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

DATA MINING APPLICATION IN SUPPORTING FRAUD DETECTION ON MOBILE COMMUNICATION: THE CASE OF ETHIO-MOBILE

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTERS OF SCIENCE IN INFORMATION SCIENCE

BY
JEMBER GEBRESELASSIE

JANUARY 2005

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

Faculty of Informatics
Department of Information Science

DATA MINING APPLICATION IN SUPORTING FRAUD DETECTION ON MOBILE COMMUNICATION: THE CASE OF ETHIO-MOBILE

BY

JEMBER GEBRESELASIE AYALEW

Name and Signature of Member of the Examining Board

Ato Nigussie Tadesse, Chairman, Examining Board	Nigussie Tadesse
Ato Tesfaye Biru, Advisor	
Dr. B.L.Desai, Member	B.L.D
Dr. Lishan Adam, External Examiner	Lishan Adam

Getachew Temaneh	Getachew Temaneh	22/02/05
Chairman, Faculty	**Signature**	**Date**

Chairman, Graduate Council	**Signature**	**Date**

ACKNOWLEDGEMENTS

It will be unlikely to finish this research work with out the support and encouragement of many people during the course of this research work. First and for most I would like to thank my families for their understanding, full support and help.

My special thanks go to my advisor Ato Tesfaye Biru for his willingness to be my advisor at the time when it is hard to find one.

I also would like to express my sincere gratitude for Ato Workshet Lemenew and Ato Sirak Alemayehu for their valuable advice, support and push to complete this research successfully.

This work would have been impossible with out the full support of the staffs of Ethiopian Telecommunication Corporation specially those in the Mobile division, with special credit to Ato Wondimu Girma , Ato Alias Terefe, and Ato Dereje Getachew , and all others who helped me during the research work.

Finally, thanks also to my friends with out their day to day push and encouragement I couldn't finish this research. And of course I would like to thank all others who helped me in completing this study.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ABTSRACT

The application of data mining methods and tools that can help to explore large quantities of data generated by the Call Detail Record (CDR) of telecommunication switch machine will be the art of the day to address the serious problems of telecommunication operators. The CDR consists of a vast volume of data set about each call made and it is a major resource of data for research works to find out hidden patterns of calls made by customers in addition to the typical use for bill processing activities.

More importantly, data mining technology has enabled telecommunication companies to utilize this generic source of data for different kinds of customer relationship management and marketing activities including fraud detection and prevention strategies.

The methodology used for this research had three basic steps. These were collection of data, data preparation and model building and testing. The required data set was selected and extracted from the billing data set of Ethio-mobile. Neural network data mining technology were employed to build and test the models. The data mining method used in this research work have proved to yield comparably sufficient results for practical use as far as supportive mechanisms are employed for the misclassification of fraudulent customers as non-fraudulent and vice versa. Due to this fact the telecommunication operators should take serious and intensive follow ups for unusually high frequency and long duration of international calls made by some possibly fraudulent customers. However the selection of representative sample data for the whole database of the problem under consideration needs a series consideration and care to properly select the training data set. Otherwise due to the rare occurrences of possibly fraudulent calls, the training of representative data set will be a tedious activity.

In this research work, the researcher has proved that the CDR is a major resource of crucial knowledge about customers of Telecommunication Corporation. Beside that the number of calls made and the duration of each call should be traced to properly know unhealthy customers of the corporation. Finally it is a rich field for research for other interested researchers to make further and in-depth research on this area.

CHAPTER ONE

INTRODUCTION

1.1 Back ground

In today's world, the telecommunication industry has quickly evolved from offering local and long distance telephone services to providing many other comprehensive services including voice, fax, pager, cellular phone, images, e-mail, computer and web data transmission and other data traffic. The integration of telecommunication, computer network, internet and numerous other means of communication and computing is also underway. Moreover, with the deregulation of the telecommunication industry in many countries and the development of new computer and communication technologies, the telecommunication market is rapidly expanding and highly competitive [1].

The principal sector of the telecommunication industry is telephone communications. Establishments in this sector operate both wire line and wireless networks. The cellular technology exploded over the last few years providing telecommunication any time and anywhere. Besides being very exciting and profitable business, wireless and mobile communications has been an extremely rich field for research, due to the many difficulties that the wireless environment presents and due to the ever increasing user's demand for more, newer and better services.

1.1.1 Ethiopia Scenario

The introduction of telecommunication in Ethiopia dates back to 1894.In those early years, the new technological scheme contributed to the integration of the Ethiopian society when the extensive open-wire line system was laid out linking the capital with all the important administrative cities of the country [2].

Starting from that dates, the organization had undergone through series 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.

ETC's development programs are not only meant to expand and improve the telephone, tele fax and the other relatively old types of services to the rural and urban areas. Though the various transmission systems, plans were to provide Internet and telemedicine as well as Interactive Distance Learning access to more than 10 regional towns, including higher education institutions with many colleges in the Regional states located far from the center of the country.

Currently the corporation offers service such as telephone, telex, tele fax, internet, cellular mobile telephone, data communication and the like.

Based on the report from ETC, the provision of GSM mobile telephone service had began in Addis Ababa in April: 1999.In Addis Ababa, the capital city of Ethiopia, the number of mobile subscribers was about 51777 as of June, 2003. In spite of the fact that the repot is not yet released it is estimated that more than 100,000 subscribers are now become the users of mobile phone in Addis Ababa, Nazareth, Jimma, Bahir Dar and the neighboring cities. The service coverage has expected to reach 200,000 subscribers in Addis and other major towns of the country according to the development plan of the corporation with short time period [2].

1.1.2 GSM-Mobile Telephone Service

Mobile telephone is a service with which you can access both the fixed and mobile networks when you are within the coverage area of the GSM network. It gives access for social, business and emergency calls virtually twenty four hours a day. GSM (Global System for Mobile communication) network is a digital cellular mobile telephone system which is originally adopted in Europe. It has been commercially available since 1992 and has penetrated the world's cellular market ever since. To day the number of operators drastically increases in over 109 countries by having millions of subscribers throughout the world according to the information released from Ethiopia Telecommunication Corporation [2].

Ethio-Mobile presently operates in the GSM 900 MHZ frequency ranges covering Addis Ababa, Debre Zeit, Nazareth, Modjo and Sodere. Supplementary services available at

extra charges, which includes the value added services such as Voice Mail, Call Forwarding, Call Barring, Call Waiting and Advice of Charge.

1.2 Statement of the Problem and Its Importance

The underlying research problem that necessitated this research is the existence of high level of fraud in telecommunication industry in general and more importantly the high level of uncollectable mobile bills in Ethio-Mobile service division of Ethiopian Telecommunication Corporation with in few years after its introduction.

More specifically the problem of telecommunication fraud is a very painful one causing massive damage to telecommunication companies worldwide. Based on the review of Bolton R. and Hant D. [3].

> *Telecom fraud is a global phenomenon. Cox et al (1997) give a figure of $1 billion a year. Telecom and Network Securely Review also (vol. 4(5), April 1997) gave a figure of between 4% and 6% of US telecom revenue being lost due to fraud. Call et al (2002) suggests that international, figures are worse, with several new service providers reporting losses over 20%. Moreau et al (1996) give a value of 'several million ECUs per year'. According to a recent report (Neural Technologies, 2000) 'the industry reports a loss of £ 13 billion each year due to fraud. Mobile Europe (2000) gives a figure of US $ 13 billion. The latter article also claims that it is estimated that fraudsters can steal up to 5% of some operators' revenues, and that some expect telecom fraud as a whole to reach $ 28 billion per year within three years.*

In the same token, in Ethiopia, based on the information from ETC, there are more than 11 million birr uncollected bills from post-paid mobile phones only in the past five years [4].

Despite the variety in these figures, it is clear that they are all very large. Apart from the fact that they are simply estimates, and hence subject to expected inaccuracies and variability based on the information used to drive them, it is clear that fraud causes substantial losses to telecommunication industry.

Thus to overcome the current high level fraud in the industry, an effective means of controlling and preventive mechanisms should be employed. But due to the nature of the technology it self and the restless human being discovery of new types of fraud at any time, preventive methods alone will not be successful. So fraud detection is necessary to take any sound and strong action on fraudulent customers before they create a huge loss in the corporation earnings and also in the country. Especially in a country like Ethiopia where the gross annual income of the population is less than one hundred dollars per annum (which is absolutely less than the poverty line), fraud and misappropriation of public resources will not be only a financial loss for the country but also a major calamity and dishonorable act that cannot be tolerable at the back of the poor people.

It is in this context that this research has sought to assess and experiment the potential applicability of data mining technology in supporting fraud detection activity in the mobile telephone line.

1.3 1.3 Objectives of the Research undertaking

1.3.1 General objective

The general objective of this research is to explore the potential application of data mining in supporting fraud detection in mobile telephone service of Ethio-Mobile.

1.3.2 Specific Objectives

In order to achieve the above stated general objective, the following specific objectives are formulated.

- To review literature on application of Data Mining in fraud detection
- To select and extract the data set required for analysis from the large volume Call Detail Record (CDR).
- To identify an appropriate Data Mining algorithm and software that would do the main task of the research project.
- To build and train a model
- To evaluate/test the model
- To report the result and forward recommendations for further study.

1.4 Research Methodology

For the purpose of this research undertaking the researcher has opted to use the methodology suggested by Berry and Linoff [5].This methodology assumes that business

problem has already been identified and hence directly proceeds to the different data mining steps that need to be carried out in order to develop a model for the data mining project.

The different steps suggested for a data mining project and how they are applied to the current research project are provided below:

1.4.1 Identifying Source of Pre classified Data

The sources of data were identified based on the discussion with the staffs of the mobile service division of ETC. As a result the researcher identified three main sources of data stored in electronic and manual documents. The first and major source of data was the Call Detail Record (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. This data source have billions of records which is stored in the form of binary digits for a period not more than six months due to the high space required to maintain it. The second source of data was the billing data kept by the IT and Data Service Department of the corporation for billing purpose. This data is a conversion of some important and relevant fields from the CDR for billing customers for the services rendered by the Corporation.

The third source of data for this research was the Customer Data Base maintained by the Finance Department for follow up purpose after bills are returned back from cash collection desks through out the regional offices of the corporation. This file shows the

status of those customers who were not paying their bills and also actions taken by the corporation to claim its service fees from the customers.

Thus for this research work three month data for October and November of year 2003 and March of year 2004 were taken from the IT and Data Service Department. More over customer data set about customers who were not paying their bills for their service in the month of October and November was collected from the Finance Department. Unfortunately due to late distribution of bills and delay in reporting by cash collection desks of Zone Offices black list customer data sets for the month of March 2004 was not available. Due to this fact data collected from IT and Data Service Division about the call detail of March is excluded from the database collected for analysis.

1.4.2 Preparing Data for Analysis

Data to be mined will be collected in a new data base. This will help to apply data mining tools and algorithms on the data. As such, the collected data were cleaned in to a form that was suitable for the particular neural network software used. The data mining software techniques used for this research project were neural network. Therefore, different neural network soft wares were examined by taking into consideration their application to the problem and availability to work with them during the research period.

Thus, in this research, the researcher selects Matlab Software which has a neural network toolbox necessary for this research undertaking. In preparing data for analysis, the

collected data were summarized, inconsistent data were encoded, missed values were accounted for, and new fields were derived from the existing ones.

1.4.3 Build and Train the Model

Although 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 in to training, validation and testing before building the model and the data was classified in the proportion of 6:2:2 for training, validation and testing purpose respectively.

Based on the data available for this research numerous networks (models) were built by using Matlab software and the performance of those models were tested by using the test data set put aside by the software for this purpose.

1.5 Scope and Limitation of the Study

The scope of this research is to appraise the potential applicability of data mining in supporting fraud detection and prevention activities of ETC. While findings of the research work can fairly be considered as relevant in appraising the potential applicability of data mining technology in the ETC at large, the current scope of the experimental research undertaking is strictly limited to appraising the possible application on data mining technology to support fraud detection in mobile communication network more specifically for the post-paid mobile phone.

The time that was given to undertake this research work and the available data in ETC was a serious limitation in developing an appropriate model for this research. Besides, obtaining the data mining software needed for this research work was a challenging task. Particularly, the efforts made to search and select more appropriate and affordable data mining software that can be used to build and test models both in neural network and decision model was unsuccessful. So that to assess the applicability of data mining technology on fraud detection only neural network is employed by using the Matlab software. More significantly the unavailability of properly organized customer database in ETC and shortage of time to undertake the research limits the research to focus only on the subscription (accounting) fraud type only.

1.6 Organization of the Thesis

This thesis is divided in to five chapters. The first chapter is an introductory part, which contains background to the research work, statement of the problem addressed, objective of the research, and methodologies adopted for the study.

The second chapter deals with data mining technology, methods/techniques used and its application for fraud detection purpose.

The third chapter is devoted to give further understanding to the problem area by detail and depth investigation of the type of frauds exist in the telecommunication industry in general and mobile telephone in particular.

The forth chapter provides discussion about the different data mining steps that were undertaken in this research work. This includes data collection, data preparation, model building and testing results obtained by using Matlab neural network toolbox.

Finally, the last chapter is devoted for the concluding remarks and recommendations forwarded on the basis of the research findings.

CHAPTER TWO

DATA MINING TECHNOLOGY

In this chapter, an attempt has been made to review the literature on the concepts and techniques of data mining in general and its application in telecommunication industry.

2.1 Introduction

An important source of knowledge is the data stored in databases. Data allows us to learn from the past and to predict the future. With rapid computerization of business and organizations, a huge amount of data has been collected and stored in data bases, and the rate at which data is stored is growing at a phenomenal rate. As a result traditional ad hoc mixtures of statistical techniques and data management tools are no longer adequate for analyzing this vast collection of data. In stead, researchers begin to look for ways to intelligently assist humans in analyzing these mountains of data[3].

Data mining or Knowledge Discovery in data bases has recently emerged as a growing field of multi disciplinary research for discovering interesting/useful knowledge from large data bases.

Data mining combines research areas such as databases, machine learning, artificial intelligence, statistics, automated scientific discovery, data visualization, decision science, and high performance computing. While each of these areas contributes in its specific way data mining focuses on the value that is added by creative combination of

the contributing areas in to produce innovative solutions to the data analysis task [1], [6], and [7].

It has now been recognized that mining for information and knowledge from large data bases and documents will be the next revolution in data base systems. It is considered an important area for many cost savings and potential revenue with immediate applications in business, decision systems, information management, communication, and scientific research and technology development. We can expect the generation information systems to be more intelligent in that they are not only data intensive but also knowledge rich [8].

Every mature business has some seemingly insoluble problems arising from difficult judgments about customers, resource allocations, business strategies or organization. These are according to Heinemann [8] the "5 percent problems" .The intractable residue that remains after years of work have solved the other 95% is the principal targets of data mining. Data mining techniques can be used to learn the factors bearing on a decision and construct an application that uses those factors to help the enterprise make those decisions in an objective and consistent way.

In a sense, on the view of Heinemann, most of today's enterprise problems are no more difficult than those of fifty years ago [8]. Managers in the 1950 faced many of the same problems that decision- makers face today. But, there is one very important difference- scale. In terms of the number of customers, variety of products, array of marketing channels, speed of commerce, churn, fraud, etc, everything today is much bigger. So big,

that unaided human decision- making processes are losing their ability to keep up. There may be nothing wrong with the process themselves, other than their inability to scale up. Further complication the modern decision maker's problem is a reduction in the time available for deliberation. Fifty years ago, slower communication and distribution channels gave managers time that they just don't have today. Enterprise management is a sequence of decisions. Managers must evaluated current conditions in light of past experiences and future expectations and choose among the available courses of action that will lead to the factors bearing on a decision and construct an application that uses those factors to help the enterprise make those decisions in an objective, consistent way.

As managers grow in experience, they develop and refine decision making methods well suited to their enterprise. These methods formalize important knowledge about how the enterprise operates. Good decision makers are often successful because of the knowledge they posses. If intelligence is the engine, then knowledge is the fuel [8].

An enterprise operations become more complex, however, decision makers all forced to make compromises in the application of these methods simply because there is a practical limit to the number of factors the human mind can objectively consider. The fine distinction being made here between knowledge and intelligence is important to achieve an accurate understanding of what data mining does.

2.1.1 What Data Mining Does

Computers are not at all intelligent, but through the application of data mining and other intelligent software techniques, they can discover, store, and apply knowledge. Since

knowledge is at the heart of effective decision-making, data mining techniques for the discovery and exploitation of knowledge can aid humans in many aspects of enterprise management. Computers are easily capable of correlating 50 subtle factors as part of a decision- making process. This is something that human beings, no matter how intelligent, simply cannot do.

Data Mining discovers enterprise knowledge from historical data and combines it with data relating to current conditions and goals to reduce uncertainty about enterprise outcomes.

In practice data mining has components: discovery and exploitation. During the discovery component, enterprise facts are discovered and represented as information- bearing data. During the exploitation component, these enterprise facts are applied to the solution of a business problem. First we discover, second we act. Neither phase makes sense with out the other [5].

2.1.2 Data Mining and Major Business Problems

Data Mining is a special type of processes that can help to solve the major problems of the current business world. As a matter of fact, Data mining is most often used to help discover relationships, make choices, make predictions and improve processes [5].

2.1.2.1 Discovering relation ship

There is a well-known story about a large retailer who conducted a data mining experiment by looking at thousands of register receipts to discover which items people bought together. This kind of analysis, which looks for concurrent events, is sometimes

called "market basket" analysis after this seminal example. The technical name is 'link analysis".

The application of link analysis to direct sales is pretty clear. Elective products are offered to customers who buy other products frequently "linked" with them. History indicates that there is a high likelihood of making additional sales.

2.1.2.2 Making Choices

A decision is a choice among alternatives available right now. In the business world, decision makers have to choose between expensive materials and inexpensive materials, valid transactions and fraudulent ones, and even good and bad customers. For example, decisions must be made regarding which customers will be the beneficiaries of the allocation of scarce resources (technical, support and otherwise). Here, data mining can be used to evaluate the data available for making a decision, apply a classification technique, and suggest or prioritize the "best" choice(s) among the available alternatives (For example, the list of customers to be the focus of retention programs).

Customers who are already standing at the check – out counter are virtually guaranteed to buy. Data mining can suggest choices that optimize the sale: customers who buy ABC and DEF will get the same or better value from XYZ, which has a higher profit margin. Knowledge like this can facilitate proactive selling that customers are likely to regard as a source rather than an annoyance.

2.1.2.3 Making predictions

A prediction is a choice among alternatives available in the future. Predictions often anticipate future behaviors of individual customers, such as whether a customer is one

who will pay his bill. Predictions can also anticipate market level behavior, predicting the total market demand for a certain product in a future time frame. Predictive models can be used to extrapolate past history, in combination with current conditions, to predict future situations. History is the best crystal ball.

2.1.2.4 Improving the process

Businesses are immersed in a complex system consisting of supplies, communications infrastructure, regulatory and market pressures, and other factors, many of which are impossible to quantity. Successful businesses have optimized their processes for interaction within this system. The effectiveness of their processes will make or break them. Data mining can be used to reveal aspects of business processes that are sub-optimal can be improved and it will be easy to estimate the effects of proposed modifications to those processes.

2.2 The Data Mining Process

Like any process, data mining can be carried out haphazardly or systematically. Systematic data mining yields better results over time than haphazard data mining. It is also less likely to come up completely "dry".

A practical data mining application is often complex. It is interactive and involves a number of key steps. The steps in the process of conducting a data mining effort according to Berry and Linoff [5] and Liu Bing [9] are:

1. Understanding the application domain, and the application goals.
2. Extracting one or more target data sets from data bases.
3. Cleaning the data, example, removing noise and handling the missing data.
4. Removing the irrelevant attributes and tuples from the data.
5. Choosing the data mining task, i.e. deciding whether the goal of the data mining process is classification, association, clustering, etc, or a combination of them.
6. Choosing the data mining algorithms.
7. Data mining using the selected algorithms to discover hidden patterns in data.
8. Post-processing the discovered patterns, i.e. analyzing the patterns automatically or semi-automatically to identify those truly interesting/useful patters from the user.

As part of a rapid prototyping sequence, some or all of these steps may be repeated as knowledge is gained about the subject data being mind and based on desired and actual levels of performance. For production implementation of data mining application systems, there are two additional steps in the process that are added to the above steps by Berry and Linoff [5] .These\are:

Step 9. Implementation.

10. Return on Investment Evaluation.

In general terms the above steps can be summarized in to three major steps as follows:

2.2.1 Problem Definition:

In data mining it is possible to do a lot of things when given a clear problem statement and adequate data. It also suggests that patterns occur in all sorts of interesting places and their imperative exploitation can produce useful, and sometimes uprising, results. So for any kind of data mining process the problem should be properly identified, clearly

defined and analyzed before any activity followed to be successful in the data mining process.

2.2.2 Data Evaluation:

The nature and quality of the data, and its ability to convey information for data mining purposes is a critical element in developing successful models. It is desirable to collect information that provides insight into the problem area, but all data collected are not equally important and there is a need to evaluate the data based on its reliability and objectivity.

2.2.3 Feature Extraction and Enhancement:

It is expected that "good" customers have certain characteristics in common, they have adequate incomes, pay their bills on time, have a good empty history, have successfully carried their accounts, have etc, when these factors are preset in combination, confidence that this will be a "good" customer.

Because the goal is to divide the population of potential customers into two groups (good and bad), it makes sense to look at how these two groups have appeared in the past. In the history of the enterprise, there have been good customer and bad customers. Data mining techniques can be used to look for similarities among these two groups of historic customers. Characters which are regularly seen among bad customers, but rarely seen among good customers, suggest caution be used when observed in a particular customer. Characteristics common among good customers, but rare among bad suggest that the risk is low when observed in a potential customer.

Here data mining is being used to codify real, historical experience. The enterprise wants a decision that is conditioned by actual business history. In this way, even an inexperienced officer will have the benefit of the results of actual outcomes for the business collected over time. Of course, the human standing at the service desk will use insight and business policy to make the final determination more consistent, objective and constant with real business experience.

2.2.4 Proto typing / Model development:

What is needed in an application that actually performs the required solution for the problem identified and analyzed properly. A data mining application that ingests feature information and renders a decision of this sort is an example of a "predictive mode". Once the relevant data has been identified, a predictive model based upon the previous data set can be easily built. There are many predictive modeling paradigms currently in use, including rule- based systems, neural networks, decision trees, etc. One appropriate to the problem will be selected and implemented. Validation will be conducted using blind tests.

The validated system will then be integrated into the operational environment as a new module in the decision system, or perhaps as a separate software application. Documentation, online help and supporting business procedures will, of course, have to be created.

For any area of business problems, data mining techniques provide decision support to the managers who must make a decision based on the choices available. In any business context, data mining tools can be used to select and condition data that served as the basis for a predictive model. This predictive model makes available to the managers the accumulated business history as it might bear on the present decision by directly answering the question; what has been our experience with customers similar to this one? The development of a risk model requires a fair amount of experimentation. Work the authors have done in this area has led to the creation of risk models in as little as two months of prototyping. If the modeled population is fairly diverse, results can be improved by building separate models for different portions of the population.

2.3 Data Mining Activities

The term data mining is often thrown around rather than loosely. Data mining is more appropriately named as Knowledge Discovery due to the fact that it can be used for a many sets of activities where all of which involve extracting meaningful new information from the data. The major six activities are [1] [5]:

- Classification
- Estimation
- Prediction
- Affinity grouping or Association rules
- Clustering
- Description and Visualization

The first three tasks –Classification, Estimation, and Prediction are all examples of **Directed Data Mining**. In Directed Data Mining, the goal is to use the available data to build a model that describes one particular variable of interest in terms of the rest of the available data. The next three tasks are examples of **Undirected Data Mining**. In Undirected Data Mining, no variable is singled out as the target; the goal is to estimate some relationship among all the variables.

Data mining activities and the kind of patterns they can discover are described below:

2.3.1 Classification

Classification consists of examining the features of newly presented object and assigning to it a predefined class. The objects to be classified are generally represented by records in a data base. The act of classification consists of updating each record by filling in a field with a class code.

The classification task is characterized by a well defined definition of the classes and a training set consisting of pre classified examples. The task is to build a model that can be applied to unclassified data in order to classify it.

2.3.2 Estimation

Classification deals with discrete outcomes, yes or no, debit card, mortgage, or car loan. Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variable such as income, height or credit card balance.

In practice, estimation is often used to perform a classification task. A bank trying to decide to whom they should offer a home equity loan might run all its customers through a model that gives them each a score, such as a number between 0 and 1.This is actually an estimate of the probability that the person will respond positively to an offer. This approach has the great advantage that the individual record may now be rank ordered from most likely to least likely to respond. The classification task now comes down to establishing a threshold score. Any one with a score greater than or equal to the threshold will receive the offer.

Often classification and estimation are used together, as when data mining is used to predict who is likely to respond to a credit balance transfer offer and also to estimate the size of the balance to be transferred.

2.3.3 Prediction

Any prediction can be thought of as classification or estimation. The difference is one of emphasis. When data mining is used to classify a phone line as primarily used for internet access or credit card transaction as fraudulent, we do not expect to be able to go back later to see if the classification was correct. Our classification may be correct or incorrect, but the uncertainty is due only to incomplete knowledge: out in the real world, the relevant actions have already taken place. The phone is or is not used primarily to dial the local ISP. The credit card transaction is or is not fraudulent. With enough effort, it is possible to check. Predictive tasks feel different because the records are classified according to some predictive future behavior or estimated future value. With prediction, the only way to check the accuracy of the classification is to wait and see.

Any of the techniques used for classification and estimation can be adapted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples. The historical data is used to build a model that explains the current observed behavior. When this model is applied to current inputs, the result is a prediction of future behavior.

2.3.4 Affinity grouping or Association rules

The task of affinity grouping is to determine which things go together. The prototypical example is determining what things go together in a shopping cart at the supermarket. Retail chains can use affinity grouping to plan arrangement of items on store shelves or in a catalog so that items often purchased together will be seen together. Affinity grouping can also be used to identify cross-selling opportunities and to design attractive packages or groups of products and services.

2.3.5 Clustering

Clustering is the task of segmenting a diverse group in to a number of some similar subgroups or clusters. What distinguishes clustering from classification is that clustering does not rely on predefined classes.

In clustering, there are no predefined classes and no examples. The records are grouped together in the basis of self-similarity. It is up to the miner to determine what meaning, if any, to attach to the resulting clusters. A particular cluster of symptoms might indicate particular disease. Dissimilar clusters of video and music purchases might indicate membership in different sub clusters.

Clustering is often done as a preclude to some other form of data mining or modeling. For example, clustering might be the first step in a market segmentation effort. Instead of trying to come up with a "one-size-fits-all rule for" what kind of promotion do customers respond to best, "first divide the customers base in to clusters or people with similar buying habits, and then ask what kind of promotion works best for each cluster.

2.3.6 Description and Visualization

Some times the purpose of data mining is simply to describe what is going on in a complicated data base in a way that increases our understanding of the people, products or processes that produced the data in the first place. A good enough description of a behavior will often suggest an explanation for it as well. At the very least, a good description suggests where to start looking for an explanation.

Data Visualization is one powerful form of descriptive data mining. It is not always easy to come up with meaningful visualizations, but the right picture really can be worth a thousand association rules since human beings are extremely practiced at extracting meaning from visual scenes.

2.4 The Business context for Data Mining

Data Mining-extracting meaningful patterns and rules from large quantities of information-is clearly useful in any field where there are large quantities of data and something worth learning. We would not be surprised to learn, for example, that military

intelligence organizations use data mining techniques to process large quantities of satellite imagery in an attempt to classify things on the ground as tanks or tractors-targets or public relations disasters in the making [5].

In the business context, the same rule applies: data mining is useful wherever there are large quantities of data and something worth learning. In business, there is an explicit definition of what it means for a thing to be worth learning. For a business, something is worth learning if the resulting knowledge is worth more money than it costs to discover. Actually the definition is even stricter than that: something is worth knowing if the return on the investment required to learn it, is greater than the return from investing the same funds some other way.

2.5 The Technical context for Data Mining

Here the technical context for data mining has three main areas:

1. Algorithms and techniques
2. Data
3. Modeling practices

The field that has come to be called Data Mining has grown from several antecedents. On the academic side are Machine Learning and Statistics. Machine Learning has contributed important algorithms for recognizing patterns in data. Machine Learning researchers are on the bleeding edge, conjuring ideas about how to make computers think. Statistics is another important area that provides background for data mining. Statisticians offer

mathematical rigor, not only do they understand the algorithms, they understand the best practices in modeling and experimental design [5] [10].

The final thread is Decision Support. Over the past decades, people have been gathering data into databases to make better informed decisions. Data Mining is a natural extension of this effort.

2.6 Styles of Data Mining

There are two styles of data mining: Directed Data Mining is a top-down approach, used when we know what we are looking for. This often takes the form of predictive modeling, where we know exactly what we want to predict. Undirected Data Mining is a bottom-up approach that lets the data speak for itself. Undirected Data Mining finds patterns in the data and leaves it up to the user to determine whether or not these patterns are important [1],[5][11].

These two approaches are not mutually exclusive. Data mining efforts often include a combination of both. Even when building a predictive model, it is often useful to search for patterns in the data using undirected techniques. These can suggest new customer segments and new insight that can improve the directed modeling results.

2.6.1 Directed Data Mining

The top-part of Figure 2.1 shows a model as a black box. What this means is that we do not care what the model is doing, we just want the most accurate result possible. This is the approach used when we know what we are looking for. When we can direct the data mining effort toward a particular goal.

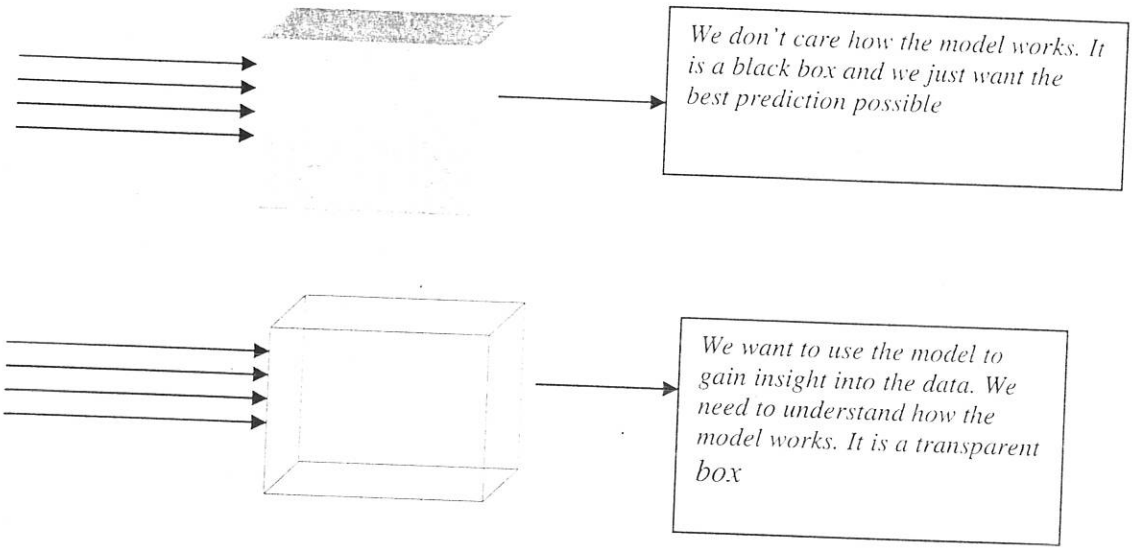

Figure 2.1 Data Mining uses both black box model and semitransparent models

Typically, we are using already known examples, such as prospects who already received an offer (and either did or did not responded),and we are applying information gleaned from them to unknown examples, such as prospects who have not yet been contacted. Such a model is called a predictive model, because it is making predictions about unknown examples.

other, with in a single industry, different companies have strategic plans and different approaches .All of this affects the approach to data mining

The virtuous cycle is a high-level process, consisting of four major business processes.

- Identifying the business problem
- Transforming data into actionable results
- acting on the results
- Measuring the results

There are no short cuts-success in data mining requires all four processes. Results have to be communicated and overtime, we hope that expertise in data mining will grow. Expertise grows as organizations focus on the right business problems, learn about data and modeling techniques and improve data mining processes based on the results of the previous efforts. In short, successful data mining is an example of organizational learning.

2.8 Data Mining for Telecommunication Industry

The rapidly expanding and highly competitive nature of the telecommunication industry creates a great demand for data mining in order to help under stand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of service [1][5],[12].

The following are a few scenarios where data mining may improve telecommunication services.

- Multidimensional analysis of telecommunication data: telecommunication data are intrinsically multidimensional with dimensions such as calling time, duration, location of caller, location of callee, and type of call. The multi-dimensional analysis of such data can be used to identify and compare the data traffic, system workload, resource usage, user group behavior, profit and so on. For example, analysts in the industry may wish to regularly view charts regarding calling sources, destination, volume, and time-of-day usage patterns. Therefore, it is often useful to consolidate telecommunication data in to large warehouses and routinely perform multidimensional analysis using OLAP and Visualization tools.

- Fraud pattern analysis and identification of unusual patterns: fraudulent activity costs the telecommunication industry millions of dollars a year. It is important to identify potentially fraudulent users and their atypical usage patterns, detect attempts to gain fraudulent entry to customer accounts, and discover unusual pattern that may need special attention, such as busy-hour fraudulent call attempts, switch and route congestion patterns and periodic calls from automatic dial-out equipment(like fax machines)that have been improperly programmed. Many of these types of patterns can be discovered by multidimensional analysis, cluster analysis and outlier analysis.

- Multi-dimensional association and sequential pattern analysis: The discovery of association and sequential pattern in multidimensional analysis can be used to promote telecommunication services.

- Use of visualization tools in telecommunication data analysis: tools for OLAP visualization, linkage visualization, association visualization, clustering and outlier visualization have been shown to be very useful for telecommunication data analysis.

2.9 Data Mining Algorithms

Central to the Data Mining process is the Data Mining Model a virtual structure that represents the grouping and analysis of relational or multidimensional data. In many ways, the structure of a Data Mining Model resembles the structure of a database table. However, while a database table represents a collection of records, or a record, Data Mining Model represents an interpretation of records as rules and patterns, composed of statistical information – as Cases. The structure of the Data Mining Model represents the case set that defines the Data Mining Model and data stored represents the rules and patterns learned from processing case data [5], [13], and [14].

Date mining algorithms are used to determine how the Data mining model analyzes the cases. These algorithms also provide the decision-making capabilities needed to classify, segment, associate and analyze the data and give predictive, variance, or probability information about the case set.

Data mining algorithms can be categorized by the general Data mining method they use. The most common methods are: -

2.9.1 Decision Trees

One of the most frequently used Data Mining methods is the Decision Tree. This is a form of data shown in a tree- like structure, in which a mode in the tree structure represents each question that further classify the data. As a result of applying this method to training set, a hierarchical classifying rules of the type." IF---- THEN ---"is created. An advantage of this method is that the representation of rules is intuitive and easily understood by a human.

2.9.2 Neural Networks

This method is based on training a model to 'learn' from data describing previous situations for input parameters and correct reactions to them are known.

2.9.3 Evolutionary programming

The underlying idea of this method is that the system automatically formulates a hypothesis that shows the dependence of the target variable on other variables.

2.9.4 Memory Based Reasoning

This method is also called the Nearest Neighbor Method because it works on the basis of outcomes by finding the nearest similar scenario that occurred in the past and selecting the one that forecast was most accurate.

2.9.5 Genetic Algorithms

The name of this method derives from its similarity to the evolutionary process of natural selection driven by three mechanisms: first selection of the strongest, second, cross-breeding, and third, multiply a number of new generations, built with the help of the

described mechanisms, a solution is obtained cannot be improved any further. This solution is taken as a final one.

2.9.6 Non- linear Regression Methods

These methods are based on searching for a dependence of the target variable on other variables of function of some predetermined form.

In general, Data Mining tool provides solutions to address a particular problem or market. The Data mining algorithms that they use are very effective for these applications but tented to be less well suited for other applications. Therefore, it is important to appreciate that there are different types of algorithms and which ones are best suited to various problem areas.

2.10 Neural Network

A Neural Network is a powerful data-modeling tool that is able to capture and represent complex input/output relationships. The motivation for the development of an artificial system was that it could perform "intelligent" tasks similar to those performed by the human brain. [15], [16], [17]. Neural network resemble the human brain in the following two ways.

1. A neural network acquires knowledge through learning.
2. A neural network's knowledge is stored within inter neuron connection strengths known as synaptic weights.

The true power and advantage of neural networks lies in their ability to represent linear and non- linear relationships and in their ability to learn these relationships directly from the data being modeled. Traditional linear models are simply inadequate when it comes to modeling data that contains non- linear characteristics.

The most common neural network model is the Multi Layer Perceptron (MLP). This type of neural network is known as a Supervised Network because it requires a desired out put in order to learn. The goal of this type of network is to create a model that correctly maps the input to the output using historical data so that the model can then be used to produce the output when the desired output is unknown. A graphical representation of an MLP is shown below.

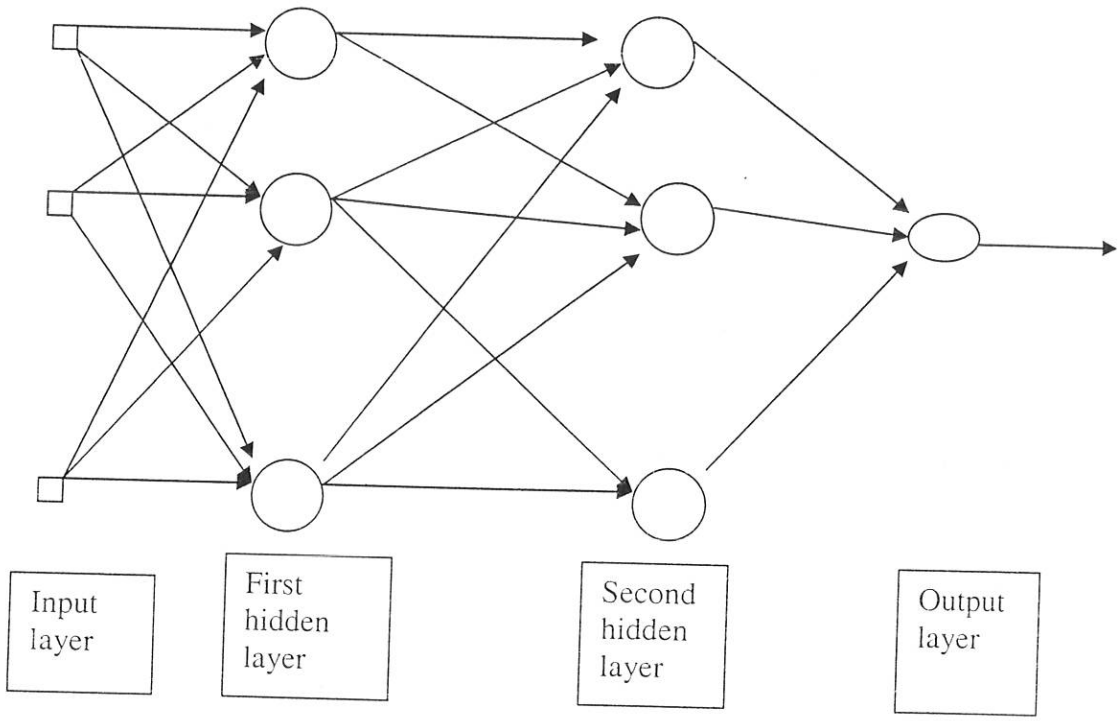

Figure 2.3 A graph representation of MLP.

The MLP and many other neural networks learn using an algorithm called Back Propagation. With back propagation, the input data in repeatedly presented to the neural network with each presentation the output (back propagated) to the neural network and used to adjust the weights such that the error decreases with each iteration and the neural model gets closer and closer producing the desired output. This process is known as "training".

A good application of neural network is the document scanners for the PC come with software that performs a task known as Optical Character Recognition (OCR). OCR software allows to scan in a printed document and then convert the scanned image into an electronic text format such as a word document, enabling to manipulate the text. Of course, character recognitions is not the only problem that neural network can solve. Neural network have been successfully applied to broad spectrum of data intensive applications such as;

- ⇨ **Process modeling and control-** creating a neural network model for a physical plant then using that model to determine the best control sufferings for the plant.
- ⇨ **Machine diagnostics-** Detect when a machine has failed so that the system can automatically shut down the machine when this occurs.
- ⇨ **Portfolios management-** Allocation the asset in a portfolio in a way that maximizes return and minimizes risk.
- ⇨ **Target recognition-** Military application, which uses video and/or infrared image data to determine if an enemy target is present.
- ⇨ **Medical Diagnosis-** Assisting doctors with their diagnosis by analyzing the reported symptoms and /or image date such as MRIS or X-rays.
- ⇨ **Credit rating-** Automatically assigning a company, or individuals credit ratting based on their financial condition.
- ⇨ **Targeted marketing-** Transcribing spoken words into ASCII text.
- ⇨ **Financial forecasting-** using the historical data of a security to predict the future movement of that security.

CHAPTER THREE

FRAUD, FRAUD DETECTION AND PREVENTION

3.1 Definition of Fraud

According to Random House Unabridged Dictionary: [18]

> *Fraud is a deceit, trickery, sharp practice, or breach of confidence, perpetrating for profit, or to gain some unfair or dishonest advantage.*

Fraud is as old as humanity itself, and can take an unlimited variety of different forms. However, in recent years, the development of new technologies (which have made it easier for us to communicate and helped increase our spending power) has also provided yet further ways in which criminals may commit fraud. Traditional forms of fraudulent behavior such as money laundering have become easier to perpetrate and have been joined by new kinds of fraud such as mobile telecommunication fraud and computer intrusion [19].

3.2 Common Types of Fraud

As a consequence of the pervasion of electronic computing systems into private and business events of everyday life a lot of targets for misuse have evolved. A very high potential of financial damage can be found in the areas of electronic banking and telecommunications. Examples of fraud in these areas are credit card fraud; insurance fraud, money laundering by use of digital payment systems, fraud in the areas of mobile communications and the unauthorized intrusion networks [3],[20].

In order to recognize fraud with in the stated areas, methods of machine learning and data mining have been applied since many years. From a computing point of view, the problem, in general, is to extract cases of fraud from a given large data base containing overwhelmingly correct data.

Current research in data mining mainly focuses on the discovery algorithms and visualization techniques. There is a growing awareness that, in practice, it is easy to discover a huge number of patterns in data base where most of these patterns are actually obvious, redundant, and useless or interesting to user. To prevent the user from being overwhelmed by a large number of interesting patterns, techniques are needed to identify only the useful/ interesting pattering and present them to the user.

The common types of fraud that are currently discovered in the modern business world are [3]:

1. Credit Card Fraud
2. Money Laundering
3. Telecommunication Fraud
4. Computer Intrusion
5. Medical and Scientific Fraud

Even if the above lists are the common types of frauds that exist in the modern business world, the focus of this research is only in mobile phone fraud. So the type of frauds that

will be discussed in this chapter will be limited only for common types of Telecommunication fraud.

3.3 Telecommunication Fraud

The telecommunication industry has expanded dramatically in the last few years with the development of affordable mobile phone technology. With the increasing number of mobile phone users, global mobile phone fraud is also set to rise.

When the first antelope mobile communication Networks were launched, the major weaknesses is in the security, particularly the lack of encryption of both the voice channel and the authentication data made the networks susceptible to eavesdropping and cloning. As the technology evolved from analogue to digital (GSM), so the nature of fraud changed as it become more difficult (and more importantly expensive) to eavesdrop and clone, and this led to a shift away from technical fraud towards more procedural and contractual types of fraud. However the possibility of technical fraud cannot be ruled out forever in GSM. As one door is closed on a fraudster, so the fraudster will attempt to open others.[3],[21][22].

It is estimated that the global mobile communication industry currently loses over 25 billion per annum due to fraud [21]. This makes the detection, prosecution and prevention of fraudulent activity important objectives for the mobile communication industry. If these objectives are to be achieved, it is clear that additional security steps need to be taken in GSM and future UMTS systems to make them less vulnerable

We need to distinguish between fraud aimed at the service provider and fraud enabled by the service provider. An example of the former is the resale of stolen call time, and an example of the latter is interfering with telephone banking instruction. We can also distinguish between revenue fraud and non-revenue fraud. The aim of the former is to make money for the perpetrator, while the aim of the later is simply to obtain a service free of charge (or, as with computer hackers, for example, the simple challenge represented by the system).

There are many different types of telecommunications fraud and these can occur at various levels. The two most prevalent types are subscription fraud and super imposed or "surfing" fraud. Subscription fraud occurs when the fraudster obtains a subscription to a service often worth false identity details, with no intention of paying. This is thus at the level of a phone number – all transactions from this number will be fraudulent [3].

Superimposed fraud is usually detected by the appearance of "phantom" calls on a bill. There are several ways to carryout superimposed fraud, including mobile phone cloning and obtaining calling card authorization details. Superimposed fraud will generally occur at the level of individual calls-the fraudulent calls will be mixed in with the legitimate ones. Subscription fraud will generally be determined at some point through the billing process - though one could aim to detect it well before that, since large costs can quickly be run up. Superimposed fraud can remain undetected for a long time.

Other types of telephone fraud include "ghosting"(technology "insiders" fraud where telecom company employees sell information to criminals that can be exploited for fraudulent gain. This of course is a universal case of fraud. Whatever the domain, 'Tumbling' is a type of superimposed fraud in which rolling fake serial numbers are used on cloned handsets, so that successive calls are attributed to different legitimate phones. The chance of detection by spotting unusual patterns is small, and the illicit phone will operate until all of the assumed identities have been spotted.

3.4 Classification of Mobile Telecom Fraud

As fraud can exist across the full range of mobile telecommunication business, discussing every example of fraud is inefficient. So this paper will simply adopt the four fraud groups that are properly and systematically classified by Gosset and Hyland [23]. The four groups are defined as:

3.4.1 Contractual Fraud

All fraud in this category generates revenue through the normal use of a service whilst having no intension of paying for use. Examples of such fraud are subscription fraud, and premium rate fraud.

Subscription Fraud: can take many guises, but can be divided into two classes, one where people enter the contract with no desire to pay for service and the other where people decided part way through a contract that they will no longer pay for service. The latter case usually results in a dramatic change in usage behavior. However, the former

case has no usage history to compare the virtual heavy usage against. In this case additional subscriber indefinite is required to try and assess the risk associated with the subscriber.

Premium Rate Fraud: involves two actions- the setting up of a premium rate service and the acquiring of a number of phones to call this number. The actual mechanism used for perpetrating the fraud will depend up on the payment showed used for the premium rate service. If the premium rate service receives a share of the revenue generated for the network, then the phones will make long duration call to the premium rate number. If the premium rate service receives money from the network according to the number of calls received by the service, then the phones will make a high number of short duration calls. The phones that have been calling this number will then not have their bills paid. The signature of such fraud is therefore dependent on the payment scheme used for the service, but will be a number of high risk phones either making repeated long duration calls or many short duration calls be certain premium rate numbers.

3.4.2 Hacking Frauds

All frauds in this category generate revenue for the fraudster by breaking in to insecure systems, and exploiting or selling on any available functionality. Examples of such fraud are PABX fraud and network attack.

In **PABX Fraud**, a fraudster repeatedly calls a PABX trying to get access to an outside line. Once the fraudster has access to this, they can then dial out, making high value

calls, whilst only paying for the low value PABX access call. Often, such attacks are associated with the use of cloned phones, so that even these low cost calls are not paid for.

In **Network Attacks**, computer networks are attacked through the access modems that are used for remote management or support. Once a modem is hit, the fraudster then tries to break in to the network and configure certain machines for this own end. Such frauds are characterized by rapid short calls to the same number in the case of PABX fraud, or short calls to sequential numbers in the case of network fraud, and it is this behavior that has to be detected.

3.4.3 Technical Fraud

All frauds in this category involve attacks against weaknesses in the technology of the mobile system. Such frauds typically need some initial technical knowledge and ability, although once a weakness has been discovered this information is often quickly distributed in a form that non- technical people can use. Examples of such fraud are cloning, and technical Internet fraud.

In **Cloning Fraud**, the phone call authentication parameters are copied on to another handset, so that the network believes that it is the original handset that is being authenticated.

In **Technical Internal Fraud**, fraudulent employees may alter certain internal information to allow certain users reduced cost access to services. The usage behavior in these frauds depends on how long the fraud is expected to remains undetected. In the situation where the fraudster believes that the fraud can be hidden for a long time, then the best approach would be to exhibit normal usage behavior, as no attention is then drawn to it. However, if the fraud has a short lifetime, then the best approach is to make as much use of the service as possible until it is stopped.

3.4.4 Procedural Fraud

All frauds in this category involve attacks against the procedures implemented to minimize exposure to fraud, and often attack the weaknesses in the business procedures used to great access to the system. Examples of such fraud are roaming fraud, voucher ID duplication, and faulty vouchers.

In the case of **Roaming Fraud**, the billing procedure may mean that the subscriber is billed a long time after the calls were made, in which case the subscriber may no longer be billable. Another aspect of this may be that the subscriber has been identified as a fraudulent roamer, but an error in the procedure of terminating his subscription means that he is able to continue making calls whilst roaming.

In the case of **Voucher Fraud** there may be problems associated with the procedures used to produce, distribute, activate and de-activate the payment vouchers. If the voucher information can be distributed to a number of people who attempt to activate the voucher

at the same time, then if the procedure provides a time window between the phone account being credited and the voucher being de-activated, this may allow a number of people to use the same voucher. Such frauds usually display normal behavior, and can only be countered by tightening up the procedures involved.

3.5 Fraud Detection and Prevention

We begin by distinguishing between fraud prevention and fraud detection. Fraud prevention describes measures to stop fraud occurring in the first place. These includes 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 telecommunication 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 [3], [24].

In contrast, 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.

Fraud detection is a continuous 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 continuously 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 earlier detection tools need to be applied as well as the latest developments.

The developments of new fraud detection methods are made more difficult by the fact that the exchange of ideas in fraud detection is severely limited. It doesn't 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.

Many fraud detection problems involve huge data sets that are constantly evolving. A telephone company carries millions of calls each day and processing this 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 huge size of the data also indicates the potential value of fraud detection: if 0.1% of a 100 million transactions are fraudulent, each losing the company $10, then overall the company loses $1 million [3].

Statistical tools for fraud detection are many and varied. Since data from different applications can be diverse in both size and type but there are common themes. Such tools are essentially based on comparing the observed data with expected values, but expected values can be derived in various ways, depending up on the context. They may be single numerical summaries of some aspects of behavior, they are often simple

graphical summaries in which an anomaly is readily apparent, but they are also often more complex (multivariate) behavior profiles. Such behavior profiles may be based on past behavior of the system being studied (for example, the way a bank account has been previously used), or by extrapolation from other similar systems. Things are often further complicated by the fact that, in some domains a given actor may behave in a fraudulent manner some of the time and not at other times.

Statistical fraud detection methods may be "supervised" or "unsupervised". In supervised methods, samples of both fraudulent and non-fraudulent records are used to construct models, which allow one to assign new observations in to one of the two classes. Of course, this requires one to be confident about the true classes of the original data used to build the models. It also requires that one have examples of both classes. Further more, it can only be used to detect frauds of any type which have previously occurred.

In contrast, unsupervised methods simply seek those account customers, etc which are most dissimilar from the norm. These can then be examined more closely. Outliners are a basic form of non-standard observations. Tools used for checking data quality can be used, but the detection of accidental errors is a rather different problem from the detection of deliberately falsified data or data which accurately describe a fraudulent pattern.

This leads us to note the fundamental point that we can seldom be certain, by statistical analysis alone, that a fraud has been perpetrated rather, The analysis should be regarded

as alerting us to the fact that an observation is anomalous, or more likely to be fraudulent than others-so that it can then be investigated in more detail. One can think of the objective of the statistical analysis as being to return a suspicion score(where we will regard a higher score as more suspicious than a lower one).the higher the score is, then the more unusual is the observation, or the more like previously fraudulent values it is. The fact that there are many different ways in which fraud can be perpetrated, and many different scenarios in which it can occur, means that there are many different ways of computing suspicion scores.

Suspicion scores can be computed for each record in the database for each customer with a bank account or credit card, for each owner of a mobile phone, for each desktop computer, and so on, and these can be updated as time progresses. These can then be rank ordered, and investigative attention can be focused on those with the highest scores, or on those which exhibit a sudden increase. Here issues of cost enter: given that it is too expensive to undertake a detailed investigation of all records, one concentrates investigation on those thought to be most likely to be fraudulent.

One of the difficulties with fraud detection is that typically there are many legitimate records for each fraudulent one. A detection method, which correctly identifies 99% of the legitimate records as legitimate and 99% of the fraudulent records as fraudulent, might be regarded as a highly effective system. However, if only one in a thousand records are fraudulent, then, on average in every 100 that the system flags as fraudulent, only about 9 will in fact be so[3]. In particular, this means that to identify those 9

requires detailed examination of all 100-at possibly considerable cost. This leads us to more general point: fraud can be reduced to low level as one likes, but only by virtue of a corresponding level of effort and cost. In practice, some compromise has to be reached, often a commercial compromise, between the cost of detecting a fraud and the saving to be made by detecting it [3].

3.5 Approaches to Fraud Detection

There are many different approaches and combinations of approaches that are available for the detection of fraud. To take a broad view, there are three general approaches which are described as follows by Gosset and Hyland [24]: learnt, taught, and investigative

3.6.1 The Learnt Approach

The Learnt Approach is typified by the use of unsupervised neural networks, where the fraud detection tool, itself learns what is the expected behavior for each user. The learnt approach is useful for detecting changes in behavior, and hence is most efficient at detecting Subscription and Hacking fraud.

In the Learnt approach, a tool will learn what a typical behavior is, and raise an alarm on any large variations from this behavior. In addition, the tool will generally evolve this typical behavior over time as the user's behavior changes. The tool's ability to monitor the behavior of the user makes it very useful for detecting frauds about which nothing is known, as in nearly all cases of hacking or contractual fraud, these will result in changes

of behavior. If little is known about the fraud that exists in a system, this is a good tool to begin obtaining examples of fraudulent behavior.

However, there are a numbers of drawbacks with such an approach. It is not possible to teach such a tool what to look for, and if the evolution parameters are not set correctly, clever fraudsters will learn how 'ramp up' usage so as not to trigger an alarm.

An example of such a tool is the Unsupervised Neural Network. Here, the inputs to the tool are measured to determine a set of parameters that describe the behaviors of the user. The tool will typically maintain a measure of recent behaviors and longer-term behavior, and is then able to assess the current behavior against both recent and long-term behavior profiles. These profiles will evolve in time according to some evolution parameters that will either be dependent on time or the number of calls.

3.6.2 The Taught Approach

The Taught Approach is typified by the uses of Supervised Neural Net works or Rule-based fraud tools. Such tools are taught what fraudulent behavior looks like, and then try to discover that are in some way similar to these frauds. The Taught approach is useful for detecting signatures of fraud, and hence is useful in detecting Subscription and Hacking fraud. In addition, once a technical fraud is discovered, these can sometimes be detected using the taught approach.

In the taught approach, examples of fraud have been obtained. These are then used to 'teach' the tool what it is looking for. In the case of a Rule-Based System, the fraud examples are analyzed for their fraud signatures, and these are then translated into rules using thresholds or relative measures. In the case of a Supervised Neural Network, examples of fraud are used along with examples of non-fraudulent behavior to teach the tool which behaviors are good and which are fraudulent. Both approaches should identify behaviors in some sense similar to the fraud examples used as fraudulent, and behaviors in some senses similar to good behavior should be deemed non- fraudulent.

There are some differences between the Rule Based and Supervised Neural Network approaches that should be noted. In the case of the Rule-Based System, work has to be done at the beginning of the fraud process to identify the signatures and there by the rules required to detect fraud. However, having done this work, the meaning of and reason for any alarm is immediately apparent. For example, if a call of duration grater than 1,000 hours is a good indicator of fraud, then an alarm raised by a rule- Based System will say that this alarm was raised because the call duration was greater than 1,000 hours.

In the case of the Supervised Neural Network, far less work has to be done at the beginning of the fraud process, as the tool, simply needs to be taught against good and bad behavior. However, any alarm that is raised will simply inform you that the user's behavior is some measure of distance away from a fraudulent example. The analysis then has to be performed at the end of the process, by doing some further exploration of the

behavior. Although some Supervised Neural Network tools simplify this process, it is still more labor intensive than a Rule-Based approach.

However, both of these approaches have the drawback that new types of fraud are not being worked for, and may therefore remain undetected by the taught' tool whilst it maybe detected by the 'learnt' tool.

3.6.3 The Investigative Approach

The Investigative Approach looks for weaknesses in procedures and technical specifications. Clearly such approach is useful in countering technical and procedural fraud.

The Investigative approach is, as it suggests, a matter of auditing the procedures and technology that is employed. This can be performed either internally or by using an external company. Such an approach is extremely useful, and will become more important as exploiting procedural and technical frauds become more attractive.

One example of its usefulness is in ensuring that an employee can not modify a data base on the network so as to grant credit or free usage to certain subscribers. Another is that the machines on the network cannot be accessed remotely in an unauthorized way. As cloning becomes more expensive, such approaches to obtaining fraudulent access to services will become more attractive.

3.6 Related Researches made on Fraud Detection

Telecommunications network generate vast quantities of data, sometimes of the order of several gaga bytes per day, so that data mining techniques are of particular importance. Due to this fact there are several researches made in the area of telecommunication fraud detection. Among them only very few of them are discussed below.

According to Bolton and Hand, [3], as with other fraud domains, apart from some domain specific tools, methods for detection hinge around outlier detection and supervised classification, either using rule-based method or based on comparing statistically derived suspicion scores with some threshold. At a low level, simple rule-based detection systems use rules such as the apparent use of the same phone in two very distant geographical locations in quick succession calls which appear to overlap in time and very high value and very long calls. At a higher level, statistical summaries of call distribution (often called profiles or signatures at the user level) are compared with thresholds determined either by experts or by application of supervised learning methods to known fraud/non-fraud cases.

Murad and Pinkal [3] and Rosset et al [3] 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 effectively. A particularly interesting description of profiling methods is given by Cortes and Pregibon [3]. Cortes et al [3] describes the Hancock language that writing programs for processing profiles, basing the signatures on such quantities as average call duration, longest call day, and

mixture models and Bayesian Networks models for detection fraud based on the data from Call Record Detail.

From the above few but very influential researches we can generalize that the telecom market will become even more complicated over time- with more opportunity for fraud. At present the extent of fraud is measured by taking account of factors such as call lengths and tariffs. The third generation of mobile phone technology will also need to take account of such things as the content of the calls (because of the packet switching technology used, equally log data transmissions may contain very different numbers of data packets).and the researches will also continue in the future too by taking in to consideration new advancements and developments in the industry.

CHAPTER FOUR

MODEL BUILDING AND TRAINING

This chapter details the different data mining steps that were carried out in this research work. The major steps undertaken were:

- Identifying the goal of the data mining task
- Identifying sources of pre- classified data
- Building and test the model.

The first step in any data mining task (i.e. Clear definition of the problem) has been already addressed in the first chapter of this study under section 1.2, so it is not discussed in this chapter. Before going in to the particulars of what was carried out for the different data mining steps, this chapter begins with an attempt to introduce the software used with the steps involved for building a model by using the tools.

In this research work, the data mining task was undertaken by using artificial neural Network. The specific neural network tool used is the Matlab neural network tool. Thus before going to the discussion of the previous specified steps that were carried out in this study, the researcher would like to give an overview of Matlab Neural Network tool.

4.1 Matlab Neural Network software

MATLAB is a high- performance language for technical computing. It integrates computation, visualization, and programming in an easy-to -use environment where

problems and solutions are expressed in familiar mathematical notation. Typical uses include [28]:

- Math and computation
- Algorithm development
- Modeling, simulation, and prototyping
- Data analysis, exploration, and visualization
- Scientific and engineering graphics
- Application development, including graphical user interface building

MATLAB is an interactive system whose basic data element is an array that does not require dimensioning. This allows you to solve many technical computing problems, especially those with matrix and vector formulations, in a fraction of the time it would take to write a program in a scalar non-interactive language such as C or FORTRAN. MATLAB toolboxes currently installed:

- Communications
- Control Systems
- Image Processing
- Optimization
- Signal Processing

4.1.1 Neural Network Toolbox

There are many toolboxes inbuilt within Matlab soft ware. One of them is the Neural Network tool box which helps for designing and simulating Neural Networks [29]. The

Neural Network toolbox extends the MATLAB computing environment to provide tools for the design, implementation, visualization and simulation of neural networks. Neural networks are uniquely powerful tools in applications where formal analysis would be difficult or impossible, such as pattern recognition and nonlinear system identification and control. The Neural Network Toolbox provides comprehensive support for many proven network paradigms, as well as a graphical user interface that allows you to design and manage your networks. The toolbox's modular, open, and extensible design simplifies the creation of customized functions and networks.

Working with Neural Networks

Inspired by the biological nervous system, neural network technology is being used to solve a wide variety of complex scientific, engineering, and business problems. Commercial applications include investment portfolio trading, data mining, process control, noise suppression, data compression, and speech recognition. Neural networks are ideally suited for such problems because, like their biological counterparts, a neural network can learn, and therefore can be trained to find solutions, recognize patterns, classify data, and forecast events.

Unlike analytical approaches commonly used in fields such as statistics and control theory, neural networks require no explicit model and no limiting assumptions of normality or linearity. The behavior of a neural network is defined by the way its individual computing elements are connected and by the strength of those connections, or

weights. The weights are automatically adjusted by training the network according to a specified learning rule until it properly performs the desired task.

The neural network tool box used for this research has the following key features:

- Graphical user interface (GUI) for creating, training and simulating neural networks.
- Support for the most commonly used supervised and unsupervised network architectures.
- A comprehensive set of training and learning functions
- Automatic generation of Simulink models from neural network objects.
- Modular network representation, allowing an unlimited number of input sets, layers, and network interconnections.
- Pre- and post- processing functions for improving network training and assessing network performance.
- Routines for improving generalization. and
- Visualization functions for viewing network performance.

Because neural networks require intensive matrix computations, MATLAB provides a natural framework for rapidly implementing neural networks and for studying their behavior and application.

The Neural Networks Toolbox GUI helps to import potentially large and complex data sets. The GUI also allows for creating, initializing, training, simulating, and managing

networks. The simple graphical representations allow visualizing and understanding network architecture easily.

The neural network tool box has the following supported Network Architectures.

Supervised Networks

Supervised neural networks are trained to produce desired outputs in response to example inputs, making them particularly well suited for modeling and controlling dynamic systems, classifying noisy data, and predicting future events. The Neural Network Toolbox supports the following supervised networks:

- **Feed- forward networks** have one- way connections from input to output layers. They are commonly used for prediction, pattern recognition, and nonlinear function fitting. Supported feed- forward networks include feed- forward back propagation, cascade-forward back propagation, feed-forward input-delay back propagation, linear, and perception networks.
- **Radial basis networks** Provide an alternative fast method for designing non-linear feed-forward networks. Supported variations include generalized regression and probabilistic neural networks.
- **Recurrent networks** use feedback to recognize both spatial and temporal patterns. Supported recurrent networks include Elman and Hopfield.
- **Learning Vector Quantization (LVQ)** is a powerful method for classifying patterns that are not linearly separable. LVQ allows to specify class boundaries and the granularity of classification.

Unsupervised Networks

Unsupervised neural networks are trained by letting the network continually adjust itself to new inputs. They find relationships within data as it is presented and can automatically define classification schemes. The Neural Network toolbox supports two types of self-organizing unsupervised networks:

Competitive layers -recognize and group similar input vectors. By using these groups, the network automatically sorts the inputs into categories.

Self-organizing maps- Learn to classify input vectors according to similarity. Unlike competitive layers, they also preserve the topology of the input vectors, assigning nearby inputs to nearby categories.

Supported Training and Learning Function

Training and learning functions are mathematical procedures used to automatically adjust the network's weights and biases. The training function dictates a global algorithm that affects all the weights and biases of a given network. The learning function can be applied to individual weights and biases within a network.

The following table easily summarizes the supported Training functions within the toolbox [29]:

Trainb	Batch training with weight and bias learning rules
Trainbfg	BFGS quasi-Newton back propagation
Trainbr	Bayesian regularization
Trainc	Cylical order incremental update
Traincgb	Powell- Beale conjugate gradient backpropagation
Traincgf	Fletcher-powell conjugate gradient backpropagation
Traincgp	Polak- Ribiere conjugate gradient backpropagation
Traingd	Gradient descent backpropagation
Traingda	Gradient descent with adaptive learning rate(lr) backpropagation
Traingdm	Gradient descent with momentum backpropagation
Traingdx	Gradient descent with momentum& adaptive Ir backpropagation
Trainlm	Levenberg-Marquardt backpropatation
Trainoss	One step secant backpropagation
Trainr	Random order incremental update
Trainrp	resilient backpropagation(Rprop)
Trains	Sequential order incremental update
Trainscg	Scaled conjugate gradient backpropagation

Table 4.1 Supported Training Functions of Matlab neural network toolbox.

4.2 Building Classification Models

Fortunately, the basic process for building classification models is the same, regardless of the data mining technique being used. Success depends more on the process than on the technique. And this process depends critically on the data being used to generate the model. Garbage -in garbage -out is an adage that applies especially well to classification and predictive model building according to Barry and Linoff [5].

In pre classified data, the outcomes are already known. And because these known examples will be used to teach the model about the data, this set is called the model set. The basic steps in building and applying a classification model are as follows [5].

1. The model is trained using pre classified data in a subset of the model set called the **training set**. In this step, the data mining algorithms find patterns of predictive value.
2. The model is refined, using another subset called the **test set**. In order to get a good model, the model needs to be refined. Otherwise it will be hard to prevent the model from memorizing the training set, thus refining is important for ensuring that the model is more general and will work better on training data. For this purpose a test data set is required.
3. We can estimate the performance of the model, or compare the performance of several models, by using a third set, entirely distinct from the first two. This holdout set is called the **evaluation set**.

4. The model is applied to the **score set**. The score set is not pre classified and is not part of the model set. We do not know the outcomes for this data. Presumably, we will use the predictive scores to make more informed business decisions of cause, the details of these steps do depend on which data mining technique that is being used and on which tool is being used, but the overall process remains the same.

4.3 Models built using Matlab Soft Ware

To build neural Network model, the first task to be performed at this step of the data mining process was importing the cleaned and prepared data set, which was in Excel format, into Matlab Neural network tool. This prepared data set consisted of 900 sample records about fraudulent and non-fraudulent customers from the original two months billing data of ETC.

As it was explained in the first section of this chapter, the Matlab soft ware needs the following research method steps to be followed.

4.4 Identifying sources of Pre-classified Data

Berry and Linoff [5] state that the primary requirement for data mining task is availability of data in any format. And in most cases the ideal sources of such data is the corporate data warehouse (where data warehouse refers to the collection of data from many different sources and it stores in a common format with consistent definition for keys and fields).

Especially in supervised learning systems, we use pre-classified historical data (past data) to build a model of the future. Based on the discussion with appropriate experts with in ETC, the researcher identified three main sources of data, which are relevant for this research project.

The first and the major source of data identified was the CDR (Call Detail Record) generated by the switch machines of ETC, which have more than 50 fields. The CDR contains all information that machine can generate with regard to the call starting from its being to end of the call. The second source of data was the billing data which is filtered from the CDR for billing purpose only. The billing data contains only major and relevant fields assumed to be necessary for billing by the IT and Data service division of ETC. The last data set was the Customer data base used by Finance Department for follow up of bad debt expenses (uncollectible bills).The nature and content of each sources are as follows:

4.4.1 Call Detail Records (CDR)

A Call Detail Record (CDR) is a single record for each call made over the telephone network. Because so many telephone calls are made, CDRs are a very large data source which contains information related to billions of calls made daily.

A Call Detail Record contains one call module which in turn contains tagged call data related to a call. All charging output data is stored in Call Data Record [30]. Every record has a Tag and a Length indicator that precede the contents of the record. The Tag is defined by an exchange parameter, and the Length contains the number of octets which

comprise the contents. All call data records have the same structure though the record length may vary. i.e.

Tag = Cell Data Record
Length = Number of octets
Contents = Charging Data

Table 4.2 Call Detail Record format

Call Modules: The tag indicates the type of call module which consist of one of the following modules: either Land to Land (L-L) module: which is a call module for Land-to -Land calls, or Mobile to Land (M-L) module: which is a call module for Mobile-to -Land calls, or Land to Mobile(L-M) module -which is a call module for Land -to -Mobile call, or Mobile to Mobile (M-M) module: which is a call module for Mobile -to -Mobile calls, or Inter Exchange Hand off Calling Subscriber (IHA) :which is a call module for inter-exchange hand off for calling subscriber or the Inter Exchange Hand off Called Subscriber (IHB): which is a call module for inter-exchange hand off for called subscriber.

The Length indicates the number of octets that follows with in the call module. All charging data is out put according to the Abstract Syntax Notation Number 1 (ANS.1) Formal Description.ANS.1 is specified in International Standards of Organization (ISO) Recommendation 8824 and 8825 [30].

The CDR stores data in binary form and it is kept for a period of not more than six months due to the requirement of large data ware house. In spite of the fact that this is the

major and very important data source for identifying different kinds of telecom fraud, for this research work, the researcher used the converted CDR data which is commonly known as the billing data for this research. The Billing data, which is currently used by the IT and Data Service Division of ETC to generate bills for customers is converted by a billing software to charge customers for the service the corporation provide to them. so therefore, as a data source, the data included in the research are only those calls that are billable to the caller, with the intention that they didn't include incoming calls (since the called person typically does not pay for these), toll -free calls, or calls made by network intrusions. Beside this there are also lots of data excluded from the data base due to the fact that calls which have missed values are assumed by the operators as errors generated by the switch. This excluded data set misses either the date, duration or both for the calling party.

4.4.2 The Bill Data Record Format

The following table describes the typical call detail Record format used for billing purpose by the IT and Data Service division of ETC.

Field Name	Description
From_Number	is a string representation of the telephone number originating the call. Usually this is a 9- digit number where 3-digits(i.e 009) for code and 6-digits(xx-xx-xx) for the telephone number
To_Number	is a string representation of the telephone number. This usually shows the specific telephone number to whom the call is made and also the area code. This field has very inconsistent data size which ranges from 6 to 18 digits based on the type of call and the technology used by the other party to whom the call is made.
Date_Of-Call	is the data that the telephone call is made which has a 6-digits in a yy-mm-dd format.
Start_Time	is the time that the telephone call started which has a 6-digits in a hh-mm-ss format.
Duration_Of_Call	represents the length of the telephone call, which has a 6-digit hh-mm-ss format.

Table 4.3 CDR Record Format for the billing data

Since this data is accumulated for about six months only and the monthly data set is very large in size, the researcher took three month calls for the month of October and November of 2003 and March of 2004 as an initial data source for the research.

Due to the fact that the raw data is very big to handle in terms of the time and space required, (i.e. 687 MB data of October 2003, 814MB of November 2003 and 986 MB of march 2004), the researcher use a file splitter (version 1.1, 1999) by Steve D.Perkins [31], to split the data and to make the exploration more efficient. Thus, using the file splitter the CDR was divided into prepaid and post paid mobile calls and the prepared mobile phone CDRs are excluded from the database required for the research.

After excluding the prepaid CDR to make the processing more efficient and also to select appropriate sample for the study out of the call details of all post-paid mobile customers any call which didn't fulfill the relevant caller number, destination number, call date, call time, and call duration is excluded from the data base by using **egrep** program. **egrep** program is a text processing tool where its basic function is to be through a text file line for line, and print all lines that matching a search pattern or regular expression to standard out put [32]. After this tedious filtering activity and the discussion with the domain experts of Customer Service Department of the corporation the potential fraudulent calls are more or less assumed to be concentrated on international calls and all non-international call and also internationals calls less than one minute per call are excluded from the data base by using **egrep** program.

Based on the above pre processing step the researcher collect the following summarized data for further analysis.

4.5 Preparing Data for Analysis

As it is described in the data mining methodology of Berry and Linoff [5] data cleaning and date preparation is the second step of any data mining task. In particular, data collected from different sources has to be massaged into a form that will allow the data mining tools to be used to best advantage. This process of data cleaning and preprocessing is highly dependent on the technique (method) to be employed. In this research work, artificial neural network data mining technique was used to build a classification model. So, data preparation was conducted by taking in to consideration the requirements of the neural network tool of Matlab software.

Therefore, in preparing the data set with a form that is convenience for Matlab, the following data cleaning and preprocessing steps were undertaken on the data set created for analysis. Models built by using Matlab and its details are presented under section 4.6.

4.5.1 Data Transformation and Reformatting

In general, the way the data is represented is often crucial to the success or failure of the neutral network project. Fortunately due to the fact that the data is generated by the switch machine there is no much area of confusion that is expected to be avoided during the data transformation process except the field for destination call number and calls which didn't fulfill the relevant fields for billing purpose.

To create the data model the researcher took the raw data sets that were collected for analysis into the format required by the data models. So, from the CDR fields a weekly connection view containing aggregated data is generated containing all information useful for indication of fraudulent call patterns.

Preparing and aggregating the data to build a weekly connection view is done by using SQL Server 2000 by defining several tables collecting measured values for the whole week (as number of calls, average number of calls, sum duration of calls, average duration of cases, and so on).

Getting enough samples of fraudulent calls is a tedious and challenging task to train the neural network. So based on the recommendation of Berry and Linoff [5],for rare occurrences of fraudulent data it is necessary to use over sampling technique to make the rare occurrences of fraudulent calls on the average range between 10 to 40 percent of the total sample size. Thus by properly identified 210 fraudulent calls from customer data set, it limits the total sample data to be 900. Because of the fact that to train the model properly, according to Berry and Linoff recommendation the sample size for the rare instants like fraud as compared to the total size of the population. is in between 20 to 30 percent by using over sampling technique[5]. Based on this recommendation, the fraudulent customers were made to be 23 Percent for this research to make it in the recommended range. Selecting those 900 customers from the total 9153 customers who made an international call of more than one minute per call was a time consuming activity. This is because within the study period there were more than 120, 000 international calls with duration of more than one minute per call that were properly recorded in the database for postpaid customers. These all international calls were made by 9153 customers of Ethio-mobile within two months period. Out of which the proportion of the customers who made the international calls were as follows.

Category	Proportion
Employees	0.02
Govermnent Organisations	0.02
Embassies and International Organisations	0.07
Business Enterprises	0.22
Residences	0.67
Total	1.00

Table 4.5 Proportion of International calls by costumers of Ethio-mobile.

Based on the above proportion of the customers and a discussion with experts from ETC, Only 900 customers data set were taken from business and residence categories of Ethio-mobile customers who represent 89 percent of the total subscribers who made international calls and highly susceptible for making fraudulent calls. The method used to select this sample groups is stratified sampling technique to get representative data from all the prepaid mobile subscribers.

Based on the above justification for the 900 customer groups which include 210 fraudulent and 690 good customers, a summarized data set which includes the relevant data fields for training the model are prepared. The fields of the aggregated data are as follows.

Field Name	Description of the field
No_of_Calls_Per_wk	Number of all international calls made within one week interval during the study period.
Dur_Of_ Calls_Per_wk	Duration of all calls made with in one week interval of the study period
Average_Dur-Of_Calls	Average duration of calls made within the study period by aggregating the weekly duration of calls.
Average_ No_ Of_Calls	Average number of calls made within the study period by aggregating the weekly duration of calls
Max_Duration	Duration of the longest call made out of all the calls made in one week intervals with in the study period.
Min_Duration	Duration of the shortest call made out of all the calls made in one week intervals with in the study period
Max _No_Of_Calls	The largest number of calls made out of all the calls made with in one week intervals in the study period.
Total_No_Of_ Calls	Total number of international calls made during the study period.
Total_Dur_Of_Calls	Total call duration of international calls made during the study period.
STD_For_No_Of_Calls	The standard deviation of the number of calls collected in one week interval within the study period.
STD_For_Dur_Of_Calls	The standard deviation of the duration of calls collected in one week interval within the study period.

Table 4.6 Fields of the Analyzed data set format prepared fro the network.

4.5.2 Preparing Data in to a form that is acceptable to the neural network

In order to facilitate the training the tool required dividing the data set in to input and target data files for all the training, validation and testing purposes. Accordingly prior to importing the dataset in to the work space of Matlab, six files in the name of Training input, Training target, Validation input, Validation target, Testing input ,and Testing target are prepared in excel by dividing the data in the proportion of 60% for training , 20% for validation,and 20% for testing. After that the data set is imported to the work space of Matlab software for further analysis and training.

4.6 Models Built Using Matlab software

To build the neural network model, the first task performed at this stage of the data mining process was importing the cleaned and prepared data, which was in Excel format, in to the neural network workspace. This prepared data set consists of 900 sample records of post paid mobile customers of ETC.

As it was explained in the first section of this chapter, Matlab software has many independent programs and out of these the neural network tool is used for this research. The neural network tool has facilities to import data from the Excel to carry out manipulation on the imported data file, and to create Matlab files in the workspace of the tool.

More over the neural network tool has a facility to train and test networks (models) simultaneously. Thus after the data set was imported, in order to make the data

compatible to the software requirement, the dimension of the data set which was originally the columns for field names and the rows for each records in the Excel, is now reversed in the work space of the tool to make the input fields name rows for the matrix prepared by the neural network tool.. After that all the necessary data manipulations such as defining the input source, target, network algorithm type, training function, learning function, performance function number of neurons for layers, transforming functions, training parameters were selected out of the available inbuilt options for each model built during the process.

The first training was conducted by using all the input attributes that were selected during the data preparation phase. Those variables that were used to build the first network model were Category, Minimum number of calls, maximum number of calls, average number of calls, standard deviation of number of calls, total number of calls, minimum duration, maximum duration, average duration, standard deviation of durations, and total duration.

For the initial trial the algorithm selected was back propagation algorithm (the architecture of the algorithm is presented at the Annex-I. There are many variations of back propagation algorithm and it is difficult to know which training algorithm will be the fastest for a given problem. It will depend on many factors, including the complexity of the problem, the number of data points in the training set, the number of weights and biases in the network, and the error goal.

In general according to the recommendation of Professor Emeritus and et al [29], on networks which contain up to a few hundred weights **the Levenberg-Marquardt algorithm** (**trainlm**) have the fastest convergence and recommended if there is no memory problem for the training. So in this research Levenberg-Marquardat algorithm (trainlm) is first used, and the training functions selected are all the default training parameters provided

As it was discussed before list of variables used to build the model, were all numeric attributes and for the initial training all the attributes were used to train the model, but in this particular run the performance of the model was 0.1505 which had more than 15% error generating chance, and the level of accuracy was 74%. (see the figure below)

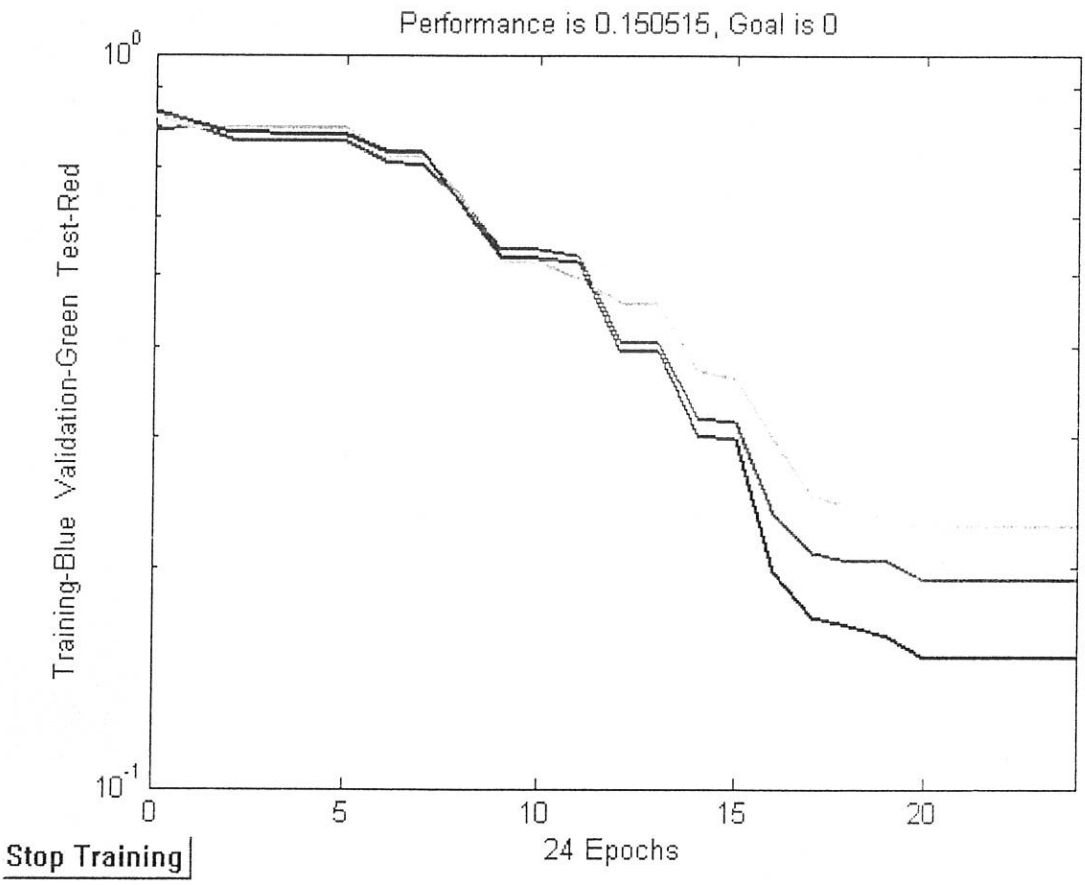

Figure 4.1 The First training model

Therefore, in order to trace the problems that forced the network to have a trouble in learning, the initial training is abandoned. However, before adjusting parameters and build another, this network has been saved and analysis of this network showed an accuracy rate of 74% in training data set and 71% in test data set. The accuracy is higher in the training set because of the fact that the network has been trained on the samples of the training set not on the samples of the test set.

To address the problem of the above network, the first option considered by the researcher was to train and test a network by varying the default parameters of the **trainlm** algorithm and its variations and also other algorithms like BFGS algorithms like **trainbfg (BFGS quasi-newton backpropagation**) were tested but the results are more worse than the previous one. Several networks were saved during the training and the highest accuracy obtained from these networks was 86 percent.9see the follwing fifures for the selected training models)

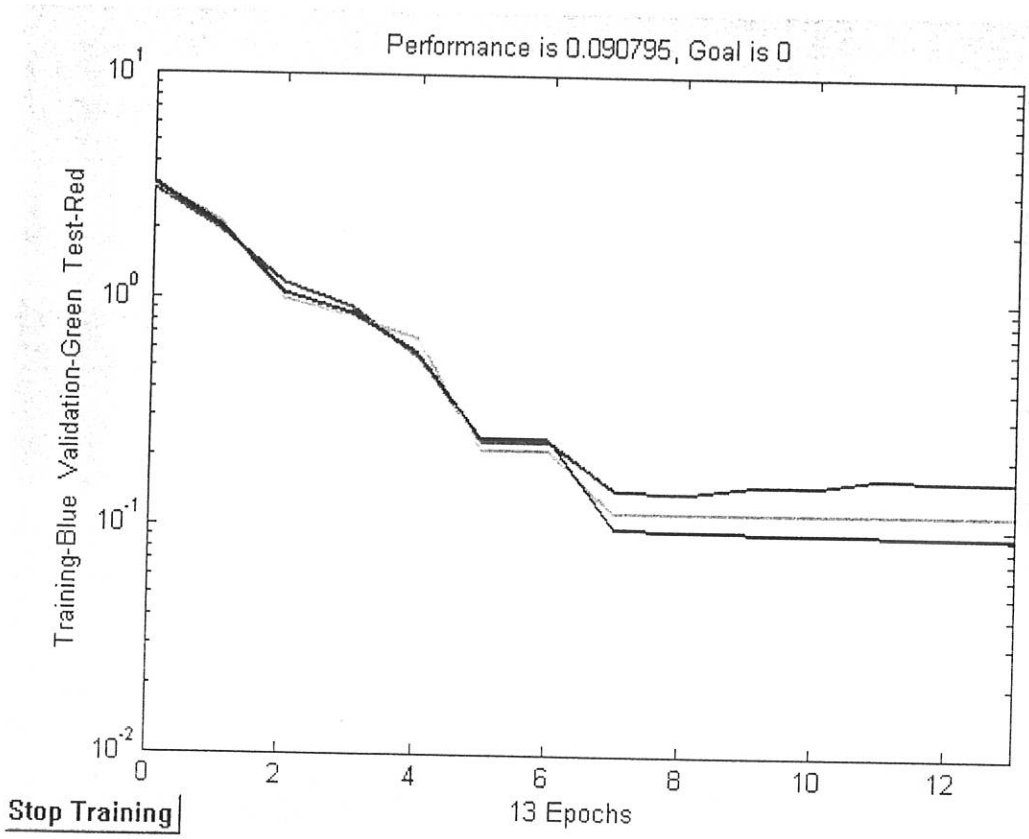

Figure 4.2 Training model with 84% accuracy.

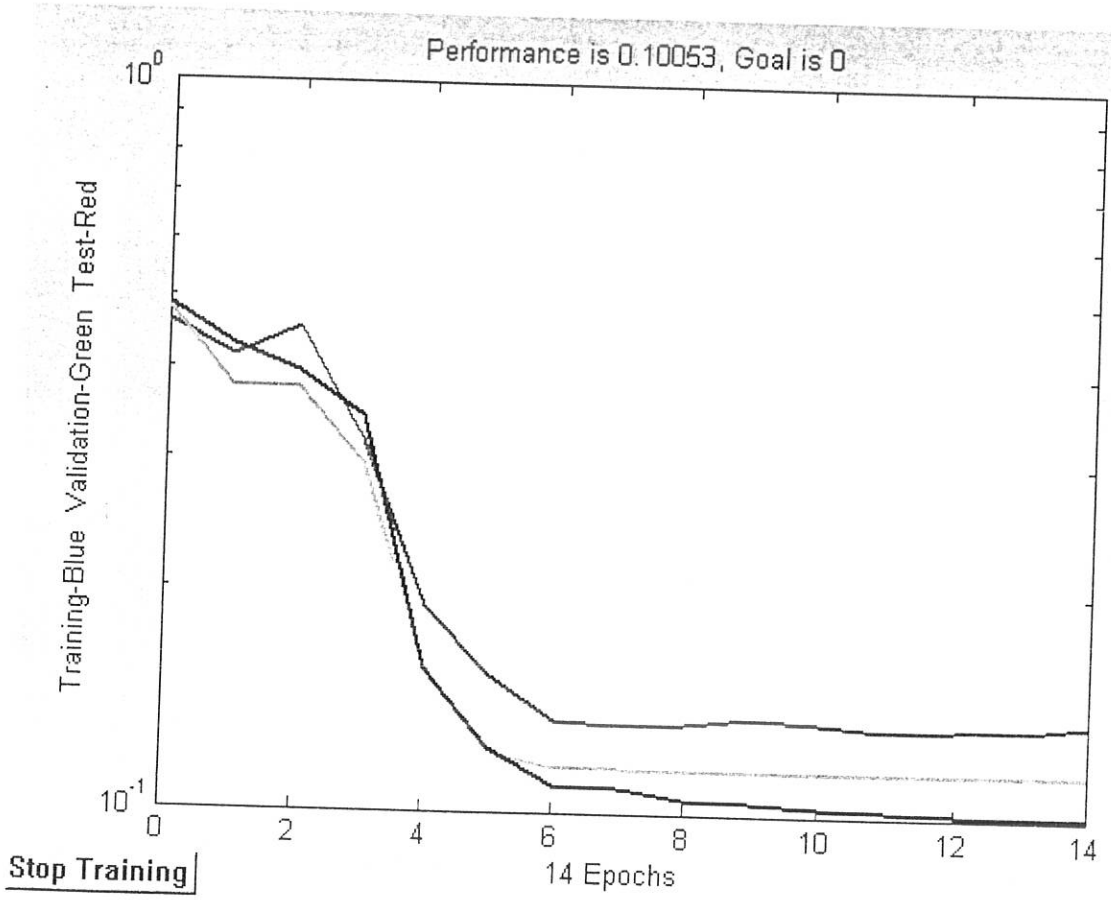

Figure 4.3 Training model with 86% accuracy

Therefore afeter the above trials, the only way left was to return back to **trainlm** and see its performance by changing its supporting training and learning functions. Following a continuous attempt, the researcher persistent the experiment by considering various options suggested to improve the performance of the neural network models.

After lots of the tiresome and tedious training sessions the maximum accuracy achieved was 89 percent with the error performance factor of 0.0686 on the training and 84.68 on

the test data set. The level of performance and accuracy for both the training and testing data sets are summarized as follows for the major trainings.

Number	Learning function	No.of Neurons for 1st layer	transformation function	performance	Accuracy on training dataset	Accuracy in test data set
1	LearnGDM	5	Tansig	0.0825	87.07	82.39
2	LearnGDM	8	Tansig	0.09	88.69	80.68
3	LearnGDM	10	Tansig	0.0804	88.69	80.68
4	LearnGDM	12	Tansig	0.0686	89.23	84.66
5	LearnGDM	15	Tansig	0.1329	84.92	80.68
6	LearnGDM	20	Tansig	0.0959	87.25	81.82
7	LearnGD	5	Tansig	0.0905	86.71	81.80
8	LearnGD	8	Tansig	0.0877	78.40	76.68
9	LearnGD	10	Tansig	0.126	74.98	72.82
10	LearnGD	5	logsig	0.0786	86.71	81.80
11	LearnGD	8	logsig	0.0862	86.36	80.68
12	LearnGD	15	logsig	0.126	86.54	81.82
13	LearnGD	20	logsig	0.1505	84.92	80.68
14	LearnGDM	12	purelin	0.07719	88.51	81.82

Table 4.7 Results of some selected trainings.

The following figure is the training result of the final training made by using the neural network for classifying fraudulent and non fraudulent customers of Ethio-mobile.

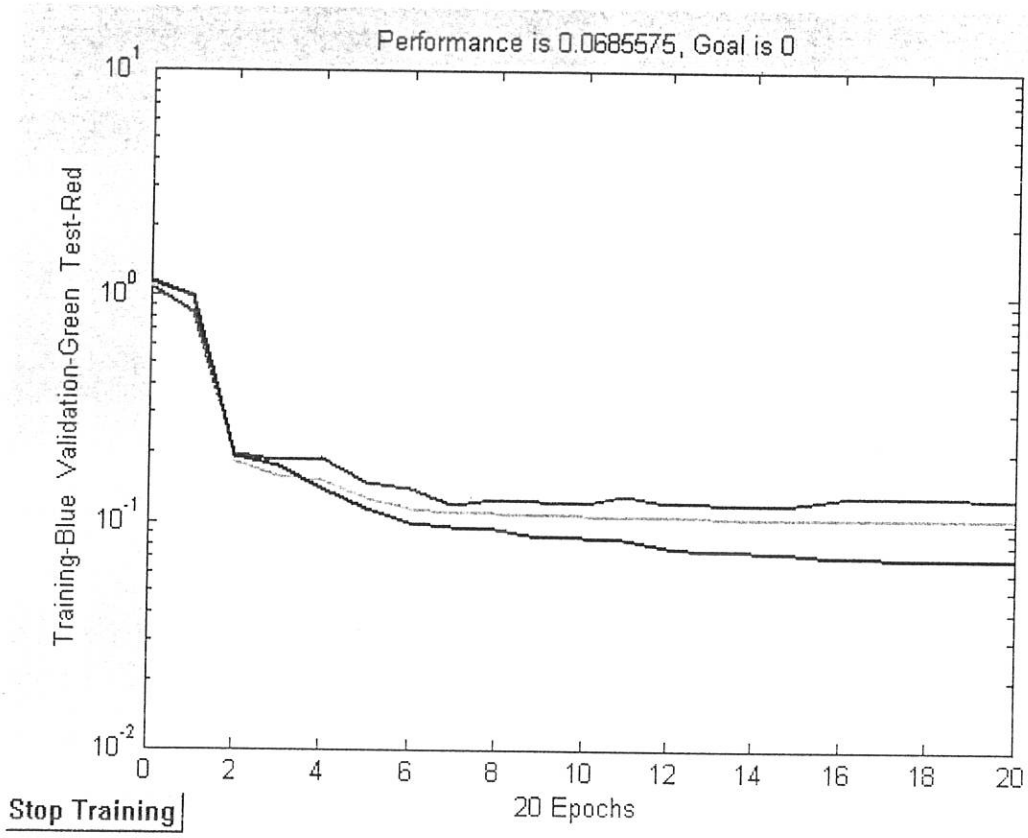

Figure 4.4 The final training model with 89% accuracy

CHAPER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

No matter what kind of organization it is, where it belongs, how it perform its activities, to day, getting data is not a major problem unlike what it was before few decades ago. How ever, the way the data is organized and the level of understanding by the organizations to use this resource beyond the common is far from what the modern and tough competitive market expects. Especially in the telecommunication industry, billions of data are generated daily due to the nature of the industry it self and there is noting done on it except billing only innocent customers.

In order to properly utilize this large volume of data generated by telecommunication operators for effective and efficient management decisions, the importance of data mining technology is unquestionable.

Data Mining technology contribute a lot for efficient customer relation ship management, predictive modeling of customers' behavior, customer segmentation, churn prediction, customer insolvency, and fraud detection in other developed and developing countries of the world. Unfortunately in Ethiopia, the importance of the technology and its contribution is not recognized, except some academic researches made in the Department of Informatics, Addis Ababa University.

Currently the telecom operators are losing millions of dollars each year due to the fraudulent customers and internal employees, and it is important to know who is doing what, who is risky and more importantly act in view of keeping the corporation growing. Thus the objective of this research undertaking was to assess the possible application of data mining technology in the telecommunication context, and particularly in mobile telephone, by developing a model that could help classify whether a customer was good or fraudulent. Such a model could then be applied in assisting the Customer Service division of ETC.

The methodology employed followed three basic steps: data collection, data preparation, and model building and testing. However since a data mining task is an iterative process, these steps were not followed strictly. There were frequent instances where there was a need to go back and forth between the different steps.

Numerous trials had to be made to come up with a model that made good classification. The best performing neural network model developed with an accuracy level of 89%. During the course of model building, there are important findings that were observed. The research revealed that some variables were consistently observed to be important variables for model building. These variables are related to the number of calls made during a specific period and the duration of each call. During the research work the researcher has come to observe that there are lots of data generated by the CDR, which were not used for the billing purpose by the IT and Data Service division. The researcher, therefore hope that the findings of this research work will help to give due emphasis for

the data set which doesn't generate attributes related to duration and time of call which are not currently used for billing purpose. During the model building and testing process, the suggestion and opinions given by domain experts was also worth doing.

In general, encouraging result were obtained by improving neural network approach and this research work have proved the potential applicability of data mining technology to classify fraudulent and non-fraudulent calls based on the analysis of call details. The encouraging result obtained from this research should be considered to support the activities performed by the customer division of Ethio-mobile to prevent and detect telecom fraud.

5.2 Recommendations

On the basis of the findings of this research work, the researcher would like to make the following recommendations in relation to the possible application of data mining technology in supporting fraud detection activities of Ethiopian Telecommunication Corporation. Although the current research work was only in one type of fraud only, subscription fraud(accounting fraud),it is the considered opinion of the researcher that the basic findings of the research work and the resulting conclusions are fairly applicable to other types of fraud.

The researcher would also like to note that this research, as an academic exercise, should only be considered as a preliminary effort to asses the applicability of data mining technology in ETC. Accordingly, the findings of this research undertaking can fairly be

properly identify the good customers from the bad ones by compensating what is missed by one system by the other.

- **Further study in other possible sources of fraud based on the pure CDR data which is automatically generated from the switch machine**.

Fraud in a continuous and broaden type of activities which are related to the need and interest of the human beings to benefit out of the weakness of the system used by the corporation. As a consequence, any type of internal and external frauds by both employees and customers are expected and uninterrupted and in-depth research based on the original CDR data is required to keep the interest of the corporation and also to get reliance and trust by the innocent customers for the fair and appropriate service charges of the corporation.

References

1. Han, J., and Kamber, M. 2001. Data Mining: Concepts and Technologies http://www.cs.sfu.ca.
2. Ethiopian Telecommunication Corporation http://www.telecom.net.et
3. Bolton, R. J.. and Hand D. J.. .2002. Statistical Fraud Detection; A Review Department of mathematics. Imperial College. London. UK.
4. Unpublished Financial Report of the Finance Department of ETC., 2003.
5. Berry, Michael J. A. and Linoff, Gordon..2000. Data Mining Techniques: for Marketing, Sales, and Customer support. New York; John Willy& Sons, Inc.
6. CRISP-DM. CRISP-DM 1.0: Step-by-step data mining guide. 2000. Online. Iternet. http://www.crisp-dm.org.
7. Two Crows Corporation. Introduction to Data Mining and Knowledge Discovery. 3rd ed. 1999. On request. Internet. http://www.twocrows.com.
8. Brutterworth-Heinemann.2001.Data Mining solve your difficult problem..www.bh.com
9. Liu, H. and Motoda, H. 1998. Feature Selection for knowledge Discovery and Data Mining Available URL: http://www.databaseheadquarters.com/bookstore/management2/079238198XAM US141630.shtml
10. Mitchell, M. (1997). *Machine Learning*. New York: The McGraw-Hill Companies, Inc.
11. Mannila, Heikki. 2002. Methods and Problems in data Mining. Available URL : http://www.cs.helsinki.fi/~mannila/
12. Anderea Andreascu and jan Zilliacus.2001.data mining Application for telecom Operators. http://www.pafis.sht.fi/~andand02/workshops/sfis_workshops
13. Raghavan, Vijay, Deogun. Jitender S. and Sover Mayri, 2002. Data Mining : Trends and Issues. Available URL: http://citeseer.nj.nec.com/138316.html
14. Kaird Ltd.2003.data Mining Explained.www.kairon.com
15. Frohlich, Jochen. Neural Net Overview. 1999. Available URL: http://rfhs8012.fh-regensburg.de/~saj39122/jfroehl/diplom/e-1-text.html
16. Grove, Tom D. 2001. Neural Nets – Part I: Why are people more intelligent than machines? Available URL: http://umtii.fme.vutbr.cz/MECH/NN/tomgrl.html
17. Bigus, J.P. (1996). *Data mining with neural networks in solving business problems: From application to Development to decision support*. New

York: McGraw-Hall.

18. Random House Compact Unabridged Dictionary, 2nd edition., RANDOM HOUSE INC. NEW YORK, N. Y. USA.
19. STS-Securing Telecommunication System: Combating Telecom Fraud http://www.sts2000.com/index.htm
20. Data Mining Telecommunication Network Data for Fraud management (1999). http://citeseer.nj.nec.com/sterritt00data.html
21. Seymour, B. (2002), How Neural Network Technology can Tackle the Growing Telecom Fraud Problem, Information Security Bulletin, CHI Publishing Ltd., UK
22. Dinkla, J. (2001), Application of Data Mining techniques for Computer-Aided Fraud Detection. http://goethe.ira.uka.de/people/dinkla/fraud/
23. Gosset, P. and Hyland, M. (2000), Classification, Detection and Prosecution of Fraud and Mobile Networks, UK.
24. Weiss, G, et al (1998), Intelligent Telecommunication Technology http://www.research.rutgers.edu/~gweiss/papers/jain98.pdf
25. Moreau, Y., Lerouge, E., Verrelst, H, Vandewalle, J., Stormann, C., and Burge, P. (1999), A Hybrid System for Fraud Detection in Mobile Communication, ESANN' 1999 Proceedings-European Symposium on ANN, Belgium
26. Hollmen, J. (2000), User Profiling and Classification for Fraud Detection in Mobile Communications Networks, Helsinki University of Technology, Finland.
27. Hollmen,j. and Volker ,T.2000.Call-based Fraud detection in Mobile Communication Networks using a Hierarcical Regime-switching model.Finland http://www.brauer.informatic.tu-muenchen.de/~trespuol/papers/nisp_books.pdf
29. The Matlab Software.The Matworks.The Users Guide, www.mathworks.com
30. Neural network tool box.What is a nueral network.www.mathsworks.com
31. The common Charging out put.telefonaktiebolaget LM Ericsson 2000.
32. Steve D.perkins .1999.A file splitter.version 1.1, http://www.StevePerkins .net
33. Nikolaj Lindberg. 2001, egrep for Linguists

Annex

Annex-1

Neural network architecture for Back propagation algorithm

Neural Network object:

architecture:

numInputs: 1
numLayers: 2
biasConnect: [1; 1]
inputConnect: [1; 0]
layerConnect: [0 0; 1 0]
outputConnect: [0 1]
targetConnect: [0 1]

numOutputs: 1 (read-only)
numTargets: 1 (read-only)
numInputDelays: 0 (read-only)
numLayerDelays: 0 (read-only)

subobject structures:

inputs: {1x1 cell} of inputs
layers: {2x1 cell} of layers
outputs: {1x2 cell} containing 1 output
targets: {1x2 cell} containing 1 target
biases: {2x1 cell} containing 2 biases
inputWeights: {2x1 cell} containing 1 input weight
layerWeights: {2x2 cell} containing 1 layer weight

functions:

adaptFcn: 'trains'
initFcn: 'initlay'
performFcn: 'mse'
trainFcn: 'trainlm'

parameters:

adaptParam: .passes
initParam: (none)
performParam: (none)

trainParam: .epochs, .goal, .max_fail, .mem_reduc,
.min_grad, .mu, .mu_dec, .mu_inc,
.mu_max, .show, .time

weight and bias values:

IW: {2x1 cell} containing 1 input weight matrix
LW: {2x2 cell} containing 1 layer weight matrix
b: {2x1 cell} containing 2 bias vectors

other:

userdata: (user stuff)

Annex-II

Sample data set prepared for the research

Category	Class	Owk1NCalls	Owk2NCalls	Owk3NCalls	Owk4NCalls					
Nwk1NCalls	Nwk2NCalls	Nwk3NCalls	Nwk4NCalls	Owk1dur						
Owk2dur	Owk3dur	Owk4dur	Owkdaydur	Owkenddur						
Odaydur	Onightdur	Ototaldur	Nwk1dur	Nwk2dur						
Nwk3dur	Nwk4dur	Nwkdaydur	Nwkenddur	Ndaydur						
Nnightdur	Ntotaldur	MaxCount	MinCount	AveCount						
STDCount	MaxDur	MinDur	AveDur	STDDur	Tdur					
TnCalls										
18	0	8	8	0	8	6	0	0	6	6944
1950	0	2626	11334	186	5547	5973	11520	528	0	0
2589	2697	420	115	3002	3117	8	0	4.5	3.817254062	
6944	0	1829.625		2363.985614		14637	36			
18	0	4	2	1	5	7	6	2	3	493
431	1678	1917	4370	149	1798	2721	4519	1636	925	291
1092	3813	131	792	3152	3944	7	1	3.75	2.121320344	
1917	291	1057.875		629.7224871		8463	30			
20	0	10	4	0	11	31	6	4	7	1987
1607	0	5080	7811	863	3846	4828	8674	9731	1181	1929
1900	13634	1107	10175	4566	14741	31	0	9.125	9.508454884	
9731	0	2926.875		3098.903098		23415	73			
18	1	6	4	7	8	3	8	3	3	274
669	1484	2676	4587	516	4493	610	5130	465	1042	1544
379	3197	233	3430	0	3430	8	3	5.25	2.251983253	
2676	274	1066.625		811.8694366		8560	42			
18	0	8	3	6	12	7	12	6	14	933
557	1083	6470	7571	1472	5708	3335	9043	2118	3065	1876
4104	10558	605	6491	4672	11163	14	3	8.5	3.77964473	
6470	557	2525.75		1979.233744		20206	68			
20	0	12	5	3	12	6	13	12	7	657
834	510	2268	4032	237	2440	1829	4269	147	784	1081
686	2608	90	1409	1289	2698	13	3	8.75	3.918819064	
2268	147	870.875		625.5471748		6967	70			
20	0	4	8	4	4	3	5	5	1	1924
4381	1957	1022	7616	1668	2903	6381	9284	19	1488	1103
1516	3543	583	2468	1658	4126	8	1	4.25	1.982062418	
4381	19	1676.25		1254.712688		13410	34			
20	0	7	2	6	7	1	0	0	0	331
284	1363	2422	3488	912	2050	2350	4400	11	0	0
0	11	0	0	11	11	7	0	2.875	3.226563851	
2422	0	551.375		884.9597146		4411	23			
18	0	3	3	0	4	3	3	4	4	509
3107	0	438	3815	239	335	3719	4054	396	199	424

185 1128 76 953 251 1204 4 0 3 1.309307341
3107 0 657.25 1004.203416 5258 24
18 0 21 1 19 29 0 0 0 0 2716
398 3659 5718 11026 1465 10334 2157 12491 0 0 0
0 0 0 0 0 0 29 0 8.75 12.13907504
5718 0 1561.375 2207.640301 12491 70
20 1 0 0 0 0 0 0 0 1 0
0 0 0 0 0 0 0 0 0 0
16 16 0 16 0 16 1 0 0.125 0.353553391
16 0 2 5.656854249 16 1
18 0 7 5 3 8 6 7 11 0 441
622 650 2122 3628 207 2523 1312 3835 832 1011 1060
0 1919 984 1633 1270 2903 11 0 5.875 3.313931631
2122 0 842.25 617.5543354 6738 47
20 1 0 1 0 0 0 0 0 0 0
169 0 0 169 0 169 0 169 0 0 0
0 0 0 0 0 0 1 0 0.125 0.353553391
169 0 21.125 59.75052301 169 1
20 0 0 0 1 0 0 1 0 0 0
0 162 0 162 0 162 0 162 0 39 0
0 0 39 39 0 39 1 0 0.25 0.46291005
162 0 25.125 56.96474474 201 2
18 0 1 6 5 1 2 5 1 4 1190
1550 1254 1123 4573 544 4644 473 5117 132 2160 90
1425 3561 246 562 3245 3807 6 1 3.125 2.100170061
2160 90 1115.5 698.9044488 8924 25
18 0 6 0 2 7 5 5 5 5 667
0 2005 2338 3701 1309 4424 586 5010 1532 434 680
1834 3957 523 4416 64 4480 7 0 4.375 2.263846285
2338 0 1186.25 848.105578 9490 35
18 0 2 3 5 2 3 4 5 3 1296
722 948 635 3601 0 3601 0 3601 1103 532 1775
543 3012 941 3799 154 3953 5 2 3.375 1.187734939
1775 532 944.25 433.4366818 7554 27
20 1 2 4 0 2 0 0 0 0 4697
2238 0 329 7264 0 4760 2504 7264 0 0 0
0 0 0 0 0 0 4 0 1 1.511857892
4697 0 908 1714.76704 7264 8
18 0 4 9 1 4 1 0 2 3 310
2918 70 769 3302 765 2236 1831 4067 20 0 103
391 413 101 514 0 514 9 0 3 2.828427125
2918 0 572.625 981.8190024 4581 24

DECLARATION

This thesis is my original work and has not been submitted as a partial requirement for a degree in any university.

Jember Gebreselassie
January 2005

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

Tesfaye Birru (Ato)