meant to expand and improve the telephone, fax and the other relatively old types of services to the rural and urban areas, but also, through the various transmission systems, it plans to provide Internet and telemedicine as well as Interactive Distance Learning access to regional towns, including higher education institutions with many colleges in the regional states located far from the center of the country.

Based on the report from Ethiopian Telecommunication, Mobile service was introduced in Ethiopia, Addis Ababa, in April 1999 with limited resource and network capacity. A

GSM technology was selected to start a mobile service in Ethiopia. This technology is a standard and field prove digital mobile system with modern features adopted by the world market for global mobility and it was in use at that time in many countries. It has an excellent speech quality achieved through digital speech transmission compared to analogue system.

However, the service quality has deteriorated as an effect of increasing the number of connected subscribers to the network. Therefore, to meet the capacity demands and clear network congestion the corporation has implemented network expansion activities. Soon it will have 1.1 million subscribers network capacity, and it will continue to aggressively develop the network and become the most successful business enterprise.

Mobile Service Division was founded in 1996. Organizationally, launches its service under the Managing Director. This Mobile Telecom Service Department is managed by a department Managers and three Division Managers responsible to lead the Technical, Resource Management, and Customer Service Division. The Technical Division is responsible for all technical matters had two deputy Divisions under it (Operation and Maintenance Deputy Division; and Engineering and Project Deputy Division). The Customer Service Division is responsible for customer care activities like sales, collection and after sales service had three Deputy Divisions under it (Sales and Collection, Domestic Customer Care; and International Customer Care). The Resource Management Division is responsible for the administration of finance, human and material resource had two Deputy Divisions under it (Administration and Finance).

The number of mobile subscribers was 53614 Post-Paid and 381417 Prepaid as of May 2005. In spite of the fact that the report is not yet released, it is estimated that the number

of subscribers might be doubled since the service coverage are now expanded to Mekele, Dessaie, Deredawa, Jijiga, Gonder, BaherDar, Jimma, Nazerat, Nekemt, Awassa and the neighboring cities.

1.2. Statement of the Problem and Justification of the study

✓ The major problem that made this research to be conducted is the existence of high level of fraud in telecommunication industry in general and in Mobile phone services in particular.

Telecommunications fraud costs carriers billions of revenue dollars annually. In fact, the Communication Fraud Control Association conducted a survey and determined that $35-$40 billion in losses is due to telecom fraud world wide. Although many forms of technical fraud are becoming less feasible and cost effective for the fraud criminal element, there are many other areas in which the carrier are exposed to fraud. Criminals are continually focused on developing new techniques and methods in order to perpetrate fraud. The telecom fraud criminal is typically motivated to avoid toll charges, to make money, to maintain anonymity, to demonstrate intellectual superiority. Because of this, fraud detection application must become more sophisticated to keep pace with the criminals.

Mobile communication has been readily available for several years, is the major business today. It provides a valuable service to its users who are willing to pay a considerable premium over a fixed line phone, to be able to walk and talk freely. Because of its usefulness and the money involved in the business, it is subject to fraud and criminal interest. Some of the features of mobile communication make it an alluring target for

criminals. Because of its newness and expensiveness provides an opportunity for the criminally inclined to try and make a profit out of the situation (Hynnien, 2003).

There are different types of mobile fraud, with new variations being developed all the time. The most common by value according to Cerebrus Solutions (the telecom equipment company) which produces fraud-detection software are:

Internal fraud: This may involve an employee altering the telecom switch to give a free mobile number or numbers to friends - who use them for commercial gain.

Premium-rate services fraud: Here bogus companies set up premium-rate services and collect revenues from the mobile operator knowing their services have been called up by fraudsters who will never pay.

Subscription fraud: This involves registering for a service under a false name or signing up for a service with no intention of paying for it.

Roaming fraud: This involves running up bills outside the home network with no intention of paying. In other words, stolen and cloned mobile phones are used to make international calls and in roaming, possibly abroad.

Prepaid fraud: This includes cloning prepaid cards, so one card's details can be used by other people.

Therefore, to overcome the high level fraud in the industry, an effective means of controlling and preventive mechanisms should be employed. Due to the nature of the technology it self and the restless human being discovery of new types of fraud, fraud detection is necessary to take any sound and strong action on fraudulent customers before

they create a huge loss in the corporation earnings. The whole point of the research will be to mine the kind of fraud (identifying patterns of behavior on different users' call and thereby detecting as many illegal calls as possible) from the very large data generated by the Call Detail Record switch machine of Ethiopian Telecommunication Corporation. This means that it will enable the corporation to detect early those calls that will make an illegal call.

The data mining function for this research problem is defined to be a classification problem, since the ultimate goal is to classify each customer call as potentially fraudulent or non fraudulent. A suitable technique for this problem will be neural network to achieve better result. Therefore, different software were examined by taking into consideration their application to the problem and availability to work with them during the research period and gradually BrainMaker neural Network software was selected.

1.3. Objectives of the Study

1.3.1. General Objective

The general objective of the research work is to explore the potential applicability of data mining technology in developing a model that can support fraud detection in Ethio-Mobile Services of Ethiopian Telecommunication Corporation.

(CDR) that is normally available in the switch machine of the corporation which records all the details of the calls made and even attempted by the subscribers daily.

1.4.3. Data Preparation

In this step, data was prepared for eventual analysis, which actually can involve a large number of sub-steps and procedures, and even preliminary analysis. Data preparation had been performed in different stages by different software systems.

Common preparation tasks include:

- Modifying the format or schema of data set so that a particular algorithm can use it as input, or so that the algorithm will run faster.
- Adding a description to each column so that it could easily be remembered what it means.
- Removing certain attributes that will not be helpful.
- Splitting the collected data into two parts. One was used to run analysis (training set) and find results. The second (testing set) has been used to verify if the results and conclusions are accurate.

1.4.4 Data Analysis

With goals defined and data prepared, it moved on to the meat of the process i.e. the use of analytical tools to generate informative conclusions.

As such, the collected data has been cleaned into a form that was suitable for the software. The type of analysis is classification and the data mining software techniques used for this research project is neural network to achieve better result. Accordingly, after examining different software by considering their application for the problem and their

availability, BrainMaker Neural Network Software has been selected as a tool for analysis for this research.

1.4.5 Model Verification

As the choice of data mining technology for classification tasks seems to be strongly dependent on the application, the data mining technique that are employed for this research work need the data to be classified into training and testing before building the model. Therefore, this part was decided after the analysis of each model on to how well the results work.

1.5. Scope and Limitation of the Study

The scope is focused on prepaid mobile phone of Ethiopian Telecommunication Corporation, where the required customer data was available. Furthermore, the study was limited in development a model that can predict customers call as potentially fraudulent or non fraudulent.

The major limitations while undertaking this research was the time of the experts at the corporation. These experts were very busy and/or out of office, and this in turn had become a constraint on the amount of identifying the types of frauds collected from the call detail record (CDR).

1.6 Research Contribution

The result of this study can have a direct or indirect impact for improving the efficiency of the Ethiopian Telecommunication Corporation. On the basis of the findings the Corporation could overcome the current high level fraud in the industry and an effective

means of controlling and preventive mechanisms could be employed. Furthermore, it can assist the corporation to plan what to do next in the corporation's long term development programs.

Additionally, the findings of this research undertaking could initiate a more in-depth comprehensive study in the area

1.7 Thesis organization

The body of this paper is structured into Five Chapters. Chapter one is the introductory part, it contains the background, problems and justifications to conduct the research, objectives and methodology to carry out the research, and its scope and limitations. Chapter Two and Chapter Three present reviews of published works on data mining and telecommunication (mobile) fraud respectively. Chapter Four deals with different pre-processing steps and reports the experiment of the research. Finally, chapter Five presents conclusions and recommendations of the research.

As a result of these development and many other research products, data mining has been flowered and become a powerful technology. As Alexander (2000) broadly categorizes, the data mining technology, given database of sufficient size and quality, can generate new business opportunities by providing these two major capabilities:

i. Automated prediction of trends and behaviors. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data - quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

ii. Automated discovery of previously unknown patterns. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Therefore, Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature: massive data collection, powerful multiprocessor computers and data mining algorithms.

2.2. Data Mining as Knowledge Discovery Process

Data, information, knowledge became focus of research and application in different field of study. We often see ***data*** as a string of bits, or numbers and symbols, or "objects" which are meaningful when sent to a program in a given format (but still un-interpreted). We use bits to measure ***information***, and see it as data stripped of redundancy, and reduced to the minimum necessary to make the binary decisions that essentially characterize the data (interpreted data). We can see ***knowledge*** as integrated information, including facts and their relations, which have been perceived, discovered, or learned as

our "mental pictures". In other words, knowledge can be considered data at a high level of abstraction and generalization (Gray, 1998).

Knowledge discovery and data mining (KDD) as pointed out by Berry and Linoff (2000) are the rapidly growing interdisciplinary field which merges together database management, statistics, machine learning and related areas—aims at extracting useful knowledge from large collections of data.

There is a difference in understanding the terms "knowledge discovery" and "data mining" between people from different areas contributing to this new field. Chung, Gray and Mannino (1998) characterize these two terms. They define Knowledge discovery in databases as the process of identifying valid, novel, potentially useful, and ultimately understandable patterns/models in data. By contrast, They define data mining as a step in the knowledge discovery process consisting of particular data mining algorithms that, under some acceptable computational efficiency limitations, finds patterns or models in data.

From the above definition, we can understand that the goal of both knowledge discovery and data mining is to find interesting patterns and/or models that exist in databases but are hidden among the volumes of data.

2.3. Data Mining and Data warehousing

When beginning work on a data mining problems, it is necessary to bring all the data together into a set of instances first. Integrating data from different sources usually presents many challenges. Different departments of an organization will use different styles of record keeping, different time periods, different degrees of data aggregation, different primary keys, and will have different kinds of error. The data must be assembled, integrated, and cleaned up. The idea of enterprise wide database integration is called data warehousing (Witten and Frank, 2000).

Gargano and Raggad (1999) point out that a data warehouse is a semantically consistent data store that serves as a physical implementation of a decision support data model, and stores the information on which an enterprise needs to make strategic decisions through integrating data from multiple heterogeneous data source. It enables each user not only to

share a common, widely distributed, diverse database but also to analytically explore, discover, and better comprehend fundamental trends and relationships using all of the available data quickly and correctly. Metadata, information concerning data, describing the warehouse are also an integral part of the system.

2.4. Data Mining and OLAP

According to Two Crows Corporation (1998), one of the most common questions from data processing professionals is about the difference between data mining and OLAP (On-Line Analytical Processing). As we shall see below, they are very different tools that can complement each other.

OLAP is part of the spectrum of decision support tools. Traditional query and report tools describe *what* is in a database. OLAP goes further; it's used to answer *why* certain things are true. The user forms a hypothesis about a relationship and verifies it with a series of queries against the data. For example, an analyst might want to determine the factors that lead to loan defaults. He/she might initially hypothesize that people with low income are bad credit risks and analyze the database with OLAP to verify (or disprove) this assumption. If that hypothesis were not borne out by the data, the analyst might then look at high debt as the determinant of risk. If the data did not support this guess either, he or she might then try debt and income together as the best predictor of bad credit risks.

In other words, the OLAP generates a series of hypothetical pattern and relationships and uses queries against the database to verify them or disprove them. OLAP analysis is essentially a deductive process. But what happens when the number of variables being analyzed is in the dozens or even hundreds? It becomes much more difficult and time-consuming to find a good hypothesis (let alone be confident that there is not a better explanation than the one found), and analyze the database with OLAP to verify or disprove it.

When contrasting data mining and OLAP, Bargain (2000) as cited in Gashaw(2004) describes that data mining is different from OLAP because rather than verify hypothetical patterns, it uses the data itself to uncover such patterns. It is essentially an inductive process. For example, suppose the analyst who wanted to identify the risk factors for loan

default were to use a data mining tool. The data mining tool might discover that people with high debt and low income were bad credit risks (as above), but it might go further and also discover a pattern the analyst did not think to try, such as that age is also a determinant of risk.

Here is where data mining and OLAP can complement each other. For instance, before acting on the pattern, the analyst needs to know what the financial implications would be of using the discovered pattern to govern who gets credit. The OLAP tool can allow the analyst to answer those kinds of questions

Furthermore, Patterson (2000) also notifies that OLAP is also complementary in the early stages of the knowledge discovery process because it can help us to explore our data, for instance by focusing attention on important variables, identifying exceptions, or finding interactions. This is important because the better we understand our data, the more effective the knowledge discovery process will be.

2.5. Data Mining and other statistical tools

People have used statistical techniques for centuries to understand the natural world. These techniques included predictive algorithms which are called regression by statisticians, sampling methodologies, and experimental design. Statistics is one of the major disciplines that have contributed to data mining. It is still an important support to the field data mining (Berry and Linoff, 2000).

Data mining does not replace traditional statistical technique; it is rather an extension of statistical methods which is the result of major changes in the statistic community. The developments of most statistical techniques are based on elegant theory and analytical method that worked well on the modest accounts of data being analyzed (Gandy, 2002).

As Brand and Gerritsen (1998) state, data miners have always used statistical tools and statisticians are now showing an interest in Data Mining problems. The interactions between the two disciplines will be very beneficial to both of them. Since similarities and connections between Data Mining and Statistics are rather notorious.

One difference as described by Luan (2004) is on the type of data. This means that while statisticians traditionally work with "first hand data" that has been collected or produced to check specific hypotheses, data miners work with "second hand data" often assembled from different sources. The idea is to find interesting facts and potentially useful knowledge hidden in the data and often unrelated to the primary purpose why the data have been collected. In addition to this, statistical data can be experimental but in data mining the data is typically observational.

Statistics is very useful in providing a language and framework for quantifying the uncertainty, which results when one tries to infer general patterns from a particular sample of an overall population. However, it does not solve all data mining problems. Moreover, the computational complexity of statistical approaches does not grow well with larger data sets (Levin and Zahavi, 1999).

In general data mining is a tool for increasing the productivity of people trying to build predictive models by making AI and statistical techniques available to the skilled knowledge workers as well as the trained professionals (Two Crows Corporation, 1999).

2.6. Data Mining Technologies

A model, according to Muenchen(2003),is the process of developing rules, which can classify or predict with an estimated level of precision. An important function of data mining is the production of a model. There are two common categories of data mining technology. These are descriptive and predictive modeling. These two data mining technologies are based on one of two kinds of learning: supervised and unsupervised (sometimes referred to as directed and undirected learning). Supervised learning functions are typically used to predict a value. Unsupervised learning functions are typically used to find the intrinsic structure, relations, or affinities in a body of data but no classes or labels are assigned prior.

Therefore the key distinction between these two data mining technology (Predictive and Descriptive) is that prediction has as its objective a unique variable (the market's value, the disease class, the brittleness, etc), while in descriptive problems no single variable is central to the model.

2.6.1. Descriptive modelling

Descriptive modeling tries to find models for the data. The aim of these models are to describe, not to predict models. As a consequence, descriptive models are used in the setting of unsupervised learning. Typical methods of descriptive modeling are density estimation, smoothing, data segmentation, and clustering. Clustering is a well-studied and well-known technique.

Many different approaches and algorithms, distance measures and clustering schemes have been proposed. The most widely used method of choice is k-means clustering.

The reasoning behind cluster analysis is the assumption that the data set contains natural clusters which, when discovered, can be characterized and labeled. While for some cases it might be difficult to decide to which group they belong, we assume that the resulting groups are clear-cut and carry an intrinsic meaning. In segmentation analysis, in contrast, the user typically sets the number of groups in advance and tries to partition all cases in homogeneous subgroups.

2.6.2. Predictive modelling

As noted above, predictive modeling falls into the category of supervised learning. Hence, one variable is clearly labeled as target variable y and will be explained as a function of the other variables x. The nature of the target variable determines the type of model: classification model, if y is a discrete variable, or regression model, if it is a continuous one.

Many models are typically built to predict the behavior of new cases and to extend the knowledge to objects that are new or not yet as widely understood. Predicting the value of the stock market, the outcome of the next governmental election, or the health status of a person etc use classification schemes to group their customers into different categories of risk.

In general, the aim of predictive modeling, as Berry and Linoff (2000) stated, is to build model that will permit the value of one variable to be predicted from the known values of

other variable. In classification, the variable being predicted is categorical while in regression the variable is quantitative.

2.6.2.1. Data mining techniques for data classification

As West(2004) defined , A data mining function for predicting target values for new records using a model built from records with known target values is known as classification .

Classification models, as Bauwens etal(2002) state, follow one of three different approaches: the discriminative approach, the regression approach, or the class-conditional approach

The discriminative approach aims in directly mapping the explanatory variables X to one of the K possible target categories y1,….,yk. The input space X is hence partitioned into different regions which have a unique class label assigned. A neural network is examples for this.

The regression approach (e.g. logistic regression) calculates the posterior class distribution P(Y|X)for each case and chooses the class for which the maximum probability is reached. Decision trees (CART, C5.0, and CHAID) classify for both the discriminative approach and the regression approach, because typically the posterior class probabilities at each leaf are calculated as well as the predicted class.

The class-conditional approach starts with specifying the class-conditional distributions (X|Yi,θi) explicitly. After estimating the marginal distribution P(Y), Bayes' rule is used to derive the conditional distribution P (Y|x). The name Bayesian classifiers is widely used for this approach. The class-conditional approach is particularly attractive, because they allow for general forms of the class-conditional distributions. Parametric, semi-parametric, and non-parametric methods can be used to estimate the class-conditional distribution. The class-conditional approach is the most complex modeling technique for classification. The regression approach requires fewer parameters to fit, but still more than a discriminative model.

There is no general rule which approach works best, it is mainly a question of the goal of the researcher whether posterior probabilities are useful, e.g. to see how likely the "second best" class would be.

From the above discussion, we can understand that the data mining techniques for classification are different in nature depending on the goal of the researcher and data types. Some of them are: Decision tree, Naive Bayes, Neural network, Adaptive Bayes' Networks and Model Seeker. In this section, however, the first three data mining techniques will be briefly discussed i.e. Decision Tree, Naïve Bayesian Classification and Neural Networks.

2.6.2.1.1 Decision tree

According to Quinlan (2003), decision trees are powerful and popular tools for classification (predicting what group a case belongs to), and for regression (predicting a specific value). The attractiveness of tree-based methods is due in large part to the fact that, in contrast to neural networks, decision trees represent *rules*. Rules can readily be expressed so that human can understand them. In other words, the visual presentation makes the decision tree very easy to understand and assimilate. As a result, the decision tree has become a very popular data mining technique.

The decision tree method encompasses a number of specific algorithms; Nelson (1998) gives some examples. These are Classification and Regression Trees (CART), Chi-squared Automatic Interaction Detection (CHAID), and C4.5. CART requires less data preparation than CHAID, but produces only two –way splits. CHAID can produce tree with multiple sub nodes for each split. C4.5 comes from world of machine learning and is based on information theory.

2.6.2.1.2. Naive Bayes

The Naive Bayes algorithm (NB) makes predictions using Bayes' Theorem, which derives the probability of a prediction from the underlying evidence, as described below. NB affords fast model building and scoring.

NB can be used for both binary and multiclass classification problems to answer questions such as "Which customers will switch to a competitor? Which transaction patterns suggest fraud? Which prospects will respond to an advertising campaign?" For example, suppose a bank wants to promote its mortgage offering to its current customers and that, to reduce promotion costs, it wants to target the most likely prospects. The bank has historical data for its customers, including income, number of household members, money-market holdings, and information on whether a customer has recently obtained a mortgage through the bank. Using NB, the bank can predict how likely a customer is to respond positively to a mortgage offering. With this information, the bank can reduce its promotion costs by restricting the promotion to the most likely candidates.

2.6.2.1.3. Neural networks

Artificial neural networks are one of the most prominent data mining techniques. Their fame is two-fold: famous for astonishing good prediction results, unfamous for their black box behavior and lack of reproducibility of achievements. Neural networks are a classical method for predictive modeling. They have been used for classification and prediction to a similar extent.

Artificial neural networks are a form of computation that is modeled on brain processes. In the brain, billions of neurons are interconnected through synapses to form a biological neural network. Information is transmitted through the network by electrochemical signals. In artificial neural networks this process is simulated. There are many types of artificial neural networks (or ANN), but they all share common characteristics.

As CEREBRUS (2002) pointed out, neural network have been successfully used in a variety of industries such as Finance, Retail, Manufacturing, Energy, Health, Telecommunications and Security. Their application ranges across many fields such as financial market prediction, sales forecasting, mineral exploration, process control, speech recognition, marketing and, of course, fraud detection.

- A large consumer package goods company can apply data mining to improve its sales process to retailers. Data from consumer panels, shipments, and competitor activity can be applied to understand the reasons for brand and store switching. Through this analysis, the manufacturer can select promotional strategies that best reach their target customer segments.

To sum up, data mining is an important tool to solve real-world problems in business or research areas (engineering, science, and business etc)

2.7.2. Application of Data Mining in Telecommunication

Telecommunications networks generate and store vast quantities of data, sometimes of the order of several gigabytes per day. With such enormous amount of data stored, it is increasingly important to develop powerful tool for analysis of such data and mining interesting knowledge from it. Therefore, data mining techniques are important for a process of inferring knowledge from such huge data.

The first step in the data mining process is to understand the data. Without such an understanding, useful applications cannot be developed. In this section we describe the three main types of telecommunication data, Weiss (1998). These three main types of telecommunication data are discussed below. These are: Call Detail Data, Network Data and Customer Data.

2.7.2.1 Call Detail Data

Every time a call is placed on a telecommunications network, descriptive information about the call is saved as a call detail record. The number of call detail records that are generated and stored is huge. Call detail records include sufficient information to describe the important characteristics of each call. At a minimum, each call detail record will include the originating and terminating phone numbers, the date and time of the call and the duration of the call. Call detail records are generated in real-time and therefore will be available almost immediately for data mining.

2.7.2.2 Network Data

Telecommunication networks are extremely complex configurations of equipment, comprised of thousands of interconnected components. Each network element is capable of generating error and status messages, which leads to a tremendous amount of network data. This data must be stored and analyzed in order to support network management functions, such as fault isolation. This data will minimally include a timestamp, a string that uniquely identifies the hardware or software component generating the message and a code that explains why the message is being generated.

Due to the enormous number of network messages generated, technicians cannot possibly handle every message. For this reason data mining technology is now helping identify network faults by automatically extracting knowledge from the network data.

2.7.2.3 Customer Data

Telecommunication companies, like other large businesses, may have millions of customers. By necessity this means maintaining a database of information on these customers. This information will include name and address information and may include other information in order to improve results. For example, customer data is typically used to supplement call detail data when trying to identify phone fraud.

The telecommunications industry was an early adopter of data mining technology, largely because of the amount and quality of the data it collects. This has resulted for many data application. Some typical applications which are stated in Weiss (1998) are adopted and described in this section. These applications are divided into three application areas: fraud detection, marketing/customer profiling and network fault isolation.

i. Fraud Detection

Fraud is a serious problem for telecommunication companies, leading to billions of dollars in lost revenue each year. For this reason, the applications of data mining focus on identifying the type of fraud. These applications should ideally operate in real-time using the call detail records and, once fraud is detected or suspected, should trigger some action. This action may be to immediately block the call and/or deactivate the account, or

may involve opening an investigation, which will result in a call to the customer to verify the legitimacy of the account activity.

ii Marketing/Customer Profiling

Telecommunication companies maintain a great deal of data about their customers. In addition to the general customer data that most businesses collect, telecommunication companies also store call detail records, which precisely describe the calling behavior of each customer. This information can be used to profile the customers and these profiles can then be used for marketing and/or forecasting purposes.

iii. **Network Fault Isolation**

Telecommunication networks are extremely complex configurations of hardware and software. Most of the network elements are capable of at least limited self-diagnosis, and these elements may collectively generate millions of status and alarm messages each month. In order to effectively manage the network, alarms must be analyzed automatically in order to identify network faults in a timely manner—or before they occur and degrade network performance. A proactive response is essential to maintaining the reliability of the network. Because of the volume of the data, and because a single fault may cause many different, seemingly unrelated, alarms to be generated, the task of network fault isolation is quite difficult. Data mining has a role to play in generating rules for identifying faults.

2.7.3. Application of Data Mining in Supporting Telecommunication Fraud

Frauds have plagued telecommunication industries, financial institutions and other organizations for a long time. As Cao, etal (2004) indicate the problem of telecommunication frauds has been getting more and more worse not only in western countries but also in some developing countries. Holmen and Tresp (1998) indicated that the telecommunication industry looses approximately 2-5% of its total revenue to fraud. In China, for example, it is reported that loses amounted to 20 billion RMB YUAN in 2001. As for situation in western countries , De Jager cited in Cao etal(2004),it is estimated that the revenue losses total 3% of 6% annul depending on specific service and

CHAPTER THREE: FRAUD

3.1. Introduction

Many literatures define fraud in different ways. For instance, Funk and Wagnall's New Standard Dictionary of the English Language (1963) defines Fraud as "an act of deliberate deception practice with the object of securing something to the prejudice of another; a trick or stratagem intended to obtain an unfair advantage." Lundin(2002) takes fraud as "an intentional deception or misrepresentation that an individual knows to be false or does not believe to be true and makes, knowing that the deception could result in some unauthorized benefit to himself/herself or some other person"

In the context of this thesis, fraud can be defined as a deliberate act of obtaining access to mobile services with no intention of paying or obtaining unbillable services and undeserved fees.

From the above definition it is easy to state the losses caused by fraud can be taken as a primary motivation for fraud detection. As Holmen and Tresp (1998) indicated fraud is costly to a network carrier both in terms of lost income and wasted capacity. They also added that it has been estimated that the telecommunication industry looses approximately 2-5% of its total revenue to fraud. The true losses are expected to be even higher since telecommunication companies are unwilling to admit fraud in their systems.

Although prevention technologies are the best way of reducing fraud, fraudsters are adaptive and, given time, will usually find ways to get out of such measures. Methodologies for the detection of fraud are essential if we are to catch fraudsters once fraud prevention has failed.

Bolton and Hand (2002) distinguishing between fraud prevention and fraud detection, state that fraud *prevention* describes measures to stop fraud occurring in the first place. These include elaborate designs, fluorescent fibers, multitone drawings, watermarks, laminated metal strips, and holographs on banknotes, PINs for bankcards, Internet

security systems for credit card transactions, SIM cards for mobile phones, and passwords on computer systems and telephone bank accounts. Of course, none of these methods are perfect, and in general, a compromise has to be struck between expense and inconvenience (for example, to a customer), on the one hand, and effectiveness on the other.

On the contrary, fraud *detection* involves identifying fraud as quickly as possible once it has been perpetrated. Fraud detection comes into play once fraud prevention has failed. In practice, of course fraud detection must be used continuously, as one will typically be unaware that fraud prevention has failed. We can try to prevent credit card fraud by guarding our cards assiduously, but if nevertheless the card's details are stolen, then we need to be able to detect, as soon as possible, that fraud is being perpetrated.

Fraud detection is an endlessly evolving discipline. Whenever it becomes known that one detection method is in place, criminals will adapt their strategies and try others. Of course, new criminals are also constantly entering the field. Many of these will not be aware of the fraud detection methods which have been successful in the past, and will adopt strategies which lead to identifiable frauds. This means that the earlier detection tools need to be applied as well as the latest developments.

The development of new fraud detection methods is made more difficult by the fact that the exchange of ideas in fraud detection is severely limited. It does not make sense to describe fraud detection techniques in great detail in the public domain, as this gives criminals the information that they require in order to evade detection. Data sets are not made available and results are often censored, making them difficult to assess (for example, Leonard 1993).

Many fraud detection problems involve huge data sets that are constantly evolving. For example, the credit card company Barclaycard carries approximately 350 million transactions a year in the UK alone (Hand *et al*, 2000), The Royal Bank of Scotland, which has the largest credit card merchant acquiring business in Europe, carries over a billion transactions a year, and AT&T carries around 275 million calls each weekday (Cortes and Pregibon, 1998).

Processing these huge data sets in a search for fraudulent transactions or calls requires more than mere novelty of statistical model, and also needs fast and efficient algorithms: data mining techniques are relevant. The above numbers also indicate the potential value of fraud detection: if 0.1% of a 100 million transactions are fraudulent, each losing the company just £10, then overall the company loses £1 million (Kvarnstrom, Lundin and Jonson, 2002).

Because of the rising needs of protection against frauds, research in this area is ever so important. We can not make computer systems and services 100% secure only by using mechanisms that prevent or stop attacks. Hence, data mining provide effective technologies for fraud detection and has been applied successfully to detect telecommunication fraud activities.

Therefore, this thesis focuses on data mining application in supporting Fraud Detection on Mobile Communication at Ethiopian Telecommunication Corporation

3.2. Telecommunication fraud

The term telecommunications was first used for wired telephony. Today, telecommunications are one of the most important of the contemporary ICTs. They include wired and wireless telephony; different mobile services, such as cellular telephones and paging; voice and data transmission; and Integrated Services Digital Networks (ISDN), which provide a very high quality of voice as well as high data communication rates. Existing telephone networks are now also used as a complement to computer networks, including the Internet and other wide area networks (WAN).

Yet since their very origins at the beginning of this century, the main goal of telecommunications has remained that of providing better, faster and more reliable person-to-person communication.

The development of the global situation concerning the availability of telecommunication services may be characterized by the growth in the number of telephone lines (both main and cellular) per capita in different geographical regions.

According to UNESCO (2000), there is a smooth and almost linear growth in the number of main telephone lines per 100 inhabitants in different regions of the world during the last twenty years and an exponential growth of the number of cellular subscribers per 100 inhabitants over the last 10 years. There is a sharp difference in the availability of telephone services (both wired and cellular) between developed and developing countries. Cellular telephony is becoming more and more popular throughout the world, mostly in the developed countries.

In this telecommunications industry there is a fraud losses and revenue leakage which can have a tremendous impact. Many studies show that billable events simply 'get lost' or 'become polluted', often because an operator's processing technology fails to accurately convert events into billable records(Wu and Park,2002).

The problem of telecommunications frauds has been getting more and more serious for many years, and is even getting more and more worse not only in western countries but also in some developing countries. Therefore, to be competitive in telecommunications market, it is essential for telephone companies to solve these problems by targeting their customers and optimize the performance of their networks.

KDD technology opens new avenues of opportunity in these areas thanks to the wealth of data available for exploration and analysis. In fact, telecommunications companies store vast amounts of data such as customer accounts, call data, equipment records, and fault logs, which represent an invaluable source of information that can be exploited through data mining for a number of purposes (e.g. to combat telecommunication fraud).

There are as many definitions of telecom fraud as there are fraud managers employed in the industry. However, there does seem to be a general consensus that telecom fraud, as the term is generally applied, involves the theft of services or deliberate abuse of voice and data networks. Furthermore, it is accepted that in these cases the perpetrator's intention is to completely avoid or at least reduce the charges that would legitimately have been charged for the services used. On occasion, this avoidance of call charges will be achieved through the use of deception in order to fool billing and customer care systems into invoicing the wrong party.

In addition to financial losses mentioned above, telecommunication fraud may cause distress, loss of service, and loss of customer confidence (Hoath 1998). As noted by Barson et al.(1996), it is difficult to provide precise estimates of on fraud losses , since some fraud may be never detected, and the operators are reluctant to reveal figures. As Parker (1996) and O 'Shea (1997) as cited in Holmen(2000),indicated, since the operators are facing increasing competition and losses have been on the rise telecommunication fraud has gone from being a problem carriers were willing to tolerate to being one that dominates the front pages of both trade and general press.

Mobile communication networks, which are the focus of this research, are particularly appealing to fraudsters as the calling from the mobile terminal is not bound to a physical place and a subscription is easy to get. This provides means for an illegal high profit business requiring minimal investment and relatively low risk of getting caught. Fraud is usually initiated by a mobile phone theft, by cloning the mobile phone card or by acquiring a subscription with false identification. After intrusion the subscription can be used for gaining free services either for the intruder himself or for his illegal customers in form of call-selling. In the latter case, the fraudster sells calls to customers for reduced rates (Holmen and Tresp1998).

3.3. Mobile fraud

According to Computer Desktop Encyclopedia (2000), mobile phones have various names /terms in different languages. For instance, they are known as: *cell phones* or *cells* in Canada, India, the Philippines, Pakistan, South Africa, USA ; *celulares* (singular form *celular*) in Brazil, Chile, Mexico,Puerto Rico and other spanish speaking countries as the spanish word for *Cellular* ;*Fón Póca So-Gluiste*, literally meaning "phone-pocket-that-moves" in Irish; *GSMs* in Belgium; *hand phones* in many Asian countries such as South Korea ; *mobiles* in Australia, India, Ireland, New Zealand, UK etc

Mobile phones have a long and varied history that stretches back to the early 1970s. Due to their low establishment costs and rapid deployment, mobile phone networks have since spread rapidly throughout the world, outstripping the growth of fixed telephony. Such

networks can often be economic, even with a small customer base, as mobile network costs are mostly call volume related, while fixed-line telephony has a much higher subscriber related cost component.

In most of Europe, wealthier parts of Asia, and Australasia, mobile phones are now virtually universal, with the majority of the adult, teenage, and even child population owning one.

Mobile phones are designed to work on cellular networks and contain a standard set of services that allow phones of different types and in different countries to communicate with each other.

Before the phone can be used, a subscription to a mobile phone operator (a.k.a. carrier) is required. For phones on GSM networks, the operator will issue a SIM card which contains the unique subscription and authentication parameters for that customer; alternatively, the carrier will put the customer's handset identifier into its subscriber database so that the handset can make calls on the network. Once the SIM card is inserted into the phone, services can be accessed. Mobile phones do not only support voice calls; they can also send and receive data and faxes (if a computer is attached), send short messages (or "text messages"; see SMS), access WAP services, and provide full Internet access using technologies such as GPRS. Mobile phones usually have a clock and a calculator and often one can play some games on them.

Many mobile phones support 'auto-roaming', which permits the same phone to be used in multiple countries. For this to work, the operators of both countries must have a roaming agreement.

Nowadays, mobile communication becomes major business today and provides a valuable service to its users who are willing to pay a considerable first-class over a fixed line phone. Because of its usefulness and the money mixed up in the business, it is subject to fraud and illegal interest. More over, some of the features of mobile communication make it an alluring target for criminals. As a result of these and other factors, there are many kinds of mobile phone fraud

Therefore, it is obvious that it is necessary to have a clear definition of the types of Mobile Fraud in order to be able to successfully apply data mining (to discovery of meaningful pattern in the data (Call detail record)).

As Hollmen (2000) indicated, historically, earlier types of fraud used technological means to acquire free access. Cloning of mobile phones by creating copies of mobile terminals with identification numbers from legitimate subscribers was used as a means of gaining free access (Davis and Goyal 1993).In the era of analog mobile terminals, identification numbers could be easily captured by eavesdropping with suitable receiver equipment in public places, where mobile phones were evidently used. One specific type of fraud, tumbling, was quite prevalent in the United States (Davis and Goyal 1993).It exploited deficiencies in the validation of subscriber identity when a mobile phone subscription was used outside of the subscriber's home area. The fraudster kept tumbling (switching between) captured identification numbers to gain access. Davis and Goyal (1993) also state that the tumbling and cloning fraud have been serious threats to operators 'revenues. First fraud detection systems examined whether two instances of one subscription were used at the same time (overlapping calls detection mechanism)or at locations far apart in temporal proximity (velocity trap).Both the overlapping calls and the velocity trap try to detect the existence of two mobile phones with identical identification codes, clearly evidence in cloning. As a countermeasure to these fraud types, technological improvements were introduced.

However, new forms of fraud came into existence. A few years later, O 'Shea (1997) reports the so-called subscription fraud to be the trendiest and the fastest-growing type of fraud. In similar spirit, Hoath (1998) characterizes subscription fraud as being probably the most significant and prevalent worldwide telecommunications fraud type. In subscription fraud, a fraudster obtains a subscription (possibly with false identification) and starts a fraudulent activity with no intention to pay the bill. It is indeed non-technical in nature and by call selling, the entrepreneur-minded fraudster can generate significant revenues for a minimal investment in a very short period of time (Johnson 1996).From the above explanation it is evident that the detection mechanisms of the first generation soon became inadequate. The more advanced detection mechanisms must be based on the behavioral modeling of calling activity, which is also the subject of this thesis.

There are many types of fraud being committed around the world today. Of course with new variations being developed all the time, Hynninen (2000) and Gosset and Hyland (2000), Shillingford (2002) described some of them. These are:

- **Roaming fraud**

In this type of fraud, stolen and cloned mobile phones are used to make international calls and in roaming, possibly abroad. Once a suitable subscription has been acquired, it can be used for call selling locally or it can be used to place calls in a roaming network.

In roaming a subscriber to operator A can use operator B's network and services, provided that the operators have made a roaming agreement. Roaming, especially international roaming, and international calls in general, are usually expensive, and therefore subject to criminal interest and fraud. Roaming fraud is a hard currency problem because the roaming user's operator has to pay to the operator of the roaming network for the roaming user's use, whether or not the user pays his bills. Therefore, operators have taken measures to limit the costs of roaming fraud.

The main problem behind roaming fraud is the delay in the communication of billing information between the operators. The importance of timely communication between the roaming operators will become critical in avoiding fraud losses.

- **SIM Cloning**

Cloning is the process of replicating an existing customer's hardware or firmware, allowing calls to be made on their account. The legitimate customer will not become aware of the deception until they receive an inflated bill at the end of the month or this includes cloning prepaid cards, so one card's details can be used by 100 or so people.

- **Subscription fraud**

Subscription fraud is currently a major form of fraud. There are several forms of subscription fraud: signing up for a mobile phone service and pretending to be a nonexistent person, or some existing person other than oneself, and just being oneself but with no intention of paying the service fees. Subscriptions can also be acquired by

stealing the phones. Once the subscription has been acquired, it can be used as such or it can be used for call selling.

In some countries there is heavy competition between the operators in attracting customers. Operators also pay dealers for every subscription they sell, so some unscrupulous dealers will sell subscriptions without properly authenticating the buyer. Call selling can be done by renting the phone for a fixed sum, or by setting up a shop where customers can use it as a payphone. GSM has a few features that have been abused by fraudsters. In conference calling, more than two parties can talk to each other at the same time. Using conference calling, the fraudster acts as an operator and sets up calls for his clients by calling the client and the third party, and then dropping off the call, which leaves the client and the third party connected. After this, the fraudster may set up another call.

Call forwarding allows calls directed to a mobile phone to be automatically transferred to some other phone number. Using call forwarding the fraudster sets the forwarding to a third party, and then the client calls the fraudster's phone and is transferred to the number he wishes to call. After this call is connected, the fraudster is free to set up another call. The caller pays for the call to the fraudster and the fraudster is charged for the transferred call

- **Billing fraud**

The network operator can also defraud the customer. Most commonly this would mean overcharging the customer or, in other words, charging for services that the customer has not used. For instance, the operator might round up the durations of calls to full minutes, even when the call lasted only a few seconds.

- **Internal fraud**

This involves an employee altering the telecoms switch to give a free mobile number or numbers to friends - who use them for commercial gain.

- **Hacking**

Switches and billing systems provide administrative ports for configuration and maintenance access. Quite often fraudsters compromise these ports in order to hack the service provider systems. When they gain access they can make unauthorized adjustments to account balances, obtain voucher recharge numbers, add unauthorized numbers to toll free lists, and add privileges not originally granted for an account, such as roaming or international calls. In addition, accounts or CDRs can be modified to act as post-paid accounts, so the account balance increments every minute rather than decrements, which results in unlimited calling durations on a single prepaid wire line or prepaid wireless service.

Cao, etal(2004) classified the above various types of mobile fraud into three categories: (1) technical frauds: such as pay phone or prepaid, tumbling and magic phones, PBX feature abuse, stolen credit cards or numbers, stolen or counterfeit handsets, clip-on, cloning or home and roaming, PRS fraud; (2) subscription frauds: e.g., accounting fraud, content sells, pre-paid fraud, call sells, eavesdropping, identity theft, SIM card cloning, IP fraud, bad debt, call forward, roaming; and (3) internal frauds: for instance, ghosting, telecommunications data theft, security breach of systems, commissions on fraudulent sales, unauthorized provisioning of services.

Since every provider providing mobile services today is almost certainly experiencing some of these types of fraud as well as others that may be specific to a provider's particular network.

CHAPTER FOUR

PREPARING DATA FOR ANALYSIS AND MODEL BUILDING

4.1 Data Collection

Data is a precious stone for data mining. There is no data mining without data. Therefore, in order to get a maximum out of the data, to improve the quality of the data and to improve the efficiency and ease of mining process, the data need to be cleaned and organized. There are a number of data processing techniques. Data collection, data cleaning, deciding the right attribute, data transformation and aggregation; and data reduction are the most important activities which finally result to improve the overall quality of the target data set.

Since the data set created will determine the final result of the research work, based on the discussion with appropriate experts within Ethiopian Telecommunication Corporation, the researcher identified the main sources of data that is Call Detail Record (CDR) which is relevant for this research.

This Call Detail Record is generated by the switch machines of Ethiopian Telecommunication Corporation and it has 19 fields. The fields are all information that machine can generate with regard to the each call like the time, the calling number, the called number and duration of the call are some to mention.

Due to the fact that the raw data is very big to handle in terms of the time and space required, and is accumulated for about six months only, the researcher took two month calls for the month of February and March of 2005 as an initial data source for the research (436Mb, and 485Mb respectively). Thus, the researcher uses a file splitter (Windows Commander 32 bit international version, file manager replacement for Windows) to split and combine the data and to make the exploration more efficient.

In the process of splitting and combining the call detail record to make the preprocessing more efficient and also to select appropriate sample for the study many filtering activity has been done. First, any calls made other than an international call were excluded. Secondly, the data set of these international call records (134,086 call records or 10MB) were grouped into week level and then into day level based on their calling date. Finally, to train and make the training comfortable the numbers of records are decided by the researcher to be 30,000 call records. Thus, taking these numbers of records to be an initial data set, 508 customer call records which have high calling frequency at each day level were selected. After this tedious activity, the researcher reached 29,972 records for further analysis. For the purpose of all the above mentioned data preparation procedures, SQL server and egrep shell of the cygwin tool were also used.

In the following section, an attempt has been made to describe the nature the call detail record's data and its structure.

4.1.1. Call Detail Record

A Call Detail Record (CDR) is a single record for each call made over the telephone network. Descriptive information about each call is saved as a call detail record .This means that the CDR contains each calls information related to its telephone call such as the origination and destination of the call, the time of the call started and ended, the duration of the call, the time of the day the call was made are some of the details of the call. Since so many telephone calls are made, the number of call detail records that are generated and stored is huge.

In the application of data mining technology and in developing a model that can support fraud detection, the goal of this research is to discover the presence of illegitimate calling activity of telecommunications customers. The illegitimate calling activity cannot be observed directly, but they are reflected in the calling behavior. The calling behavior is collectively described by the call detail record, which in turn can be observed. Therefore, it is reasonable to use call detail record to apply data mining technology and to formulate model through learning and evaluate the accuracy of the model using the selected software

4.1.2 Description of the data collected

Description of the data is very important in data mining process to understand the data. Without such an understanding, useful application cannot be developed. In this section, from the source described above, the attributes with their data types and descriptions are shown in the following table.

S.N	Attribute name	Type	Description
1	PPS(Pre-paid Service)	integer	Use to distinguish different service Values: 1=PPS(Prepaid-service)
2	SubsriptionType	integer	Use to distinguish different types of subscribers. Values: 1=OTC(OneTwoCAll),3=(TEST)
3	CallType	char	Referring to the originating end or terminating end. Values:0=Mobile Calling, 1= Fixed Phone Calling, 2=Mobile Called, 3= From Mobile to Fixed Phone Calling, 4= From Fixed to Mobile Phone Calling
4	ChargeType	char	Referring to Local call or Long distance call, etc 0=International long distance call, 1=Domestic long-distance call, 2=Local call, 3=Called party roaming, 4=Free Call 5=Forwarding and charge to the calling party, 6= Forwarding and charge to the called party, 7=Charge to the opening of the voice mail box, 8= Charge to the opening of the forwarding service, 9=Forwarding to the voice mail box.
5	Roamflag	char	Roam or not. Values: 0=Not roaming,1=Roaming
6	CallingPartyNumber	char	Caller number, with area code. Example: Mobile 2519310000 or Fixed 2511505081
7	CalledPartyNumber	Text	Called number, with area code. Example: Mobile 2519310001 or Fixed 2511505050
8	RoamAreaNumber	char	Roam area number. If the Calling party is charged, the number is the area number of the Calling party. If the Called party is charged , the number is the area number of the Called party;
9	CallBeginTime	char	The Time when the call begins.
10	CallDuration	Integer	The Duration of the call in second.
11	CallCost	integer	The Unit in cents.
12	ChargePartyIndicator	char	Which Party is Charged? 1=The Calling party is charged, 2= The Called party is charged.
13	CallingAreaNumber	char	The Visiting area number of the caller. This field is of no use in Called party bills, and is sometimes not known by the system ,so generally "0000" is filled in.
14	CalledAreaNumber	char	The Visiting area number of the Called party. The visiting area number of the Called party, which can be obtained from the reported IDP message.
15	TerminationReason	integer	The Reason for the termination of the call. 1=On hook, 2=The Balance is used up.
16	Balance	integer	The Balance after the user finished the call
17	ServiceType	char	The Service type which use in this call. Value: 1=PPS
18	OriginalCalledParty	char	The Original called number of the forwarding flow
19	Discount	integer	The Discount rate of the additional service which subscriber enjoys, it does not include the time-discount during the call(100 means no discount)

Table: 4.1.1 Attributes with their description and data type

As it is clear from the above list and sample raw data (see Annex), the CDR has much information about the characteristics of each individual call. In addition to input variables

mentioned above, the output variable (target attribute) which shows the call behavior of each call is also another attribute. This target attribute has two classes: fraudulent and non fraudulent which classifies each call's behavior.

4.2 Data Pre-Processing

Data pre processing is a stage where the data set is prepared to be used by the data mining tools and techniques. It can improve the accuracy and efficiency of the subsequent mining process. It is an important step since quality decisions must be based on quality data.

4.2.1 Data Cleaning

Real-world data tend to be incomplete, noisy, and inconsistent and most of the time they have some errors, and irrelevant attributes which are not necessary for the goal of a data mining research at hand. Thus, data cleaning is just increasing the quality of the data so as to reach to meaningful result.

According to Han and Kamber (2001), the data cleaning can be used to fill missing value, ignore tuples, smooth noisy data, and identify outliers and data inconsistency.

In order to get a relevant output, relevant input should get due consideration. In line with this, an in-depth exploration of the data and frequent consultation with the domain expert has revealed that a good part of the variables were irrelevant to this specific research.

Accordingly, based on the domain expert's opinion and the researcher's own observation, attributes like PPS (since all records are prepaid, the value are «*1*»), Subscriber Type (since the records are not test, all values in the record are «*1*»), Charge party indicator (since the calling party is charged, all values in the record are «*1*»), Called area number (since there are no values into the selected record), Service type (since all records are

prepaid, the values are «*1*»), Original Called party (since the values in this field are null) and Discount(since their was no discount values/ the values in this fields are «*100*» - meaning their is no discount) were removed from the data set since one type of value has no contribution in the classification process. Moreover, Termination reason and Balance were removed since these attributes do not have significance for this research purpose.

In additions to these processes of data cleaning, the dataset with the remaining attributes were also explored to look for missing values, inconsistencies and other interpretable observations. Thus, there were missing values on Calling area indicator. Since it was very difficult to use other methods of dealing with these missing values and also there is sufficient data, such cases were discarded. Besides, a record with such missing values was irrelevant for the research problem undertaking.

There were also erroneous values in the attribute Call duration and Call cost. These inaccurate values were meaningless numbers. The exact cause for such inaccurate values couldn't be recognized with the expert of the corporation; therefore, there was no option than to ignore such records.

These data cleaning process showed a significant decrease of the total volume of data. Consequently the size of the dataset was reduced to 29,463 records

Specifically, with respect to this research, the general objective of the research work is to explore the potential applicability of data mining technology in developing a model that can support fraud detection in Mobile Services of Ethiopian Telecommunication service, the call detail record data set, records of both kinds of behavior (non fraudulent and fraudulent calls) are needed. Gathering normal call data is relatively easy as this

As a result, 9 attributes out of 10 candidate attributes were selected as the most important to discriminate fraudulent from non fraudulent call.

Accordingly, the selected attributes with their description and data type are presented as follows.

S. No	Attribute name	Type	Description
1	CallType	char	Referring to the originating end or terminating end.
2	ChargeType	char	Referring to local call or long distance call, etc
3	Roamflag	char	Roam or not.Values:0=Not roaming,1=Roaming
4	CallingPartyNumber	char	Caller number, with area code.
5	CalledPartyNumber	Text	Called number, with area code.
6	CallBeginTime	char	The time when the call begins.
7	CallDuration	Integer	The duration of the call in second.
8	CallCost	integer	The unit in cents.
9	CalledAreaIndicator	char	The visiting area number of the called party.

Table: 4.2.2 Selected attributes with their data type and description prepared for the network.

4.2.3 Data Transformation

Data transformation and aggregation involves constructive data preparation operations such as the production of derived attributes that contain new records or transformed values for existing attributes. To create the data model the researcher took the above raw data sets into the format required by the data models which is an Excel format.

The dependent attribute in this experiment was taken to be the characterization of the customers call as fraudulent or non fraudulent; thus the attribute customers call was assigned two values, 0 in the cases of non fraudulent and 1 for the cases of fraudulent.

After the pre-processing was completed, the final dataset used for modeling had 29,463 records described by 10 attributes. And with respect to the dependent variable, 20,276 are with non fraudulent and 9187 with fraudulent.

In order for a neural network to function properly, the values of independent attribute is normalized by scaling the values so that they fall within a range 0.0 to 1.0. Normalizing input values for independent attribute will help speed up the learning phase.

Thus, using linear equation(y=mx+b, where m is the Gradient and b is the y intercept), each independent values' original data are transformed into a range 0.0 to 1.0. For example, 0 and 4 are the minimum and maximum values of an attribute Call type.

The method used to calculate the Gradient is:

$$\text{Gradient}(m) = \frac{\text{Change in y}}{\text{Change in x}}$$

(x1, y1) (x1, y2) , where x1= original minimum value
(0, 0) (4, 1) x2= original maximum value
y1= target minimum value
y2= target maximum value

$$m = \frac{1 - 0}{4 - 0}$$

$$m = \frac{1}{4}$$

The method used to calculate the y intercept is:

$$Y = 1/4x + b$$
$$0 = 1/4 * 0 + b$$
$$b = 0$$

Therefore, the values in the Call type are normalized using the equation y= 1/4x, -for the original value 0, the new normalized (target) value becomes 0.

So far, in this chapter, different pre-processing tasks undertaken on the collected data has been reported in detail. And this enables the researcher to come up with clean, reduced and manageable dataset to be used for experiment, which is the focus of the next chapter.

4.2.4. Defining the Data Mining Function

Data mining functionalities include the discovery of concept/class descriptions, association, classification, prediction, clustering, trend analysis, deviation analysis, and similar analysis. The researcher's task here was to develop intelligent models that could classify customers' individual call in one of the two classes: fraudulent and non fraudulent. Each individual record in the data set is input/output pair with each record has an associated output. The output variable, the call's behavior, as described above has two classes. A supervised learning algorithm employed maps an input vector to the desired output class. Accurate results of such data analysis could provide crucial information for the identifying fraudulent call.

4.3. Model Building, Training and Evaluation

According to StatSoft, neural network have seen an explosion of interest over the last few years, and are being used successfully across extraordinary range of problem domains, in areas such as medicine, geology, physics etc. Indeed, anywhere that there are problems of prediction and classification, neural networks are being applied. This sweeping success can be attributed to its power and ease of use. Neural network are very sophisticated modeling techniques capable of modeling extremely complex functions. It learns by examples.

Therefore, in this research work, the data mining task was undertaken by using Artificial Neural Network. The type of training used in neural network is supervised learning. As it was discussed above, a set of training data is assembled. The training data contains inputs together with the corresponding outputs. The specific neural network software that is used for model building and testing purposes is BrainMaker software. Thus, before going

to the details of specific steps that are carried out in this study, the researcher would like to give an overview on BrainMaker software.

BrainMaker Neural Network software is developed by California Scientific Software. This software uses back propagation algorithm in developing neural network model. The network is trained by presenting a set of facts (records) over and over again. BrainMaker goes through all the training lists (records) addressing each fact (record) in turn and making necessary corrections. After the entire list of facts has been presented (when one epoch or one run is completed), BrainMaker starts again from the beginning of the list. The training process is repeated until the network gets all facts (records) correct or until training is interrupted.

4.3.1. Data Organization for Model Building

At this stage, it was important to organize the data into a format suitable for training, or model building in general. For this purpose BrainMaker Neural Network software has two programs called NetMaker and BrainMaker.

Netmaker reads data from Lotus, ASCII, Excel, Text, dBase or Binary files and creates all of the files BrainMaker needs for training and testing. NetMaker can also manipulate the imported data, calculating moving averages, differences, etc. Both numeric data and text data can be accepted by Netmaker and converted in to a representation that the neural network can understand. Therefore, the whole dataset can be imported into NetMaker; and then the input fields (independent variables) and pattern fields (dependent variables) are determined and labeled. The prepared file is then saved with a .dat extension.

NetMaker will create the necessary neural network training and testing files from this data.

The BrainMaker program has the facility where data are classified into training and testing sets. By default, the software puts aside 90% of the records and 10% of records into training set and test set respectively. Thus, the above *. dat file is the basis to create the three brain maker files.

The first BrainMaker file is the definition file (*.def) that has the definition information for training such as what columns are inputs and patterns, and how the information is displayed. The second BrainMaker file is the fact file (* .fct), which by default, constitutes 90% of the prepared data for taining. The third one is the test file (*.tst), which by default, constitutes 10% of the prepared data for testing. We can also create running fact file (*.in) to predict future records.

Thus, as it is mentioned above, to build the neural network model, the 29463 total numbers of records in the dataset is imported into NetMaker program. After importing the data set into NetMaker, the input fields and the patterns are labeled and saved with a .dat extension. This data extension file contains all records and therefore, by default, the software puts aside 90% of the records (26516 records) and 10% of records (2946 records) into training set and test set respectively. Reserving test set helps to add confidence in the performance of the final model. The final model is going to be tested with the test set of data to ensure the results on the training set are real and not artifacts of the training process.

Following splitting the data set, BrainMaker files are created. These are the definition file (*.def) that has the definition information for training such as what columns are inputs and patterns, and how the information is displayed, the fact file (* .fct) constitutes the 90% of the prepared data for training, the test file (*.tst) constitutes 10% of the prepared data for testing and the running fact file (*.in) to predict future records.

4.3.2. Creating and Training the Network

After the above files are created the model developer moves from Netmaker program to the Brain Maker Program.

The Brain Maker Program has different parameters with different possible values for each parameter that are essential in neural network training and testing. The most important parameters are training tolerance, learning rate, smoothing factor (momentum), number of hidden layers, number of neurons in hidden layer, type of transfer function. After determining these parameters, training is started by using the Operate/Train Network command. While training progresses statistical information are provided on the screen such as which fact the BrainMaker is processing at a specific time, the number of facts which met and did not met, the training tolerance, the number of run (epoch) etc.

There are also two graphs that display the progress of training. The first is a histogram that shows the distribution of error over an entire run. The horizontal axis represents the error level and the vertical axis signifies the number of out put values at that particular level. As training progresses and fewer facts are classified as incorrect, the bars (solid boxes) move to the left. The second graph shows the progress of the error rate as network trains. In this graph, the horizontal axis shows the number of runs while the vertical axis

represent the over all error level (root mean square error, RMS). For a good training, the value of RMS would decrease as the number of runs increases as shown below.

Training can be stopped at any time before the instruction to stop training is met. The default for stopping training process is when the incorrect classification of facts in a single run (epoch) is zero. Better network model can be obtained before the criteria for stopping training are met. Hence, it is advisable for model developer to test and save a network model periodically.

Once the above files are created, the experiment is conducted using BrainMaker program. In creating the network the different parameters in the selected training function were set.

Initially, training was started based on the default network parameters. The default parameters for the network are: learning rate=1.0; smoothing factor=0.9; training tolerance=0.100; testing tolerance=0.400: training noise=0.00; number of hidden layers=1; number of neurons in hidden layer=10; type of function=sigmoid are some to mention a few. The overall performance of the first test was encouraging. The constructed model, which is by using the above default values, predicts 24466 correctly out of 26516 call records. Thus, the model works with an accuracy of 92.27%. After 20 runs of facts (epochs), there is no improvement in the training process because the distribution of errors and the root mean square error value stopped to decrease.

Since the test result of the above network model was encouraging, the researcher continued the experiment by considering various options suggested to improve the performance of neural network models. For example, the vendors of Brainmaker software suggested that when a network model got a sufficient number of training facts correct and

performing fairly well in testing process, try to build another network model by reducing number of nodes in the hidden layer, by adding noise to the network, by adding number of hidden layers, by shuffling the records, by changing training tolerance and learning rate etc (California Scientific Software, 1998).

Second attempt was made whether additional nodes in hidden layer (twelve nodes) in the network would improve the accuracy of the network by keeping the default values of the number of hidden layers and other parameters constant. After 24 runs of facts (epochs), this network model also fails to learn all the training facts, however the accuracy was slightly increased in comparison to the above network. It predicts 24901 records correctly out of 24516 and the model works with an accuracy of 93.91%.

Third, by keeping the default value for the number of hidden layer and other parameters constant but increasing the number of hidden nodes to fifteen, the accuracy becomes 94.34%.

Fourth, by keeping the number of hidden layers one and number of nodes 17, the default values of training tolerance and learning rate (the most important parameters) were changed in to 0.300 and 0.700 respectively. But in this experiment, the default value of smoothing factor 0.900 had not been changed because the providers of this software stated that adjusting the smoothing factor has not been found to reduce training time or improve prediction power of the network in every case. After 12 runs of facts, this network model starts to rise. This is an indication that the network is starting to over fit the data, and therefore training was ceased.

Since model building is an iterative process, the researcher forced to build different models by changing the different parameters. Lastly, from these models, based on relative accuracy level, the three model's parameters summarized below.

Model	Learning Rate	Training Tolerance	Hidden Layers	Nodes in Hidden Layer	Accuracy on training dataset
One	1.000	0.900	1	10	92.71%
Two	1.000	0.900	1	12	93.91%
Three	1.000	0.900	1	15	94.34%

Table 4.3.2 Parameters for the three models

The model which shows the best accuracy level is shown in the following figure.

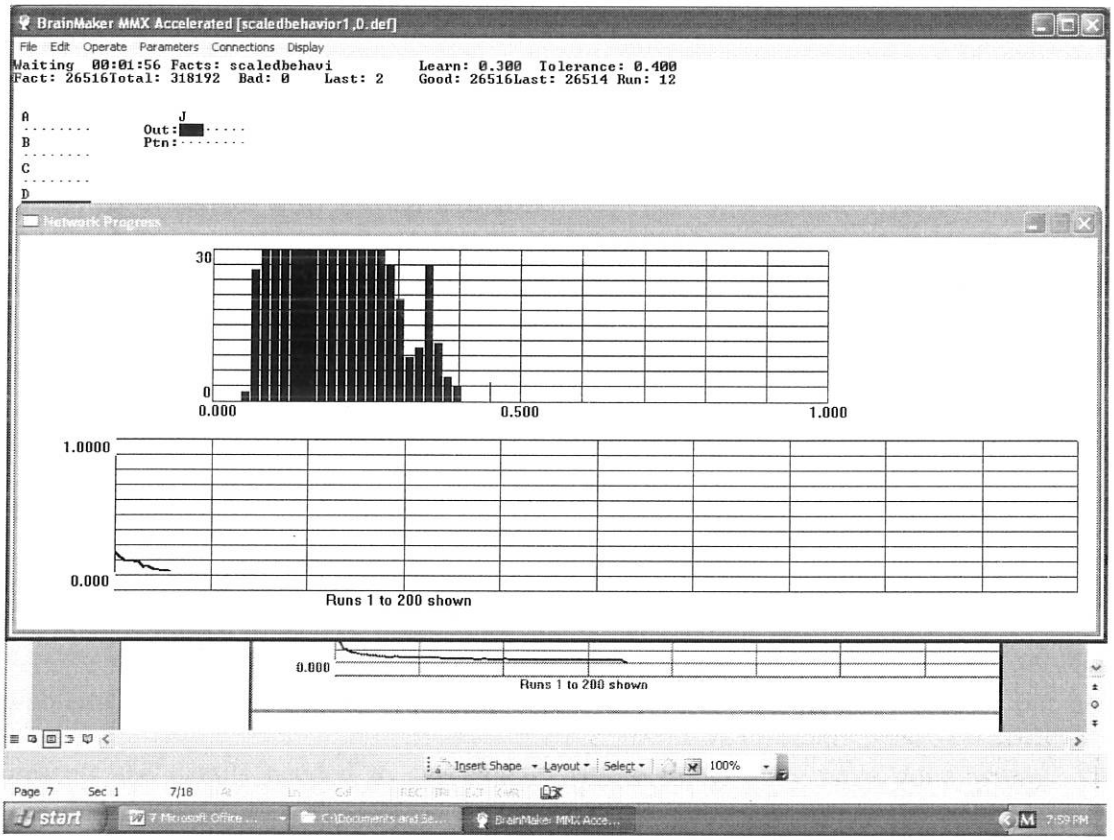

Figure 4.3. The input, and display formats, network progress display for the highest accuracy model.

Model		Predicted		Total	Score
		Non fraudulent(0)	Fraudulent(1)		
One Actual	Non fraudulent (0)	1900	127	2027	93.75%
	Fraudulent(1)	153	766	919	83.35%
	Total	2053	893	2946	
Two Actual	Non fraudulent(0)	1975	52	2027	97.43%
	Fraudulent(1)	128	791	919	86.07%
	Total	2103	843	2946	
Three Actual	Non fraudulent(0)	1936	91	2027	95.51%
	Fraudulent(1)	106	813	919	88.46%
	Total	2042	904	2946	

Table 4.3.3: Confusion matrix for Model One, Two and Three

From the three models the best model was selected taking into consideration of maximizing the accuracy for fraudulent calls and minimizing the false alarms i.e. falsely predicting the non fraudulent as fraudulent. Model three was the best since it has high accuracy for fraudulent calls and relatively moderate error rate for non fraudulent calls (88.46% Accuracy for fraudulent and 4.49% Error rate for non fraudulent calls).

CHAPTER FIVE
CONCLUSION AND RECOMMENDATION

5.1. Conclusion

Information and communication technologies (ICT) pursue two main goals: processing information (i.e. presenting it in various forms, storing it, searching for it, reproducing it, etc) and transmitting information from one geographical point to another, from one person to another, to a group of people or to the whole community.

The second half of this century has witnessed the global phenomenon of an information explosion. Telecommunications are one of the most important of the contemporary ICTs and which have huge amount of data. The new generation of computerized methods is also helping the ventures in analyzing this very large datasets automatically and efficiently, thereby extracting useful information and knowledge which are valuable for decision making. The telecommunications industry has been one of the earliest adopters of data mining technology, largely because of the amount and quality of the data that it collects. Thus, data mining offers great advantage to the telecommunication corporation helping them uncover patterns hidden in the data that can be used to predict the behavior of customers.

The objective of this research was to explore the possible application of data mining technology in supporting fraud detection on Ethio-Mobile service by developing a model. To achieve this objective, literature review was conducted to get detail understanding about the concept raised, the data was identified i.e. the call detail record which has on the average 460.5 MB for a single month, this data was prepared using different preprocessing techniques, and model was built and evaluated.

In order to support the fraud detection several models were built. The best performing model was chosen taking into account its accuracy for fraudulent calls and error rate for non fraudulent calls. Such a classification model could support in minimizing fraudulent activities, increase the corporation's profitability and thereby improve efficient management decisions

To conclude, results from the study have shown that detection of mobile fraud in Ethio-Mobile service could be supported by the use of data mining, especially with the use of neural network.

5.2. Recommendation

On the basis of the results obtained by the methodologies applied on the collected data, the following recommendations are forwarded:-

- Data mining techniques could contribute a lot in identifying potential fraud customers/calls. Therefore, it could be more important to use data mining technique as a tool for the detecting mobile fraud.
- In this research work, an attempt has been made to assess the possible application of data mining technology to support mobile fraud detection by using some set of attributes that were considered important by experts. However, additional research could be done by including other attributes of the call detail record so as to build models with better performance and accuracy than the models built in this research work.
- Although there is an assumption, by experts at the corporation, about the cause of on the types of fraud that was found on call detail record as an error on the switch

machine, the amount is found to be reasonably high. Therefore, it needs further investigation.

- The fraudsters are very skilled with regard to "flying under the radar." They figure out the rules and the thresholds set in the detection systems, and they simply modify their behavior to proceed with their criminal activity undetected. Therefore, though the call detail record detail has an incredibly large size of data, it needs to be explored frequently in depth. Besides, fraudsters adapt to new prevention and detection measures, so the Corporation's fraud prevention mechanism should be adaptive and evolve over time.

REFERENCES

Abbot, D. et al. (1996). An Evaluation of High-End Data Mining for Fraud Detection.. Available At URL:http://www.kdnuggets.com

Andreescu, A. and Zilliacus, J. (2002). Data Mining: Applications for Telecom Operators. Available At URL:http://www.pafis.shh.fi/~andand02/workshop/html

Berry, M. J. and Linoff, G. (2000). Mastering Data Mining: the Art and Science of Customer Relationship Management. New York: John Wiley & Sons, Inc.

Berry, M. J. and Linoff, Gordon..2000. Data Mining Techniques for Marketing, Sales, and Customer Support. New York: John Willy& Sons, Inc.

Bolton, R. J. and Hand D. J. .2002. Statistical Fraud Detection: A Review Department of Mathematics. Imperial College. London. CUP.

Brand and Gerritsen(1998) DBMS, Data Mining Solutions Supplement. Available At URL: http://www.dbmsmag.com

Cao etal.(2004).Hybrid Strategy of Analysis and Control of Telecommunications Frauds. Proceedings of the 2nd International Conference on Information Technology for Application. Available At URL: http://www.staff.it.uts.edu.au

Carbone.,Clifton and Thuraisingham.(2001). Data Mining. Newsletter Available At URL:http://www.kdnuggets.com

Carey, B. and Collier, K. (1999). A Methodology for evaluating and Selecting Data Mining Software. Center for Data Insight. Northern Arizona University. Available at URL: http://www.insight.cse.nau.edu

CEREBRUS .(2002) .An advanced fraud detect ion system that places the raud analyst at the heart of the fraud management operation. Available At URL: http://www.CerebrusSolutions.com

Chung M., Gray P. and Michael Mannino. (1998). "Introduction to Data Mining and Knowledge Discovery" in Information Systems Research Vol. 6, No.4, 1995, pp 328-356.

Computer Desktop Encyclopedia(2000).

Cortes and Pregibon(1998). Giga-mining.Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining; 174-178, 1998 August 27-31; New York, NY:

Data Mining Explained. Available At URL: http://www.kairon.com

Data Mining Telecommunication Network: data for fraud management. Available At URL: http://citeseer.nj.nec.com/sterrit00data.html

Ethiopian Telecommunication Corporation Available At URL:http://www.telecom.net.et

FairIsaac(2003).Prepaid Telecommunication Fraud Techniques and Detection. Available At URL: http://www.fairisaac.com

Fayyad, U. (2003). Optimizing Customer Insight. Intelligent Enterprise Magazine. Available At URL: http://www.dmreview.com/potal/cfm/html

Financial Report of the Finance Department of ETC. (2003). Unpublished

Funk Wagnall's New Standard Dictionary of the English Language. (1963).New York: Funk and Wagnall Publishing Ltd

Gandy, Oscar .(2002). Data mining and surveillance in the post-9.11 environment. Available At URL:http://www.asu.upenn.edu

Gashaw Mulatu (2004). Application of Data Mining Technology to Support Customer Insolvency Prediction at Ethiopian Telecommunication Corporation. Unpublished Master's Thesis. Addis Ababa University. Addis Ababa.

Giudici, P. (2003). Applied Data Mining. New York: John Wiley & Sons, Inc.

Gray,P. (1998). The New DSS; Data Warehousing,OLAP, MDD, and KDD. Tutorial, Thirty First Hawaii International Conference on Systems Sciences, Kohala Coast, Hawaii, January 6-9, 1998. Available At URL:http://www.csdl.computer.org

Han, J., and Kamber, M. 2001. Data Mining: Concepts and Technologies. Available At URL :http://www.cs.sfu.ca

Holmen and Tresp(1998) Call-based Fraud Detection in Mobile Communication Networks using a Hierarchical Regime-Switching Model Available at URL http://www.brauer.informatic.tu-muenchen.de/~trespuol/papers/nisp_books.pdf

Holmen(2000) User Profiling and Classification for Fraud Detection in Mobile Communications Networks. PhD thesis, Helsinki University of Technology.

Hynninen,J(2003) Experiences in Mobile Phone Fraud. Available at URL http:www.Hynninen@hut.fi

Jaccobs, Riaan(2003).Telecommunication Fraud. Available at URL http://www.dimensiondata.com

Jember Gebrselasie (2004). Data Mining Application in Supporting Fraud Detection on Mobile Communication: the Case of Ethio-Mobile. Unpublished Master's Thesis. Addis Ababa University. Addis Ababa.

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

Kein, Daniel .(2002). Information Visualization and Visual Data Mining. Available At URL:http://www.portal.acm.org

Kou etal.(2000).Survey of Fraud Detection Techniques. Available At URL:http://www.lee.org

Kvarnstrom,Lundin and Jonson. Combining fraud and intrusion detection meeting new requirements. Available At URL:http://www.ce.chalmers.se

Levin,N.and J.Zahavi.(1999).Data mining.Available At URL: http://www.kdnvggets.com

Luan,Jing .(2004). Data Mining Applications in Higher Education. Available At URL: http://www.spss.com.

Lundin,E (2002). Aspects of employing fraud and intrusion detection systems. Available At URL:http://www.cs.kau.se

Mitchell, M. (1997). Machine Learning. New York: The McGraw-Hill Companies, Inc.

Moxon, B. (1996). Defining Data Mining. California. Miller Freeman, Inc.

Nikolaj Lindberg. (2001). egrep for Linguists. Available At URL: http://www.ida.liu.se

Nilsson Nills. (1996). Introduction to Machine learning. Available At URL:http://www.robotics.stanford.edu

Quinlan, .J (2003). Programs for Machine Learning, Morgan Kaufmann.

Seymour, B. (2002), "How Neural Network Technology can Tackle the Growing Telecom Fraud Problem" Information Security Bulletin, CHI Publishing Ltd., UK

Shapiro, G. (2000). From Data Mining to Knowledge Discovery in Databases. Available At URL: http://citeseer.nj.nec.com/fayyad96from.html

StatSoft.(2003). Neural Networks. Available At URL:http://www. Statsoft.com

Two Crows Corporation. (1998).Introduction to Data Mining and knowledge Discovery Third Edition. Available At URL:http://www.twocrows.com

Weiss, G, et al (1998). Intelligent Telecommunication Technology Available At http://www.research.rutgers.edu/~gweiss/papers/jain98.pdf

Weiss, Gary (1998). Data Mining In Telecommunications. Available At URL:http://www.citeseer.1st.psu.edu

West,J(2004). Mutual Value Enhancement through Data Mining. Available At URL:http://www.wvu.edu

Declaration

I, the undersigned, declare that this thesis is my original work and has not been presented as a partial degree requirement for a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.

Geberemeskel Girma

March 2006

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

Dr. B. L. Desai

March 2006