
# ADDIS ABABA UNIVERSITY

## COLLEGE OF NATURAL SCIENCES

## SCHOOL OF INFORMATION SCIENCE

# CONSTRUCTING PREDICTIVE MODEL FOR SUBSCRIPTION FRAUD DETECTION USING DATA MINING TECHNIQUES: THE CASE OF ETHIO-TELECOM

BY

TESFAY HADDISH

JUNE 2013

Addis Ababa, Ethiopia

# ADDIS ABABA UNIVERSITY

## COLLEGE OF NATURAL SCIENCES

## SCHOOL OF INFORMATION SCIENCE

# CONSTRUCTING PREDICTIVE MODEL FOR SUBSCRIPTION FRAUD DETECTION USING DATA MINING TECHNIQUES: THE CASE OF ETHIO-TELECOM

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

BY

TESFAY HADDISH

JUNE 2013

Addis Ababa, Ethiopia

# ADDIS ABABA UNIVERSITY

## COLLEGE OF NATURAL SCIENCES

## SCHOOL OF INFORMATION SCIENCE

CONSTRUCTING PREDICTIVE MODEL FOR SUBSCRIPTION FRAUD DETECTION USING DATA MINING TECHNIQUES: THE CASE OF ETHIO-TELECOM

BY

TESFAY HADDISH

Name and signature of members of the examining board

| Name | Title | Signature | Date |
|---|---|---|---|
| _____ | Chairperson | _____ | _____ |
| Workshet Lamenew | Advisor | _____ | _____ |
| _____ | Examiner | _____ | _____ |

# DEDICATION

This research is dedication to the ― Almighty God‖ who is always there for me in all hard time and challenge of my life.

# ACKNOWLEDGMENT

First and foremost extraordinary thanks go for my Almighty God, for giving me the ability to be where I am now and for giving me my lovely son Ziada Tesfay while I was studying this postgraduate program.

Secondly, I would like to express my gratitude and heartfelt thanks to my advisor, Ato Workshet Lamenew for his keen insight, guidance, and unreserved advising. I am really grateful for his constructive comments and critical readings of the study.

Thirdly, I would like to express my sincerest gratitude and heartfelt thanks to my instructor Dr. Million Meshesha for his support and I would like to articulate my appreciation for his academic commitment. I would also like to express my sincerest gratitude and heartfelt thanks to Dr. Tibebe Bashah.

Fourthly, my special thanks goes to my friend and classmate Tsegaye Semere and Berihu Yohannes, your support and advice has been worth considering starting from the beginning up to the completion of the program.

Fifthly, I am very grateful to the management and staff of Ethio Telecom, especially Ato Hailu Haftu, Ato Asfeteaw Abay, Ato Ephrem Abebe and Ato Abraham. My special thanks also goes to Information Network Security Agency (INSA), Capt. Binyam Tewelde and Ato Ahmed Ibrahim for providing me appropriate professional comments and for their cooperation to give me valuable resources which are relevant for my study.

Last but not least, I would like to thank all my families and friends for the constant assistance and encouragement during the time of postgraduate program.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ACRONYMS

| | |
|---|---|
| CRISP-DM | Cross-Industry Standard Process -Data Mining |
| SEMMA | Sample, Explore, Modify, Model, Assess |
| CSV | Comma Separated Values |
| ARFF | Attribute Relation File Format |
| RF | Random forests |
| ANN | Artificial Neural Networks |
| DM | Data Mining |
| WEKA | Waikato Environment for Knowledge Analysis |
| FMS | Fraud Management Systems |
| UMTS | Universal Mobile Telecommunications System |
| CFCA | Communications Fraud Control Association |
| ITU | International Telecommunication Union |
| PSTN | Public switching telephone network |
| GSM | Global System for Mobile Communication |
| CDR | Call Detail Record |
| PRF | Premium Rate Fraud |
| IREG | International Roaming Expert Group |
| VoIP | Voice over Internet Protocol |
| INSA | Information Network Security Agency |
| 1G/2G/3G | First / Second /Third Generation |
| PBX | Private Branch Exchange |
| PIN | Personal Identification Number |
| PRS | Premium Rate Service |
| SIM | Subscriber Identity Module |
| MIN | Mobile Identification Number |
| ESN | Electronic Serial Number |

# ABSTRACT

Access to telecommunications is critical to the development of all aspects of a nation's economy including manufacturing, banking, education, agriculture and government.

However, the telecommunication services are not free from problem. Telecommunication fraud is the main problem of all telecom operators. Telecommunication fraud is the theft of telecommunication service (telephones, cell phones, computers etc.) or the use of telecommunication service to commit other forms of fraud. Victims include consumers, businesses and communication service providers.

The subscription fraud is the most prevalent since with a stolen or manufactured identity, there is no need for a fraudster to tackle a digital network's encryption or authentication systems.
This study is initiated with the aim of exploring the potential applicability of the data mining technology in developing models that can detect and predict pre paid mobile subscription fraud in Ethio-telecom service provision.

The researcher selected around 25,000 records from six months collection of Call Detail Record data. After eliminating irrelevant and unnecessary data only a total of 21367 datasets are used for the purpose of conducting this study. The researcher also selected 14 attributes for this study based on their relevant for this research. The collected data has been preprocessed and prepared in a format suitable for the DM tasks.

The study was conducted using WEKA software version 3.7.9 and four classification techniques namely J48, PART, Random forest and Multilayer perceptron of artificial neural network. As a result the Random forest algorithm registered better performance of 99.9251% accuracy running with 10-fold cross validation and used 14 attribute for this experimentation of this research. Future works are also implicated in this work.

# CHAPTEER ONE

# 1. INTRODUCTION

## 1.1. Background of the Study

Globally, the development of telecommunication industry is one of the important indicators of social and economic development of a given country. In addition to this, the development of communication sector plays a vital role in the overall development of all sectors related to social, political and economic affairs. This sector is very dynamic in its nature of innovation and dissemination (Taye, 2010).

Telecommunication Fraud is the intentional and successful employment of any deception, cunning, collusion, artifice, used to circumvent, cheat, or deceive another person, whereby that person acts upon it to the loss of his property and to his legal injury. However, there does seem to be a general consensus that telecom fraud, as the term is generally applied, involves the theft of services or deliberate abuse of voice and data networks. Furthermore, it is accepted that in these cases the fraudster's intention is to completely avoid or at least reduce the charges that would legitimately have been charged for the services used. On occasion, this avoidance of call charges will be achieved through the use of deception in order to fool billing and customer care systems into invoicing the wrong party (Gary, 2005).

Fraud is different from revenue leakage. Revenue leakage is characterized by the loss of revenues resulting from operational or technical loopholes where the resulting losses are sometimes recoverable and generally detected through audits or similar procedures. Fraud is characterized with theft by deception, typically characterized by evidence of intent where the resulting losses are often not recoverable and may be detected by analysis of calling patterns.
Telecommunications is an attractive target for fraudsters. In terms of volume, it is now measured in the billions worldwide. Recent highly sophisticated schemes are employed by organized crime using hackers and self learning; Estimated that telecommunications fraud is more attractive than the drug trade (Brown, 2005).

The Communications Fraud Control Association conducted a survey and determined that $72–$80 billion in losses are due to telecom fraud worldwide (CFCA, 2009). While many large operators have developed sturdy, Fraud Management Systems (FMS) to combat fraud, others have not. The motivation behind crime is attributed to migration and demographics, penetration of new technology, staff dissatisfaction, the 'challenge factor', operational weaknesses, poor business models, criminal greed, money laundering and political and ideological factors (Brown, 2005).

According to Brown (2005), there are so many types of fraud. However, the most common types of fraud are listed here. Subscription Fraud, SIM Box (Bypass) Fraud, Premium Rate Fraud (PRF) Internal Fraud, Prepaid Fraud, Post paid Fraud, Cloning Fraud, Roaming Fraud, PBX (Private Branch Exchange) fraud etc. Subscription is among the most common frauds due to the low technical knowledge required to perform the fraud.

Subscription Fraud: The subscription fraud is the most prevalent since with a stolen or manufactured identity, there is no need for a fraudster to tackle a digital network's encryption or authentication systems. Subscription fraud is one of the fraudster's preferred methods for digital roaming fraud. The modus operandi of subscription fraudster is posing as a credit worthy person or company; the fraudster can gain access to any network, anywhere 1G, 2G or 3G. Typically the first step for fraudsters is to use subscription fraud to gain access to the home network (Brown, 2005).

Industries like telecommunication produce and accumulate huge amount of data. These data comprise call detail data, network data and customer data. Because of its hugeness the data is difficult to analyze manually. Therefore, we need a mechanism to handle this data. As a result the development of knowledge-based expert systems and automated systems performed important functions such as identifying telecom frauds (Gary, 2005).

Data Mining also called Knowledge Discovery in Databases (KDD) can be defined as the science of extracting useful information from large datasets or databases. Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful

information, i.e. information that can be used to increase revenue, cut costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases (Ojuka, 2005).

The telecommunication industry was an early adopter of data mining technology and therefore many applications exist. Four major applications include: marketing customer profiling, fraud detection, churn management and network fault isolation (Gary, 2005).

Telecommunication companies utilize data mining technique to improve their marketing efforts, identify fraud and better manage their telecommunication network connection (Abidogun, 2005). In this paper the researcher studied an effective solution to identify the fraud detection in telecommunications using Data Mining classification techniques.

These days, the use of DM techniques as a support in business decision-making is growing fast. Different attempts have been made to apply the DM techniques in solving business problems in Ethiopia. Some of the attempts are made for insurance, Banking, Airlines, and telecommunication.

### 1.1.1. Ethiopian Scenario

The Ethiopian Telecommunication Corporation (ETC) is the oldest public telecommunication operator (PTO) in Africa. The first long-distance telephone line was established in 1894 between Addis Abeba and Harar (ITU, July 2002). The network has begun to expand starting from then.

Although the ETC was initially private company, it was placed under government control at the beginning of the twentieth century and was later brought under the control of the Ministry of Post and Communications (ITU, July 2002).

Currently, ETC is providing all types of services (PSTN, cellular Mobile, Internet and data communication services) in all parts of the country. The government of Ethiopia has aggressively been moving and implementing development strategies aimed at reducing the poverty prevailing

in the country. In this line, telecommunication plays a key role in facilitating the poverty reduction and development strategy being implemented by the government. To this end, the government has designed strategies to expand telecommunication national network infrastructure (Taye, 2010).

Now a days the Ethiopian Telecommunication Corporation or Ethiopian telecommunication agency has changed its name to Ethio telecom and its network coverage and services is improved.

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 (Redl et al., 1995).

According to Jember (2005) as cited the information released from Ethiopia Telecommunication Corporation Ethio-Mobile presently operates in the GSM 800 MHZ frequency ranges covering Addis Ababa, Debre Zeit, Nazareth, Modjo and Sodere.

However, now according to Ethiopian Information and Communication Technology (2012), Ethiopia's mobile subscribers as Dec 11, 2012 over 18 million as Debretsion Gebremichael, Information and Communication Technology Minister with the rank of Deputy Prime Minister, said that the government is trying to increase access to Information and Communication Technology in the country as well as improve its quality. Opening the 2nd Annual Ethiopian ICT Entrepreneurship Conference on Monday (December 10$^{th}$ 2012), Debretsion said ICT infrastructure was expanding. The number of mobile subscribers in the country now reached over 18 million and of these, the Minister said, 3.4 million mobile subscribers used the system to access internet service. The aim is able to reach 45 million subscribers by the end of the 2014. He noted that wireless coverage of the country had now reached 73 percent (Ethiopian ICT, 2012).

In the Ethiopia scenario the government of Ethiopia gives more attention to the situation of telecom fraud. One example is the new declaration or recently adopted Proclamation 761/2012 on Telecom Fraud Offences. The Ethiopia House of Peoples' Representatives passed the Proclamation on 11 July 2012. It creates new offences related to the use and provision of telecommunications services, and increases sentences for a number of existing offences.

The Proclamation identifies "telecom fraud offences" as a "serious threat to national security beyond economic losses". The Proclamation on Telecom Fraud Offences must be repealed in its entirety. And new law is needed because of the existing laws are not adequate to combat these threats ARTICLE 19 (Government of Ethiopia, 2012).

## 1.2. Statement of the problem

Telecommunication fraud is the theft of telecommunication services such as telephones, cell phones, computers etc or the use of telecommunication service to commit other forms of fraud. Victims of telecommunication fraud include consumers, businesses and communication service providers (Olson, & Delen, 2008).

Easy ways to for fraudsters is to use subscription fraud to gain access to the home network. This way, they appear on the network as a valid subscriber accepted by the digital network and authentication system. Often fraudsters work with corrupt dealers or internal groups within the service provider in order to create the subscriber accounts estimated that telecommunications fraud is more attractive than the drug trade (Brown, 2005).

Mobile telephone had begun in Addis Ababa in April 1999. In Addis Ababa, the capital city of Ethiopia, the number of mobile subscribers was 51777 as of June 2003. In spite of the fact that the report released that more than 100,000 mobile subscribers are found in Addis Ababa, Nazareth, Jimma, Bahridar and neighboring cities (Jember, 2005). Jember attempted to predict fraud in post-paid mobile phone and the data he used restricted on three months only and that data was taken from call detail record (CDR) and he was used matlab software for windows as a tool and artificial neural network algorithm as data mining technique used for his work.

Another study was done by Geremeskel (2006). The data source of this study was taken from call detail record (CDR) and the tool he used was SPSS software and he also used artificial neural network data mining technique. This data was taken from two months record only. Additionally, the study was targeted at prepaid mobile service. The number of postpaid mobile subscriber was 53614 and 381417 was prepaid. In May 2005 the total number of mobile subscribes was 435031(Geremeskel, 2006).

Nowadays Ethiopian government is trying to increase access to Information and Communication Technology in the country as well as improve its quality. In addition ICT infrastructure was expanding with the number of mobile subscribers in the country reaching over 18 million and of these 3.4 million mobile subscribers used the system to access internet service. The aim will be able to reach 45 million subscribers by the end of the 2014. Besides, wireless coverage of the country had also reached 73 percent (ICT, 2012).

Some of the problems with previous works are: First they used short time record data, second because of small number of subscriber the test data was very small, in addition the area coverage they consider was very limited when it is compare with today's coverage and both researchers used only neural network algorithm without clear methodology.

So we can conclude that the previous researches do not represent the current situation.
Hence, this research is initiated to experiment on the current status of subscription fraud in the Ethio-telecom using data mining technique by considering the above problems.

This research attempts to answer the following research questions.

- What are the natures and traits of fraudsters in Ethio -telecom?
- How they detect the subscription frauds in the Ethio- telecom?
- What possible fraud preventive mechanisms can be setup in Ethio-telecom?

## 1.3. Objective

### 1.3.1. General objective

The general objective of this research is to develop detecting and predicting model for subscription fraud using data mining techniques in Ethiopia mobile communication services provision.

### 1.3.2. Specific objectives

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

- To review different literatures on DM techniques and their applications in the telecom industry so as
- To understand the domain area through different discussions and interviews.
- To select and use appropriate training and test data
- To identify appropriate data mining algorithm and tools for telecom frauds detection
- To prepare the data for analysis and model building, by cleaning and transforming the data into a format suitable for the selected DM algorithms
- To train and develop a classification model that will help to predict and detect telecom fraud
- To evaluate and compare the performance of different DM algorithms and recommend the overall best results of classification models.
- To report the result and forward recommendation for further research

## 1.4. Methodology of the study

Methodology is a way that deals with data collection, analysis and interpretation that shows how researcher achieves the objectives and answers the research questions. Hence, in order to achieve the general and specific objectives of the study the following method was used

### 1.4.1. Research design

For the purpose of conducting this research the six-step process model of Hybrid Model is selected. This model was developed, by adopting the CRISP-DM model to the needs of academic research community. Unlike the CRISP-DM process model, which is fully industrial, the Hybrid of process model is both academic and industrial.

The main extensions of the Hybrid of process model include providing a more general, research oriented description of the step, an introduction of a DM step instead of the modeling step, and an integration of several explicit feedback mechanisms (Cios, at el., 2007).

The overall design issues according to Cios, at el. (2007) or the Hybrid Data mining Methodology are represented diagrammatically as shown in figure 1.1.

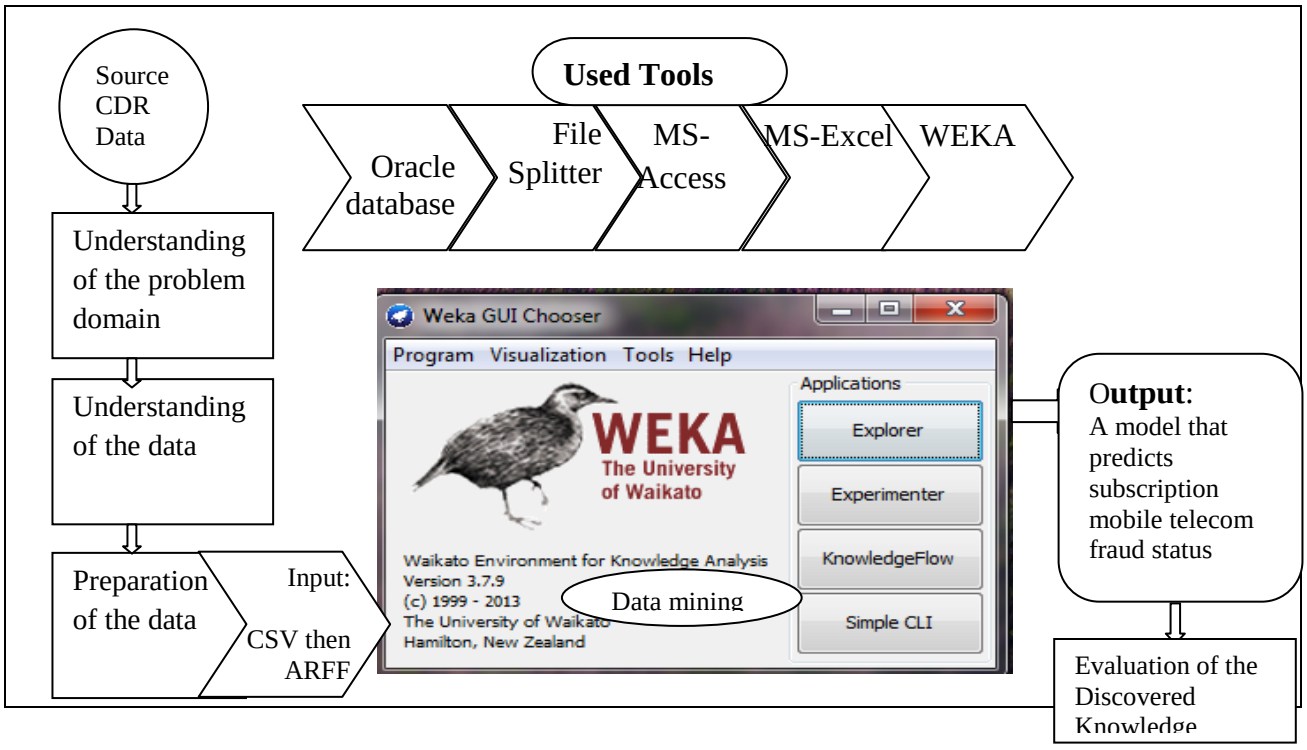



Figure 1.1 Diagrammatic overview of the overall research design & methodology

**Understanding of the problem domain**:

In ordered to identify, define, understand and formulate the problem domain the researcher used different discussion points as shown in appendix 10. Those discussion points reflects telecom subscription frauds, so that the researcher closely works with the domain experts of Ethio-telecom and INSA, then determine attribute feature selection and understand some complex business process. In collaborating with the domain experts the CDR data is selected as the main source of data collection. As a result of insight gained knowledge of the domain of telecom business and data mining problem is defined.

**Understanding of the data**

The researcher briefly described the CDR data. The description includes listing out initial attributes, their respective values and the researcher in collaborating with domain experts of Ethio - telecom evaluation was made how the CDR is importance for this research.

**Preparation of the data**

The researcher decides, together with domain experts, the CDR data is used as input for applying the DM techniques. The cleaned data is further processed by feature selection consulting the domain experts and the Weka attribute selection preprocessing techniques to reduce dimensionality and by derivation of new attributes. The result of these processes generates datasets for training and testing of the classification algorithms selected in this study.

**Data mining**

The main purpose of this study is to develop predicting model for detecting telecom subscription frauds using data mining techniques in Ethiopia mobile communication services provision. The researcher selected classification technique because the research dataset has clear and simplified labeled class.

**Evaluation of the discovered knowledge**

In this research different classification models were developed and evaluated using training and testing dataset. The experimental output of the classification models are analyzed and evaluated the performances accuracy using confusion matrix.

The classifier is also made in terms of different confusion matrices (True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), False Negative Rate (FNR), Relative Operating Characteristics (ROC)), the number of correctly classified instances, number of leaves and the size of the trees, execution time. The template confusion matrix of the model used to classify subscription telecommunication fraud status is presented in table 1.1.

| | | Predicted telecom fraud Status | | |
|---|---|---|---|---|
| | | fraud | Nonfraud | Total |
| Actual fraud Status | Fraud | TP | FN | TP + FN |
| | Nonfraud | FP | TN | FP +TN |
| | Total | TP + FP | FN+ TN | TP + FP + FN + TN |

Table 1-1 Summary of confusion matrix model

Key: TN = True Negative FP= False Positive FN= False Negative TP= True Positive

Negative: None fraudster Subscriber

Positive: fraudster Subscriber

Contextually what each the above variable stands for is stated as follows

TN: The number of fraud negative Subscribers that are classified as negative.

FN: The number of fraud positive Subscribers that are classified as negative.

TP: The number of fraud positive Subscribers that are classified as positive.

FP: The number of fraud negative Subscribers that are classified as positive.

TP + FN: The total number of actually fraud positive Subscribers.

FP + TN: The total number of actually fraud negative Subscribers.

TP + FP: The total number of predicted fraud positive Subscribers.

FN +TN: The total number of predicted fraud negative Subscribers.

TN + FP + FN + TP: The total number of all Subscribers.

Correctly Classified Instances (Accuracy): To compute the proportion of clients those are correctly classified using this formula i.e. Accuracy.

Accuracy = [(TP+TN)*100%]/TP+TN+FP+FN

Incorrectly Classified Instances (Error Rate): The proportion of clients that are incorrectly classified i.e. Error Rate.

Error Rate = [(FP + FN) *100%]/ / (TN + FP + FN + TP).

Furthermore we can also compute the effectiveness and efficiency of the model in terms of recall and precision. As a result, recall can be computed for both positive and negative classes.

Precision: Precision is the number of class members classified correctly over the total number of instances classified as class members.

Precision= (TP*100%)/ (TP+FP)

Recall: Recall also called True Positive Rate (TPR), Recall measures the number of correctly classified examples relative to the total number of positive examples. In other words the number of class members classified correctly over the total number of class members

Recall= (TP*100%)/ (TP+FN)

**F-measure**

Precision and Recall stand in opposition to one another, as precision goes up, recall usually goes down (and vice versa). F measure provides a good balance between precision and recall, or the F-measure combines the two values.

$$\mathrm{F-measure} = 2.\frac{\mathrm{precision}\ .\mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

ROC (Receiver Operating Characteristic): The performance of the classifiers with different parameters is also compared by examining their ROC curve. According to Han & Kamber (2006), ROC curves are a useful tool for comparing classification models. They further said that ROC curve shows the trade-off between the TP rate and the FP rate for a given model.

In addition, the researcher have been used the following methodology to develop data mining model that predicted the subscription fraud in the Ethio telecom mobile communication.

### 1.4.2. Literature review

Different previously related literatures books, journals, magazines, manuals,, articles and the internet have seen briefly reviewed in order to have detail understanding on the present research. Relevant materials on data mining and its application, about telecommunication and mobile fraud have been consulted. In addition, the following data collection methods have seen adopted.

### 1.4.3. Data collection

The primary source of data to conduct this research was Call Detail Record (CDR) of huge data of telecommunication record of six month that was about 300 GB per month. The data very huge

then the telecommunication company saved in damp file, so that the researcher extracted that data by using oracle database. Finally append and arranged the six month data in one file and make it suitable for the experiment and for the selected data mining tools by using preprocessing tasks like data cleaning and reduction.

### 1.4.4. Tools and Techniques

In ordered to accomplish the research process, the researcher has been used the following tools and application software.

WEKA version 3.7.9 DM tool has been used to create models using the classification algorithms (like J48, PART, Random forest and ANN).

It is chosen because:

- It is easy to use by a novice user due to the graphical user interfaces it contains
- It is very portable because it is fully implemented in the Java programming language and thus runs on almost any computing platform
- Contains a comprehensive collection of data preprocessing and modeling techniques, and
- It is freely available under the General Public License (GNU)

Oracle Database: The telecommunication record of six month records was very huge data it saved in damp file. To read this file, the researcher has been used Oracle Database and retrieve necessary data in collaborating with domain experts.

File Splitter: the researcher also used the file splitter tool because it is a free Windows program that has been used to split the large Excel file into pieces of small size. Therefore, the researcher continues it work by using MS-Access and finally used MS-Excel.

MS-Access: it was used for pre-processing and taking enough samples because the MS-Access has more rows and columns than MS-Excel.

MS-Excel: it was used for data preparation, pre-processing and analysis task because it has the capability of filtering attribute with different values. Besides, it is a very important application software to make ready the data and easily convert into of the file that accepted by the weka tool.

### 1.4.5. Model building

As indicated in the research objective, the researcher was used data mining technique to develop a model that predicts the telecom subscription fraud. Having in mind the objective, the researcher chose hybrid model, because that model combines the aspects of both academic and industrial models had led to the development of hybrid models.

So, in the experimental part predictive models has been developed by using the selected data mining technique. The research was conducted with four selected techniques of data mining algorithms Classification models that were developed in this research and evaluated using different algorithms the trained and tested dataset and compared the performances to show the better classification accuracy. Finally, a comparison of them was made to get a reasonable accuracy.

## 1.5. Scope and Limitation of the study

The scope of this research is to develop detecting and predictive model for subscription fraud in the Ethio telecom service provision. Today there are many telecom frauds among those subscription fraud is very complex and transversal to the operator structure. Hence, this research was limited to subscription fraud on pre paid mobile communication service in i.e. Ethio telecom. Due to time and financial matters this research didn't include the all types of telecommunication fraud. So, further research can be conducted to be including the other types of telecommunication fraud. Only six month data were used. Since Ethio-telecom do not keep data belong six months.

## 1.6. Research contribution

The result of this study helps to improving the efficiency of fraud detection mechanism of the current threat posed on Ethio telecom service provision. Accordingly, this research will play an important role to controlling and preventive the current threat on the mobile communication of Ethio telecom. Now a day the telecom fraud is very serious problem that is why the ARTICLE 19, declared by the Ethiopia House of Peoples' Representatives passed the Proclamation on 11 July 2012. The Proclamation identifies "telecom fraud offences" as a "serious threat to national security beyond economic losses" so that this research therefore, will open new research areas in the field of telecom fraud detection and preventive mechanism. Generally, this study is important since its application can reduce the revenue losses due to subscription fraud.

The output of the research also contributes to understanding of the theoretical views and practical problems in the detection of telecom fraud. Finally, this study can also be an input for further research in this and other types of telecommunication frauds.

## 1.7. Organization of the thesis

This thesis is organized into six chapters. The first chapter briefly discusses background to the problem area and DM technology, and states the problem, objective of the study, research methodology, scope and limitation, and significance of the results of the research.

The second chapter deals about DM technology, methods/techniques and algorithms, the different methodologies, the DM process, the different tasks of DM. This chapter discusses about different classification techniques like decision tree method of J48 algorithm, PART, Random forest, neural network algorithm.

In the second chapter critical literature review on the telecommunication fraud and which includes the common types of telecom fraud like subscription fraud, premium rate fraud, internal fraud, prepaid fraud, cloning fraud, roaming fraud SIM box fraud and PBX fraud and also provides the detection and prevention techniques of telecom fraud. Furthermore this chapter reviews local and global related works.

The fourth chapter provides introduction about telecommunication and discussions about the different DM steps that are undertaken by the methodology used in this research work. This includes the business understanding, data understanding, data cleaning, reduction and preparation of dataset to be used as input for predictive model.

The fifth chapter deals with experimentations and result interpretations. In this chapter building of model with training dataset and validating the result with testing datasets, and interpretation of the result of the experimentation were the major concern. The last chapter is devoted to concluding remarks and recommendations forwarded based on the research findings of the present study.

# CHAPTER TWO

# 2. DATA MINING AND KNOWLEDGE DISCOVERY

## 2.1. Overview of Data Mining and Knowledge Discovery Process

### 2.1.1. What is Data Mining?

Information is available at anytime anywhere because of the rapid growth of World Wide Web and electronic information services. The invented machines are faster in producing, manipulating and disseminating information. In information age, an appropriate usage and organization of the information helps to be powerful and to achieve goals. Hence, information processing mechanisms such as automatic data summarization, information extraction and discovering hidden knowledge are very important. (Osmar, 1999).

Data mining is the process of extracting or mining knowledge from large data sets. But, knowledge mining from data can describe the definition of data mining even if it is long. data mining have similar or a bit different meaning with different terms , such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging (Han &Kamber 2006).

Data mining can be also considered as an exploratory data analysis (Olson &Delen (2008). Generally, Data mining uses advanced data analysis tools to find out previously unknown (hidden), valid patterns and relationships among data in large data sets.
As can be seen from Figure 2.1 data mining is the core field for different disciplines such as database, machine learning, and pattern recognition etc.

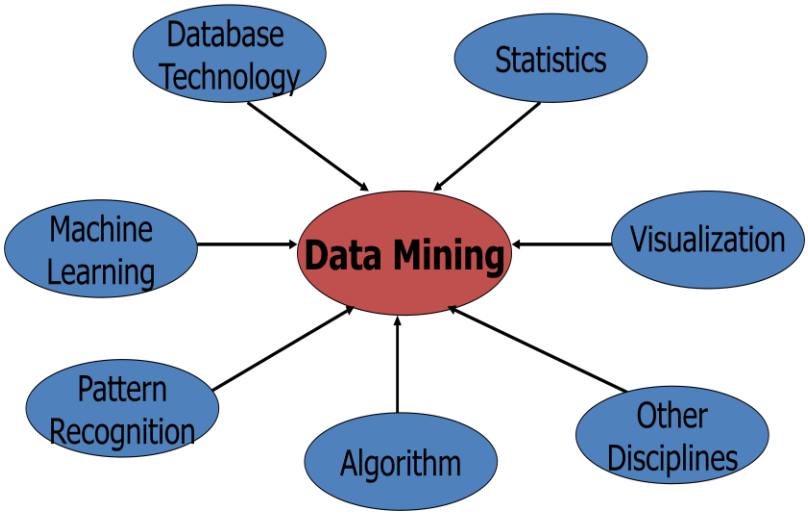



Figure 2.1 Data mining: confluence of multiple disciplines

### 2.1.2. Why Data Mining?

Nowadays, massive amount of data is produced and collected incrementally. The possibility of gathering and storing huge amount of data by different organizations is becoming true because of using fast and less expensive computers. When organizational data bases keep growing in number and size due to the availability of powerful and affordable database systems the need for new techniques and tools became very important. These tools are used for helping humans to automatically identify patterns, transform the processed data into meaning full information in order to draw concrete conclusions. In addition it helps us to take out hidden knowledge from huge amount of digital data (Seifert, 2004).

In the private sector industries such as banking, insurance, medicine, telecommunication and retailing use data mining to reduce costs, enhance research, and increase sales. Different organizations worldwide are using data mining techniques for Appling and locating higher value customers and to reconfigure their product offerings to increase sales. In the public sector, data mining applications initially were used as a means for detecting fraud and waste of materials, but it grown for different purposes such as measuring and improving program performance (Seifert, 2004).

Data mining is widely used by banking firms in seeking customers' credit card, by insurance and telecommunication companies in detecting fraud (Weng, Chiu, Wang& Su 2006). In addition it is used to improve food and drug product safety, and detection of terrorists or criminals.

### 2.1.3. Data Mining and Knowledge Discovery (DMKD)

It is common practice to refer to the idea of searching applicable patterns in data using different names such as data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. Among these terms KDD and data mining are used widely (Fayyad et al, 1996).

Knowledge discovery was coined at KDD to emphasize the fact that knowledge is the end product of a data-driven discovery and that it has been popularized in the artificial intelligence and machine learning fields. According to (Fayyad et al. 1996) KDD and data mining are two different terms. KDD refers to the overall process of discovering useful knowledge from data and data mining refers to a particular step in the process. Furthermore, data mining is considered as the application of specific algorithms for extracting patterns from data.

### 2.1.4. Overview of the Knowledge Discovery Process and Models

The knowledge discovery process (KDP), also called knowledge discovery in data base is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The process consists of many step, among which is data mining. Each step attempts to complete a particular discovery task and is accomplished by the application of a discovery method.

Knowledge discovery process includes Knowledge extraction processes (like data access and storage, analysis of massive datasets using scalable and efficient algorithms and the interpretation and visualization of the result (Cios et al., 2007).

### 2.1.5. Knowledge Discovery Process Models (KDPM)

A Knowledge discovery process model provides an overview of the life cycle of a data mining project, containing the corresponding phases of a project, their respective tasks, and relationships between these tasks.

There are different process models originating either from academic or industrial environments all of which have in common the fact that they follow a sequence of steps which more or less resemble between models. Although the models usually emphasize independence from specific applications and tools, they can be broadly divided into those that take into account industrial issues and those that do not. However, the academic models, which usually are not concerned with industrial issues, can be made applicable relatively easily in the industrial setting and vice versa (Cios et al., 2007).

**The KDD Process**

The basic steps of data mining for knowledge discovery (KDD) are: defining business problem, creating a target dataset, data cleaning and pre-processing, data reduction and projection, choosing the functions of data mining, choosing the data mining algorithms, data mining, interpretation, and using the discovered knowledge. A short description of these steps follows in the coming paragraphs (Fayyad, et al., 1996b). The KDD process is shown diagrammatically in Figure 2.2 below, source:(Fayyad, et al., 1996b).

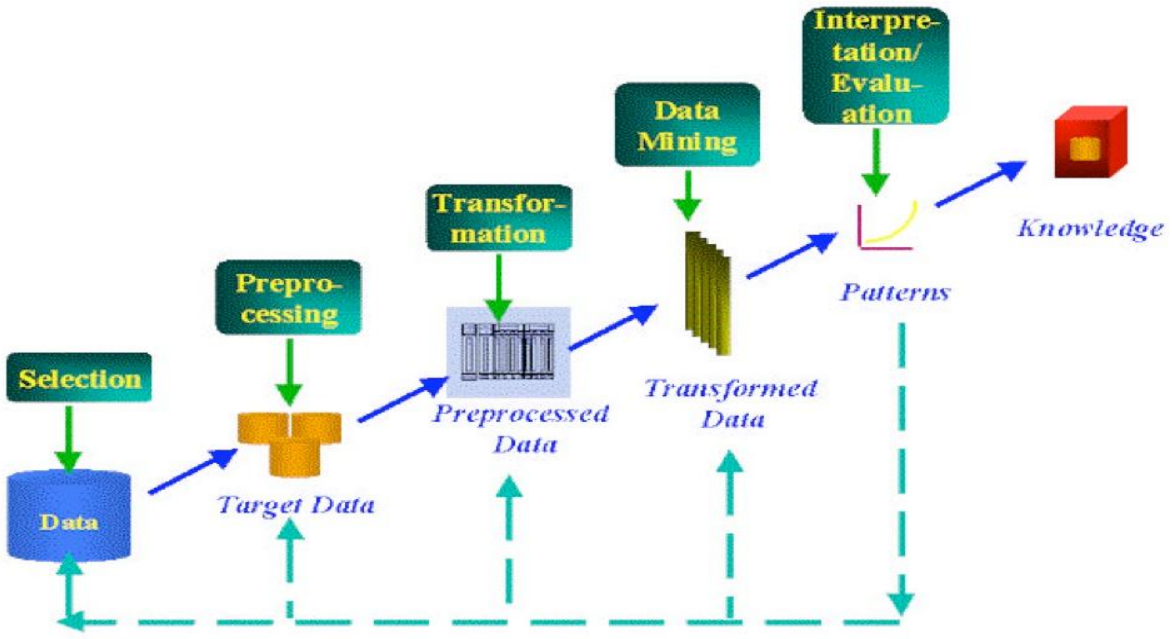



Figure 2.2 The five stages of KDD

**Selection -** This stage is concerned with creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed by Understanding the data and the business area. Because, Algorithms alone will not solve the problem without having clear statement of the objective and understanding**.**

**Pre-processing –** this phase is concerned in removing noise or outliers if any, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes. On top of these tasks, deciding on DBMS issues, such as data types, schema, and mapping of missing and unknown values are parts of data cleaning and pre-processing.

**Transformation** - The transformation of data using dimension reduction or transformation methods is done at this stage. Usually there are cases where there are large numbers of attributes in the database for a particular case. With the reduction of dimension there will be an increase in the efficiency of the data-mining step with respect to the accuracy and time utilization.

**Data Mining -** this phase is the major stage in data KDD because it is all about searching for patterns of interest in a particular representational form or a collection of such representations. These representations include classification rules or trees, regression, clustering, sequence modeling, dependency, and line analysis. Therefore, selecting the right algorithm for the right area is very important.

**Evaluation –** In this stage the mined data is presented to the end user in a Human viewable format. This involves data visualization, where the user interprets and understands the discovered knowledge obtained by the algorithms.

**Using the Discovered Knowledge -** Incorporating this knowledge into a performance system, taking actions based on the knowledge, or simply documenting it and reporting it to interested parties, as well as checking for and resolving for conflicts with previously acquired knowledge are tasks in this phase.

Knowledge discovery (KDD) as a process consists of an iterative sequence of steps as discussed above. It is also clear that data mining is only one step in the entire process, though an essential one, it uncovers hidden patterns for evaluation.

### *2.1.5.1.CRoss-Industry Standard Process for Data Mining (CRISP-DM)*

Industrial models quickly followed academic efforts. CRISP-DM is a model established by four companies: Integral Solutions Ltd. a provider of commercial data mining solutions, NCR a database provider, DaimlerChrysler an automobile manufacturer, and OHRA an insurance company. It is a Cross-Industry Standard Process for Data Mining widely used by industry members. This model consists of six phases intended as a cyclical process (Azevedo, 2008).

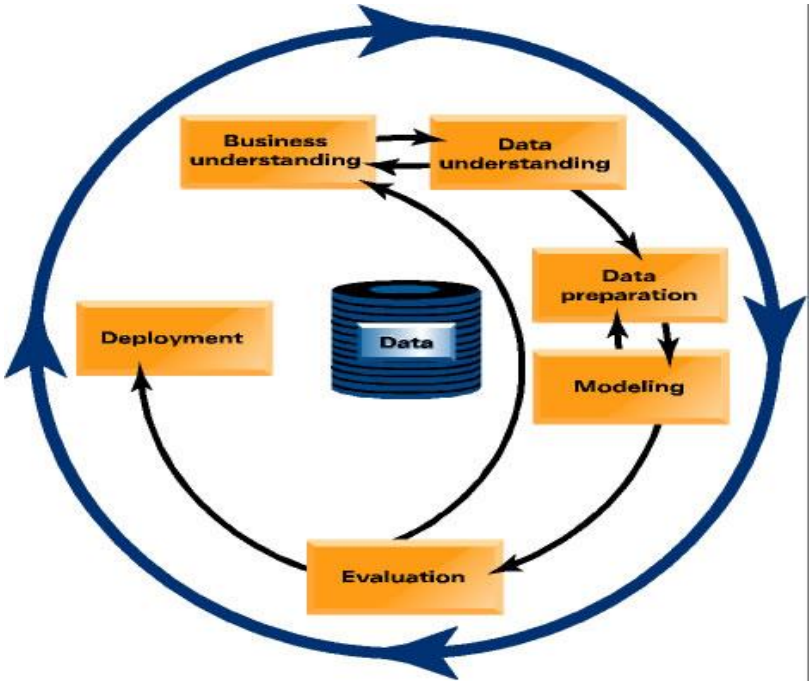



Figure 2.3 The CRISP-DM process

The sequence of the six stages is not rigid, as is schematize in figure 2.3 CRISP-DM is extremely complete and documented. All his stages are duly organized, structured and defined, allowing that a project could be easily understood or revised although the CRISP-DM process is independent from DM chosen tool, it is linked to the SPSS Clementine software (Fayyad et al., 1996).

**The discussion for all the phases of CRISP-DM are below.**

**Business Understanding:** Business understanding includes determining business objectives, assessing the current situation, establishing data mining goals, and developing a project plan.

The key element of a data mining study knows what the study is for. This begins with a managerial need for new knowledge, and an expression of the business objective regarding the study to be undertaken. Goals in terms of things such as "What types of customers are interested in each of our products?" or "What are typical profiles of our customers, and how much value does each of them provide to us?" are needed. Then a plan for finding such knowledge needs to be developed, in terms of those responsible for collecting data, analyzing data, and reporting. At this stage, a budget to support the study should be established, at least in preliminary terms (Azevedo, & Santos, 2011).

In customer segmentation models, such as Fingerhut's retail catalog business, the identification of a business purpose meant identifying the type of customer that would be expected to yield a profitable return. The same analysis is useful to credit card distributors. For business purposes, grocery stores often try to identify which items tend to be purchased together so it can be used for affinity positioning within the store, or to intelligently guide promotional campaigns.

**Data Understanding**: Once business objectives and the project plan are established, data understanding considers data requirements. This step can include initial data collection, data description, data exploration, and the verification of data quality. Data exploration such as viewing summary statistics which includes the visual display of categorical variable scan occurs at the end of this phase. (Olson, &Delen, 2008).

Since data mining is task-oriented, different business tasks require different sets of data. The first stage of the data mining process is to select the related data from many available databases to correctly describe a given business task. There are at least three issues to be considered in the data selection (Pete, 2000).

Variable independence is achieved by making sure that the variables do not contain overlapping information. Useful knowledge patterns are discovered by algorithms through a careful selection of independent variables.

Data type such as quantitative data is measurable using numerical values can be either discrete or continuous. The data can be readily represented by some sort of probability
Distribution describing how it is dispersed and shaped. For instance, normally distributed data is symmetric, and is commonly referred to as bell-shaped.

Qualitative data type, also known as categorical data, contains both nominal and ordinal data. Nominal data has finite non-ordered values, whereas ordinal data has finite ordered values. Customer credit ratings are considered ordinal data since the ratings can be excellent, fair, and bad. Qualitative data may be first coded to numbers and then be described by frequency distributions (Olson, &Delen, 2008).

**Data Preparation:** Once the data resources available are identified, they need to be selected, cleaned, built into the form desired, and formatted. The purpose of data pre processing is to clean selected data for better quality. The data selected may have different formats as the selection could be from different sources. If selected data are from flat files, voice message, and web text, they should be converted to a consistent electronic format.

In general, data cleaning means to filter, aggregate, and fill in missing values. By filtering data, the selected data are examined for outliers and redundancies. Outliers differ greatly from the majority of data, or data that are clearly out of range of the selected data groups (Azevedo, & Santos, 2011).

**Modeling:** Data modeling is where the data mining software is used to generate results for various situations. A cluster analysis and visual exploration of the data are usually applied first. Depending upon the type of data, various models might then be applied. If the task is to group

data, and the groups are given, discriminate analysis might be appropriate. If the purpose is estimation, regression is appropriate.

**Evaluation:** Model results should be evaluated in the context of the business objectives established in the first phase. This will lead to the identification of other needs often through pattern recognition, frequently reverting to prior phases of CRISP-DM. Gaining business understanding is an iterative procedure in data mining, where the results of various visualization, statistical, and artificial intelligence tools show the user new relationships that provide a deeper understanding of organizational operations.

Two issues are essential in the data interpretation stage, how to recognize business value from the knowledge patterns discovered and which visualization tool should be used to show the mining results (Olson, &Delen, 2008).

Visualization packages and tools such as pie charts, histograms, box plots, scatter plots, and distributions are available. Good interpretation leads to productive business decisions, while poor interpretation analysis may miss useful information.

**Deployment:** The results of the data mining study need to be reported back to project sponsors. The data mining study has uncovered new knowledge, which needs to be tied to the original data mining project goals. Management will then be in a position to apply this new understanding of their business environment.

It is important that the knowledge gained from a particular data mining study be monitored for change. Customer behavior changes over time and what was true during the period when the data was collected may have already changed. If fundamental changes occur, the knowledge uncovered is no longer true. Therefore, it's critical that the domain of interest be monitored during its period of deployment (Olson, &Delen, 2008).

Data mining can be used to both verify previously held hypotheses, or for knowledge discovery identification of unexpected and useful relationships. Through the knowledge discovered in the earlier phases of the CRISP-DM process, sound models can be obtained that may then be applied to business operations for many purposes, including prediction or identification of key situations (Larose, 2004).

This six-phase process is not a rigid, by-the-numbers procedure. There's usually a great deal of backtracking. Additionally, experienced analysts may not need to apply each phase for every study. But CRISP-DM provides a useful framework for data mining (Olson, &Delen, 2008).

#### *2.1.5.2. The SEMMA Process*

The SEMMA process was developed by the SAS (Statistical Analysis System) Institute. The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a data mining project. The SAS Institute considers a cycle with 5 stages for the process (Matignon, & SAS Institute, 2007).

In order to be applied successfully, the data mining solution must be viewed as a process rather than a set of tools or techniques. Beginning with a statistically representative sample of your data, SEMMA intends to make it easy to apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and finally confirm a model's accuracy. A pictorial representation of SEMMA is given in figure 2.4 (Kurgan, &Musilek, 2006).

By assessing the outcome of each stage in the SEMMA process, one can determine how to model new questions raised by the previous results, and thus proceed back to the exploration phase for additional refinement of the data. That is, as is the case in CRISP-DM, SEMMA also driven by a highly iterative experimentation cycle (Olson, &Delen, 2008).

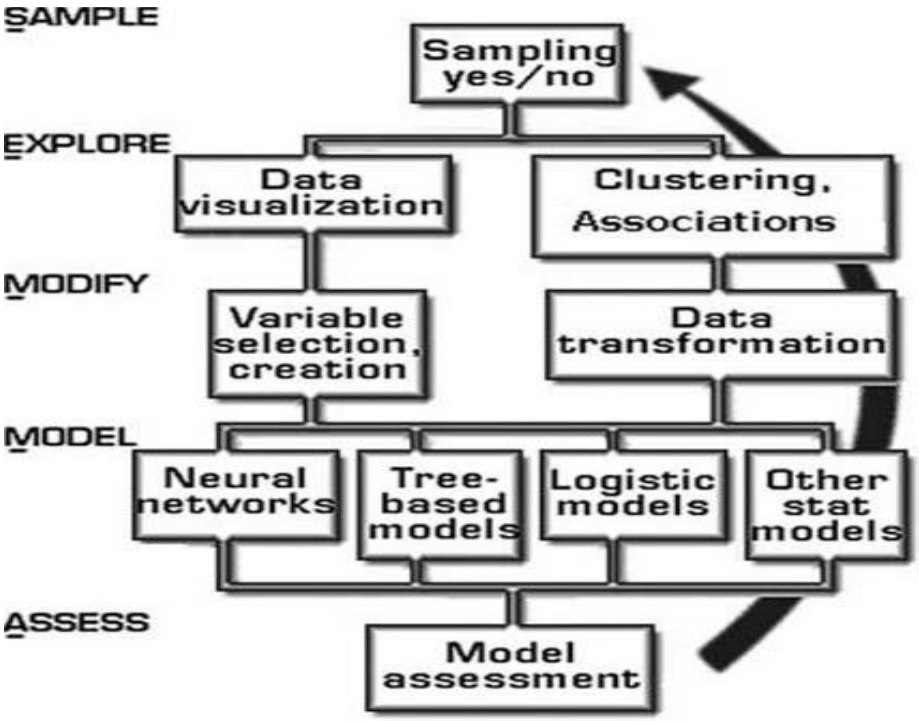



Figure 2.4 The Schematic of SEMMA

**Steps in SEMMA Process**

**Sample:** This is where a portion of a large data set big enough to contain the significant information yet small enough to manipulate quickly is extracted. A sampling strategy, which applies a reliable, statistically representative sample of the full detail data, is advocated for optimal cost and computational performance.

Mining a representative sample instead of the whole volume may drastically reduce the processing time required to get crucial business information. If a rare patternis so tiny that it is not represented in a sample and yet so important that it influences the big picture, it should be discovered using exploratory data description methods.

It is also advised to create partitioned data sets for better accuracy assessment.

- Training – used for model fitting.
- Validation – used for assessment and to prevent over fitting.
- Test – used to obtain an honest assessment of how well a model generalizes.

**Explore**: At this stage, unanticipated trends and anomalies are searched for in order to gain a better understanding of the data set. Visual or numerical exploration should be done for inherent trends or groupings. Exploration helps refine and redirect the discovery process. If visual exploration does not reveal clear trends, one can explore the data through statistical techniques including factor analysis, correspondence analysis, and clustering.

**Modify**: This is where the creation, selection, and transformation of the variables upon which to focus the model construction process happens. It may also be necessary to look for outliers and reduce the number of variables, to narrow them down to the most significant ones. One may also need to modify data because mining is a dynamic, iterative process.

**Model:** A variable combination that reliably predicts a desired outcome is searched for at this stage. Once data preparation is done, models that explain patterns in the data are constructed. Modeling techniques in data mining include artificial neural networks, decision trees, rough set analysis, support vector machines, logistic models, and other statistical models such as time series analysis, memory-based reasoning, and principal component analysis.

**Assess:** The usefulness and the reliability of findings from the data mining process are evaluated here. In this final step of the data mining process, an assessment of the models is done to estimate how well it performs. A common means of assessing a model is to apply it to a portion of data set put aside and not used during the model building during the sampling stage. If the model is valid, it should work for this reserved sample as well as for the sample used to construct the model. Similarly, models can be tested against known data. Practical applications of the model, such as partial mailings in a direct mail campaign, help prove its validity as well.

The SEMMA approach is completely compatible with the CRISP approach. Both aid the knowledge discovery process. Once models are obtained and tested, they can then be deployed to gain value with respect to business or research application.

The SEMMA process offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for its conception, creation and evolution, helping to present solutions to business problems as well as to find DM business goals (Azevedo, 2008).

#### *2.1.5.3. Hybrid Model*

According to Santos &Azevedo (2008) KDD, CRSP-DM, SEMMA and Hybrid model are commonly used for Data mining process. But in this study only Hybrid model have been used.

Hybrid models are model that combine aspects of both academic and industrial areas. As can be seen in figure 2.5 it is a six-step model. It was developed based on the CRISP-DM model by adopting it to academic research. The main differences and extensions include: Hybrid model provide more general, research-oriented description of the steps, it introduces a data mining step instead of the modeling step, in addition it introduces several new explicit feedback mechanisms, (the CRISP-DM model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms) and modification of the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in other domains.

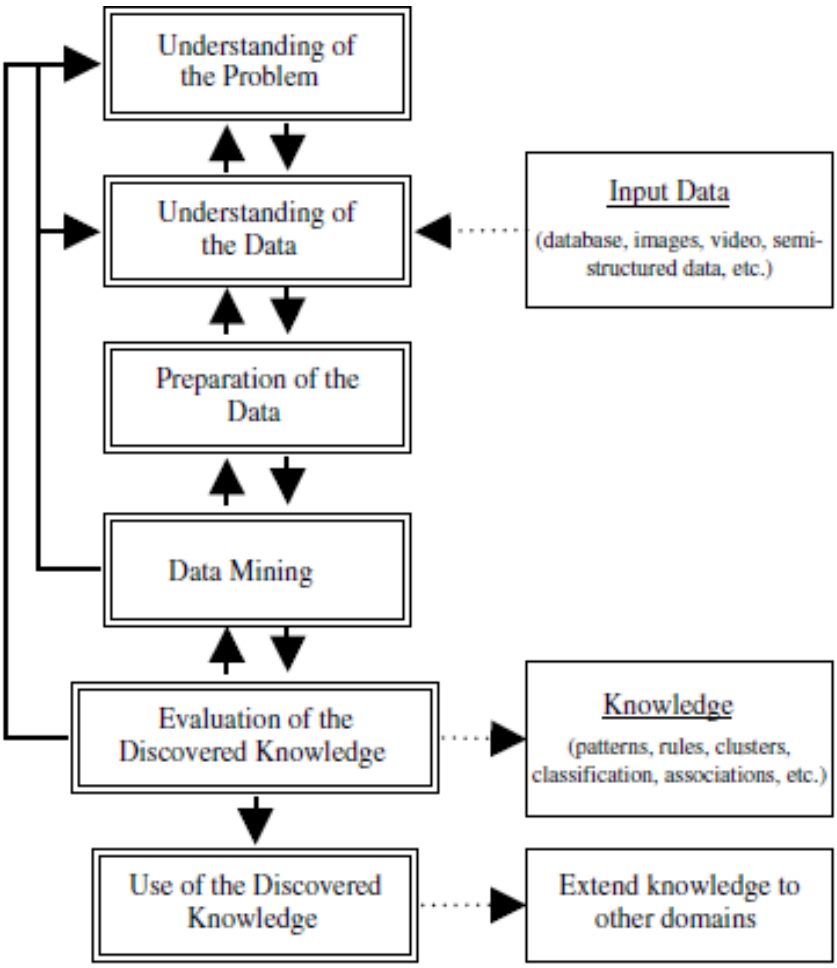



Figure 2.5 The Hybrid Models (description of the six steps follows)

**Understanding of the problem domain**

this helps to work closely with domain experts to define the problem and determine the project goals by selecting key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology. Finally, project goals are translated into DM goals, and the initial selection of DM tools to be used later in the process is performed.

**Understanding of the data: Understanding of the data**

This is used for collecting sample data and deciding which data is important. Data are checked for completeness, redundancy, missing values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.

**Preparation of the data**

this phase uses for preparing necessary data for subsequent operations. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values. The cleaned data are fed to next operations like reducing dimensionality, discretization and data granularization. The end results are data that meet the specific input requirements for the DM tools selected in first step.

**Data mining**

At this point the researcher used four DM algorithms to extract knowledge from preprocessed data.

**Evaluation of the discovered knowledge**

In this step the researcher in collaborating with domain experts of Ethio-telecom and INSA, understanding the results of the models, checking whether the discovered knowledge is new and interesting, and checking the contribution of the discovered knowledge.

## 2.2. Data Mining Models

According to Chang & Hsu (2005), the objective of data mining is identifying understandable correlations and patterns from existing data. In order to achieve the objective, the tasks of data mining can be modeled as either Predictive or Descriptive in nature (Dunham, 2003).

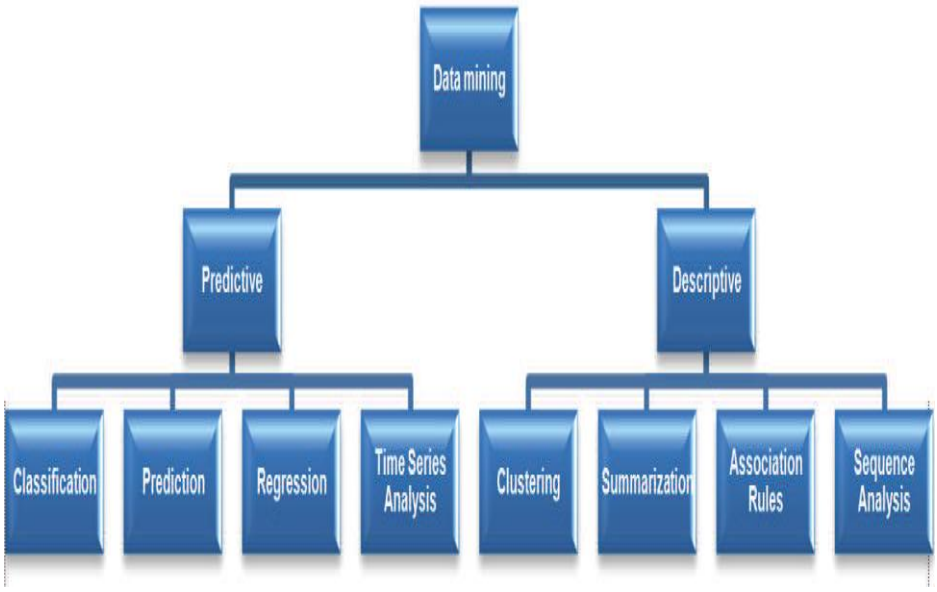



Figure 2.6 Data mining tasks and models

According to Williams (2011) Modeling is the first thing that comes to mind when someone thinks about data mining. Modeling is the process of taking some data and building implied description of the processes that might have generated it. The description is most of the time a computer program or mathematical formula. A model captures the knowledge exhibited by the data and encodes it in some language. Often the aim is to address a septic problem through modeling the world in some form and then use the model to develop a better understanding of the world.

Each model encompasses different techniques so as to achieve its task objective. According to Han (1996), different data mining techniques have been developed and used in data mining projects with the aim of achieving the data mining objectives. Some of these techniques are Association, Classification, and Clustering.

According to Han (1996) two or more of these techniques can be combined to form an appropriate process that meets the business needs. Some of the data mining models and techniques are examined as follows.

### 2.2.1. Predictive Model

A predictive model makes a prediction about values of data using known results found from different historical data. Prediction methods use existing variables to predict unknown or future values of other variables (Chan, & Hsu, 2005). As shown in figure 2.6. Predictive model includes classification, prediction, regression and time series analysis.

According to Chan, & Hsu (2005), Classification is the best understood of all data mining approaches among all Predictive models. Classification is commonly characterized as with classification tasks such as supervised learning, categorical dependant variable, and ability of assigning new data in to the set of well-defined classes.

#### *2.2.1.1. Classification*

Classification is one of the classic data mining techniques. Classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all records of employees who left the company; predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups (Han, 1996).

#### *2.2.1.2. Prediction*

According to Han (1996), prediction is one of the data mining techniques that discover relationship between independent variables and relationship between dependent and independent variables. The prediction analysis technique can be used in sales to predict profit for the future. If sale is considered as independent variable; profit could be a dependent variable. Then based on the historical sale and profit data, fitted regression curve can be drawn so that is can be used for profit prediction.

#### *2.2.1.3. Regression*

Statistical Regression is a supervised learning technique that involves analysis of the dependency of some attribute values upon the values of other attributes in the same item (Williams, 2011). This model predicts classes of new variables to which they belong.

The early research included investigations that separated people into deferent classes based on their characteristics. The regression came from modeling the heights of related people (Crano and Brewer, 2002).

#### *2.2.1.4. Time series analysis*

Time series analysis usually involves predicting numbering value for instance about market price of future. Time series are sequences of events. For example, the final sales income is an event that occurs each day of the week and each week in the month. Time series analysis can be used to identify the sales trends of organizations like Ethio telecom (Roiger, & Geatz, 2003).

### 2.2.2. Descriptive Modeling

According to Marco & Gianluca (2005), Descriptive data mining method can be defined as discovering interesting regularities in the data, to uncover patterns and find interesting subgroups in the bulk of data is normally used to generate frequency, cross tabulation and correlation.

Unlike the predictive model, a descriptive model serves as a way to explore the properties of calls examined, not to predict new fraud call properties. Descriptive task encompasses methods such as Clustering, Summarizations, Association Rules, and Sequence analysis.

Descriptive analysis is the task of providing a representation of the knowledge discovered without necessarily modeling a septic outcome. From a machine learning perspective, we might compare these algorithms to unsupervised learning (Williams, 2011).

#### *2.2.2.1. Clustering*

According to Han (1996), Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique. The clustering

technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes.

In Clustering, a set of data items is partitioned into a set of classes such that items with similar characteristics are grouped together. Clustering is best used for finding groups of items that are similar. For example, given a data set of customers, subgroups of customers that have a similar buying behavior can be identified (Roiger, & Geatz, 2003).

Clustering (which is also called Unsupervised learning) groups similar data together into clusters. It is used to find appropriate groupings of elements for a set of data. Unlike classification, clustering is a kind of undirected knowledge discovery or unsupervised learning; that is, there is no target field, and the relationship among the data is identified by bottom-up approach (Han, 1996).

#### *2.2.2.2.Association*

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction for example calling number and peak status. That's the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together. Telecom can also use association technique between calling date and time with zone to identify its customer's calling habits. (Han, 1996).

#### *2.2.2.3. Summarization*

Summarization maps data into subsets with associated simple descriptions (Dunham, 2003). Basic statistics such as Mean, Standard Deviation, Variance, Mode and Median can be used as Summarization approach.

#### *2.2.2.4. Sequential analysis*

Sequential patterns analysis is one of the data mining techniques that looks for discovering or identifying similar patterns, regular events or trends in transaction data over a business period. In financial transaction, with historical transaction data, businesses can identify a set of durations or dates that customers prefer to do their calls. Then businesses can use this information to recommend customers appropriate time for making calls so as to get better network traffic. (Han, 1996).

Sequence Analysis is used to determine sequential patterns in data. The patterns in the dataset are based on time sequence of actions, and they are similar to association data, however the relationship is based on time. In Market analysis, the items are to be purchased at the same time, on the other hand, for Sequence Analysis the items are purchased over time in some order (Dunham, 2003).

## 2.3. Classification Techniques

### 2.3.1. J48 Decision trees

Decision tree is one of the most used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it. The algorithms that are used for constructing decision trees usually work top-down by choosing a variable at each step that is the next best variable to use in splitting the set of items (Rokach and Maimon, 2005).

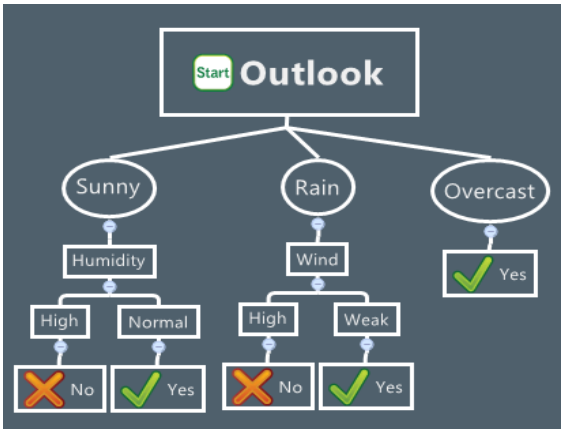



Figure 2.7 Decision tree

A number of different algorithms may be used for building decision trees including CHAID (Chi-squared Automatic Interaction Detection), CART (Classification and Regression Trees), C4.5, J48 and Random forest (Two crows corporation, 1999).

## J48 decision tree algorithm

Decision tree models are constructed in a top-down recursive divide-and-conquer manner. J48 decision tree algorithms have adopted this approach. The training set is recursively partitioned into smaller subsets as the tree is being built (Han, & Kamber, 2006).

According to Bharti et al. (2010) J48 decision tree algorithm is a predictive machine learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data.

Having the capability of generating simple rules and removing irrelevant attributes, the J48 decision tree can serve as a model for classification. (Witten and Frank 2000).

According to (Krishnaveni & Hemalatha, 2011) J48 Algorithm performs the following sequence of steps to accomplish its classification task.

- It check for base cases
- It finds normalized information gain
- It selects the attribute with the highest normalized information gain
- It Creates decision nodes

- It recourses on the sub lists obtained by splitting and add those nodes as children of node

J48 algorithm has various advantages. Some of these advantages are:-

- Gains of balanced flexibility and accuracy
- Capability of limiting number of possible decision points
- Higher accuracy

### 2.3.2. Random forest (RF)

Random forests (RF) are a combination of tree predictors that each tree depends on the values of a randomly selected vector samples and it distributes equal values of vector samples to all of the trees in the forest. The strength of individual trees in the forest and the correlation between them determines the generalized error of a forest and its tree (Breiman, 2001).

As primary form knowledge representation, the random forest algorithm is presented in terms of decision trees. Though, the random forest algorithm is thought of as a meta-algorithm. Decision tree algorism or any one of the other model building algorisms could be the actual model builder. (Breiman, 2001).

According to (Krishnaveni & Hemalatha, 2011) Random forest Algorithm performs the following sequence of steps to accomplish its classification.

- Choose T number of trees to grow
- Choose m number of variables used to split each node. m<<M,
    - where M is the number of input variables,
    - m hold constant while growing the forest
- Grow T trees. When growing each tree do
- Construct a bootstrap sample of size n sampled from Sn with the replacement and grow a tree from this bootstrap sample
- When growing a tree at each node select m variables at random and use them to find the best split
- Grow the tree to a maximal extent and there is no pruning

- To classify point X collect votes from every tree in the forest and then use majority voting to decide on the class label

Random forest algorithm has various advantages. Some of these advantages are:-

- Unexcelled accuracy and ability of running on large data bases.
- Capability of handling thousands of input variables without variable deletion and fast learning.
- Having an effective method of estimating missed data and highest maintenance accuracy.
- Ability of saving generated forests for future use.

### 2.1.1. PART (Partial Decision Trees) classifier

Frank and Witten (1998) defined PART as a rule based algorithm that produces a set of if-then rules that can be used to classify data. PART is a modification of C4.5 and RIPPER algorithms and draws strategies from both. PART adopts the divide-and-conquer strategy of RIPPER and combines it with the decision tree approach of C4.5.

PART generates a set of rules based on the divide-and conquers strategy, and then it removes all instances from the training collection that are covered by this rule, finally it proceeds recursively until no instance remains (Krishnaveni & Hemalatha, 2011).

PART builds a partial decision tree for the current set of instances and chooses the leaf with the largest coverage as a new rule. Though unlike that of C4.5 the trees built by PART for each rules are partial and incomplete, PART is advantageous because of its simplicity and its capability of generating sufficiently strong rules (Frank and Witten, 1998).

### 2.1.2. Artificial Neural Networks (ANNs)

According to Gaur (2012), neural networks represent a brain symbol for information processing. These models are not an exact replica of how the brain actually functions but they are biologically inspired. Due to their ability to "learn" from the data, flexible assumptions and ability of generalization, neural networks are found to be promising systems in different forecasting and business classification applications.

Artificial neural network (ANN) is a resulting model of neural computing. Neural computing is a key component of any data mining tool kit and it refers to pattern recognition methodology for machine learning. Pattern recognition, forecasting, prediction, and classification are some of the business applications where neural networks have been used. (Gaur, 2012).

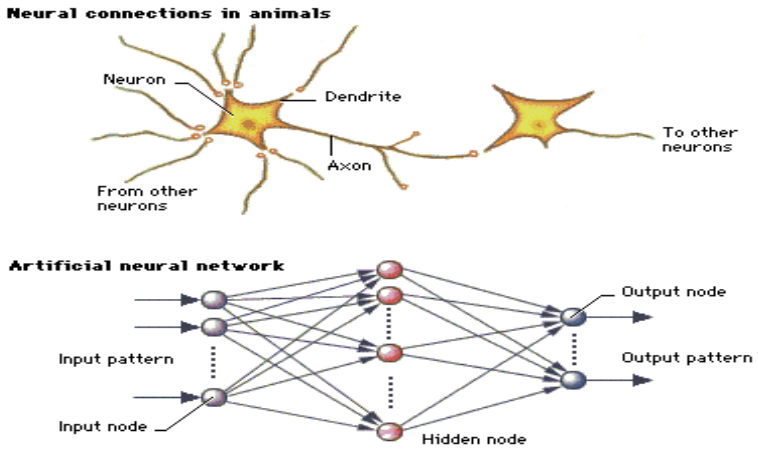



Figure 2.8 Natural and artificial neural connections

**Neural Network Topologies:**

There are two types of Neural networks, Feed forward neural network and The Multilayer perceptron. According to Singh & Chauhan, (2009), the feed forward neural network is the first and arguably simplest type of artificial neural network where information moves in only one direction from the input nodes to the hidden nodes and to the output nodes. On the other hand, according to Han and Kamber (2006), the Multilayer perceptron Building on the algorithm of the simple Perceptron, the MLP model gives a perceptron structure for representing more than two classes. It also defines a learning rule for this kind of network. Finally the MLP is divided into input hidden and output layers, where each layer in this order gives an input to the next. Figure 2.9 depicts graphical representation of both layers.

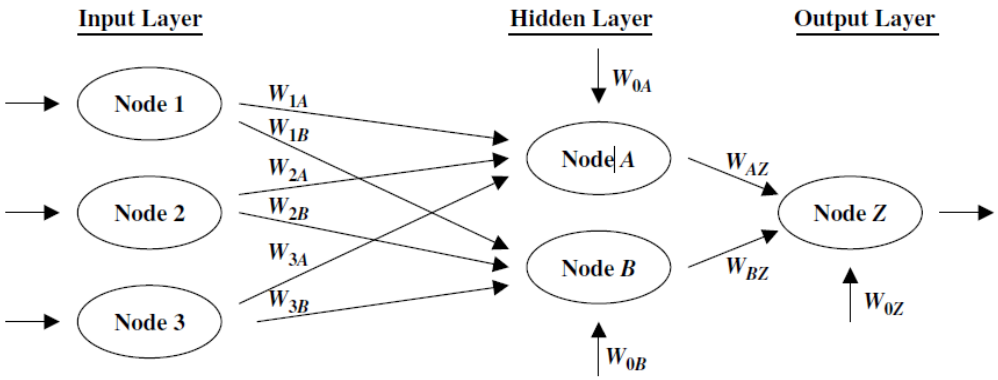



Figure 2.9 The Multilayer perceptron and feed forward neural network

As Han and Kamber (2006) stated, training is defined as the process of iterating through the training set to adjust the weights. The connection weights are estimated by the neural network, so that the output of the neural network accurately predicts the test value for a set of values fo a given input. Various aspects of training such as adjusting the speed of conversion are controlled by set of parameters found in each training method and Back propagation is the most common type of training methods.

A neural net has high tolerance to noisy data as well as ability to classify patterns on which they have not been trained. Moreover, several algorithms have recently been developed for the extraction of rules from trained neural nets. The conditions contribute towards the usefulness of the neural networks for classification in data mining (Han, & Kamber, 2006).

**The learning process of an ANN**

According to Han & Kamber (2006), in supervised learning, the learning process is inductive; that is, connection weights are derived from existing cases. Supervised learning follows the following procedure.

Compute temporary outputs => Compare outputs with desired targets => Adjust the weights and repeat the process.

Figure 2.10 depicts tasks that are involved in the process of supervised learning.

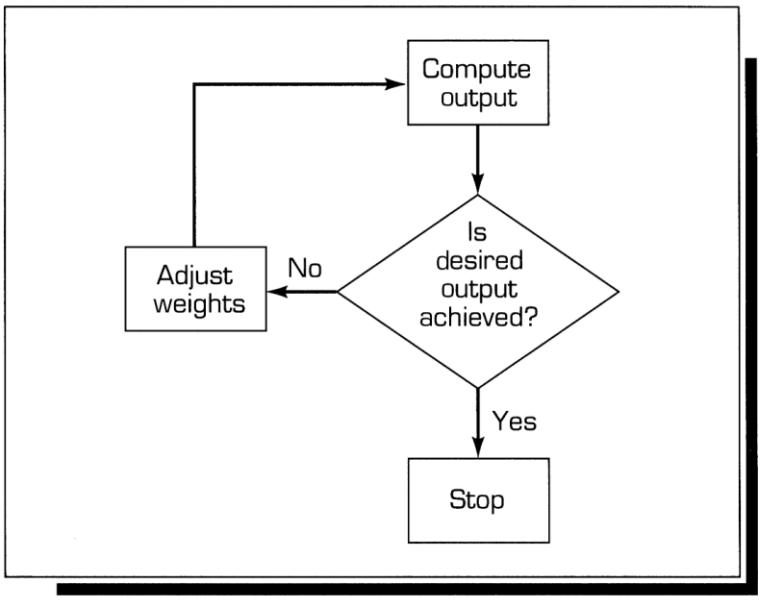



Figure 2.10 The learning process of an ANN

According to Gaur (2012), these days, neural network is very appropriate to solve the problems of data mining. Good robustness, self-organizing, adaptive, parallel processing, distributed storage and high degree of fault tolerance are some of the characteristics that makes neural network appropriate.

According to Singh & Chauhan (2009), using neural is advantageous because of its high accuracy, noise tolerance, independence from prior assumptions, ease of maintenance, and implementation capability parallel hardware

# CHAPTER THERE

# 3. TELECOMMUNICATION FRAUD

## 3.1. Introduction

In response to human communication needs, telecommunications have developed into cellular radio networks, which enable subscribers to connect and communicate regardless of their location and even if the objects in movement. But telecommunication not free from probes. As a result the telecom industries are challenged by different types and depths of fraud situations. In this of work the researcher initiated to help the Ethio-telecom company in detecting and predicating the fraudsters and to minimize revenue loss.

The Communications Fraud Control Association (CFCA) estimated that annual global fraud losses in the telecoms sector were between $54 billion and $60 billion, an increase of 52 percent from previous years and the CFCA also estimated that global annual losses to fraud account for 5 percent of the total telecom sector revenue with mobile operators seen as more vulnerable than fixed line (CFCA, 2009).

According to the CFCA (2006).47.3% of global fraud losses are from Subscription/ Identity Document (ID) Theft and Private Branch Exchange PBX/Voicemail.

The reasons for doing the fraud is to dynamic incensement distributed of new technology over the world and faced the challenges of operational weaknesses, fraudsters need money and political and ideological factors (Brown, 2005).

## 3.2. Common Types of Fraud

### 3.2.1. Subscription Fraud

The subscription fraud is the most common since with a stolen or manufactured identity, there is no need for a fraudster to undertake a digital network's encryption or authentication systems. Their low technique with less chance of detection and it is one of the fraudster's preferred methods (Brown, 2005).

According to GSM Association and the Communications Fraud Control Association, subscription fraud is the starting point for many other telecoms fraud and as such is recognized as the most damaging of non-technical fraud types (Ghosh, 2010).

Subscriber fraud involves the fraudulent individual obtaining the customer information required for signing up for telecommunication service with authorization. The usage of the service creates a payment obligation for the customer (Brown, 2005).

### 3.2.2. Premium Rate Fraud

Premium rate fraud mostly used in combination with other types of fraud, the main reasons are to make free calls to high cost numbers like competition or hot lines, and also to make money from falsely generating calls to a number owned and operated by the fraudster. The more calls generated, the higher the profit to be made (Van Heerden, 2005).

According to Van Heerden (2005) the premium rate services fraud currently considered the most financially damaging fraud type in combination with Roaming. This fraud type is international so boundaries have no relevance in this case. In well organized attacks the calls are made during weekends, holidays, etc in order to take advantage of using a bigger time, until the first Roaming High usage report (HUR) is received.

### 3.2.3. Internal Fraud

The internal fraud can be performing by an operator employee and can be carried out under different forms, for instance, applying services directly onto the switch without make changes on the billing or removing records from billing systems, creating false accounts/customers /employees, r emoving call detail records (CDRs) from the billing cycle, or just manipulating the accounting and credit processes. All these factors can mean lost or incorrect billing records (Shawe-Taylor, Howker, & Burge, 1999).

According to Brown (2005) internal fraud represents 8.2% of incidences but generates 40.3% of value lost which is equal in value of the following four types of fraud combined: roaming (11.4%), pre-paid (10.8%), subscription (11.6%) and premium (13.2%). The motivation for such a fraud is caused by company's weak management accounting practices, impractical performance targets, few checks and balances, and unhappy employees.

According to Brown (2005) follows are from the internal fraud types.

Theft of data/equipment.

Network attacks abuse.

Employee placement.

Misuse of computer systems.

Sale of sensitive information:

Customer related information.

Products, new services and equipment.

Sales figures.

### 3.2.4. Prepaid Fraud

Prepaid fraud is an attempts made by the fraudsters to use free telecommunication services by using lost or stolen credit cards. Internal engineers can access and adjust the billing-activation

systems, stealing cards, and PIN numbers and recharge codes at production and support sites, scan for data from valid phones and duplicate the information onto stolen devices. The prepaid phones appear valid but steal minutes from an honest customer (Brown, 2005).

Credit information is stored in the SIM; it is not very difficult to modify the same illegally. In order to minimize the impact of prepaid fraud, the vendor community is developing techniques such as Real time billing and rating systems, combining pre-paid and post paid systems into one, generation of logs for changes made to IN based systems; activation at point of sale systems; migration to IN (Intelligent Network) based systems(Brown, 2005).

### 3.2.5. Cloning fraud

Cloning fraud is created by reprogram techniques that mean copying the identity of valid mobile telephone to another mobile telephone. During cloning it requires access to electronic serial number and mobile identification number.

A cloned mobile phone is a form of reprogram to access to a mobile network as a valid cell phone. The legal phone user then gets billed for the cloned phone's calls. Cloning mobile phones is achieved by cloning the SIM card contained within, not necessarily any of the phone's internal data. There are various methods used to obtain the ESN and MIN; the most common are to crack the cellular company, or listen in on the cellular network (Brooks, et al., 2012).

Whenever a cellular phone is on, it periodically transmits two unique identification numbers: its Mobile Identification Number (MIN) and its Electronic Serial Number (ESN). These two numbers together specify the customer's account. These numbers are transmit unencrypted, and they can be received, decoded and stored using special equipment. Cloning occurs when a customer's MIN and ESN are programmed into a cellular telephone not belonging to the customer. When this second telephone is used, the network sees the customer's MIN and ESN and subsequently bills the usage to the customer. With the stolen MIN and ESN, a cloned phone user that is the fraudsters can make virtually unlimited calls, whose charges are billed to the customer. The attraction of free and untraceable communication makes cloned phones very popular in major metropolitan areas (Fawcett, & Provost, 1997).

### 3.2.6. Roaming Fraud

Roaming Fraud is the ability to use of telecom products or services, such as voice or data services, outside the home network with no intention to pay for it. In these cases fraudsters use the longer time-frames required for the home network. Roaming fraud can starts as an internal or subscription fraud in the home network, when obtained SIM cards are sent to a foreign network the (Brooks, at el., 2012).

The fraudster's are used different fraud methods. Subscription fraud is one of the fraudster's preferred methods for digital roaming fraud. The delay in the home provider receiving roamer call data can be anywhere from one to several days, Stealing mobile phones belonging to roamers, usually in vacation destinations (Brown, 2005).

### 3.2.7. SIM Box (Bypass) Fraud

A SIM Box is device that maps the call from voice over internet protocol (VoIP) to a SIM card of the same mobile operator of the destination mobile. So that international call terminating as home call to subscriber country and usually cheap compared to the cost of terminating the international call. This is to just bypass international traffic. Commonly, SIM boxes are used to commit bypass fraud. SIM Boxes, also known as illegal mobile communication gateways, generate significant interconnect profits losses for mobile operators by illegally bypassing official interconnection. Sim Boxes also have a negative impact on call quality and create network congestion (Cohen & Southwood, 2004).

According to Cohen & South wood (2004) the purpose of bypassing is making money by illegally terminating traffic into operator's network, without paying the interconnection fee, using VoIP technology. That is usually international traffic. They have contract with a wholesale operator for a determined number of minutes to be terminated via his SIM cards. Traffic is being received via IP and is routed through the fraudster's, SIM gateways (SIM boxes), it reaches the destination as a national call. The fraudster will pay the network for a national call but will charge the Wholesale operator for every minute he terminated; the Network Operator loses the Interconnection fee.

### 3.2.8. PBX Fraud

A PBX (Private Branch Exchange) is a switch station for telephone systems. It consists mainly of several branches of telephone systems and it switches connections to and from them, thereby linking phone lines. Most medium-sized and larger companies use a PBX for connecting all their internal phones to an external line. This way, they can rent only one line and have many people using it, with each one having a phone at the desk with different number. There are different types of PBX fraud, all the call pay the companies for including not made for the business of the company. Some of these frauds are carried out by workers of the company and others hacking into the company network (Shawe-Taylor, Howker, & Burge, 1999).

## 3.3. Related Works

Jember (2005) conducted a research to explore the potential application of data mining in supporting fraud detection on mobile communication service in the case of Ethio - mobile. According to the researcher, the application of data mining methods and tools to 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.

The data used by the researcher was a three month data from October and November of 2003 and March of 2004. The researcher used three month of data from October and November of 2003 and March of 2004. From the total of 9153 customers, 900 customers who made an international call of more than one minute per call were selected using stratified sampling technique in order to get representative data from all the postpaid mobile subscribers. The methodology used had three basic steps, data collection, data preparation, model building and testing. Matlab tool and neural network data mining technology were employed to build the models from which an accuracy of 89% was achieved.

The research selected the attributes Minimum number of calls, Maximum number of calls, Average number of calls, Standard deviation of number of calls, Total number of calls,

Minimum duration, Minimum duration, Average duration, Standard deviation of duration and Total duration.

The research scope was limited to exploring the possibility of application of the DM technology to supply fraud detection in mobile communication network using artificial neural network of the postpaid mobile phone. The researcher recommended giving sufficient time to get representative samples of the data and sufficient time to building an appropriate model to get the best performance (Jember, 2005).

Another study was done by Geremeskel (2006). In this research work, an attempt had been made to assess the possible application of data mining technology to support mobile fraud detection on Ethio-Mobile Services. The data source of this study was taken from call detail record (CDR). The researcher used a two month data from the month of February and March of 2005.

The methodology by Geremeskel (2006) included data collection, data preparation, and model building and testing. SPSS software tool and artificial neural network algorithm were used as data mining techniques. The researcher took 29,463 records as a sample and selected 9 attributes for that process. Fromthese 29,463 call records, 9186 were identified as fraudulent calls. In general, the numbers of fraudulent call became 31.18 percent of the total sample size (total record). The following attributeswere selected during his experimentation. CallType,ChargeType,Roamflag, CallingPartyNumber, CalledPartyNumber, CallBeginTime, Call Duration, CallCost and CalledArealndicator

Additionally the study was targeted at prepaid mobile service. The number of post-paid mobile subscribers was 53614 and the prepaidones were 381417. The researcher further recommended that additional research could be done by including other attributes of the call detail record so as to build models with better performance and better accuracy (Geremeskel, 2006).

Melaku (2009) also conducted a research to assess the applicability of DM techniques to CRM as a case study on Ethiopian Telecommunications Corporation (ETC). The objective of his research was to support CRM at ETC by employing appropriate DM techniques such as clustering and

classificationon customer's database and in order to achieve the research objectives he adopted the CRISP-DM model.

First the different classification model was done with J48 decision tree algorithm with the 10-fold cross-validation, and splitting the dataset in to 80% training and 20% testing, techniques by setting the cluster index formed by the cluster model as dependent variable and the rest as independent variable. Among these models a model that showed classification accuracy of 98.97% is selected. Similarly different classification model of Multilayer perceptron ANN algorithm are carried out by changing its hidden layer number of node and learning rate parameters value. A model with classification accuracy of 98.62% is chosen. Finally a comparison of decision tree and ANN model was done, in terms of the overall classification accuracy. Hence the decision tree model has excelled in the evaluation parameters and therefore selected as best classifier for CRM application (Melaku, 2009).

Fekadu (2004) did a similar research as melaku (2009), the application of data mining techniques to support customer relationship management at Ethiopian telecommunications corporation (ETC) on the post-paid mobile. He used the k-means clustering algorithm for clustering and decision tree algorithms for classification and achieved a 94.78% accuracy of prediction.

Abidogun (2005) conducted a research on data mining, fraud detection and mobile telecommunications. One of the strategies for fraud detection is to check for signs of questionable changes in user behavior. Although the intentions of the mobile phone users cannot be observed, their intentions are reflected in the call data which define usage patterns.

Over a period of time, an individual phone generates a large pattern of use. While call data are recorded for subscribers for billing purposes, we are making no prior assumptions about the data indicative of fraudulent call patterns, i.e. the calls made for billing purpose are unlabeled. Further analysis is thus, required to be able to isolate fraudulent usage. An unsupervised learning algorithm can analyses and cluster call patterns for each subscriber in order to facilitate the fraud detection process.

This research investigates the unsupervised learning potentials of two neural networks for the profiling of calls made by users over a period of time in a mobile telecommunication network. The study provides a comparative analysis and application of Self-Organizing Maps (SOM) and Long Short-Term Memory (LSTM) recurrent neural networks algorithms to user call data records in order to conduct a descriptive data mining on users call patterns.

The investigation shows the learning ability of both techniques to discriminate user call patterns; the LSTM recurrent neural network algorithm providing a better discrimination than the SOM algorithm in terms of long time series modeling. LSTM discriminates different types of temporal sequences and groups them according to a variety of features. The ordered features can later be interpreted and labeled according to specific requirements of the mobile service provider. Thus, suspicious call behaviors are isolated within the mobile telecommunication network and can be used to identify fraudulent call patterns. We give results using masked call data from a real mobile telecommunication network.

The method of research proceeds with the normalization of call data records which contained a 6-month call data set of 500 masked subscribers from a real mobile telecommunication network. The researcher extracted from the data set, Mobile Originating Calls (MOC). These are calls that were initiated by the subscribers. Within the 6 months period, a total of 227,318 calls originated from the 500 subscribers. The SOM and LSTM RNN models are then applied to unsupervised discrimination of the normalized call data set. Results are reported which estimates the performances of the two learning models.

Ojuka (2005) also conducted a research in the area of fraud detection in telecommunication. According to the researcher, detecting fraud is a challenging task and is a continuously evolving discipline. Whenever it becomes known that one detection method is in place, the fraudsters will change their tactics and try others. For example, on an industrial scale, telecommunication fraud is mainly perpetrated by organized criminal gangs, professional hackers and service providers own employees

Subscription fraud and bad debts are the major causes of loss of revenue in the telecommunication industry. This research project therefore, focused on designing a subscription fraud detection system with minimum false positive alerts using decision tree learning. The system has been trained to learn from training data and used decision tree algorithms to make classifications/predictions on future telecommunication data. Weka software was used to induce the resulting decision tree from the training data. Call Detail Record (CDR) is descriptive information about a call placed on a telecommunication network. It includes sufficient information to describe the important characteristics of each call such OPC, DPC, duration of the call, call start and end time.

This was used as training data to train the model to produce a function which mapped the variables to the classes. This model can thus be used to predict classes of future unseen data which the system/model has not been trained on, i.e. having no classes.

The variables/attributes from the past records collected were as below; Phone number, Customer segment, Balance, Pay duration, Credit limit, Ratio and Classification The dataset was then classified by choosing the J48 algorithm which is a decision tree. In weka software, C4.5 is a program for machine learning and the test option chosen was 10-fold cross-validation. From the selected attributes and 999 instances from the training data, 992 instances were correctly classified representing 99.2993% of total records and 7 instances were incorrectly classified representing 0.7007% of total records.

The researcher recommended that the future work should be extension and modification to this method of detecting subscription fraud and bad debts using decision tree learning to achieve better performance and a better degree of automation, like increased speed with a large amount of data (Ojuka, 2009).

Hilas (2012) was a 2nd Pan-Hellenic Conference on Electronics and Telecommunications - PACET'12, March 16-18, 2012, Thessaloniki, Greece. Fraud is increasing dramatically each year resulting in loss of a large amount of Euros worldwide. An invaluable tool for the detection of fraud is the modeling of telecom users' behavior. In this paper, a short description of ongoing

research is given with focus on how available data mining techniques are applied to test user profiles with the task to detect superimposed fraud cases in the telecommunications networks of large organizations

Fraud detection is important to the telecommunications industry because companies and suppliers of telecommunications services lose a significant proportion of their revenue as a result. Moreover, the modeling and characterization of users' behavior in telecommunications can be used to improve network security, improve services, provide personalized applications, and optimize the operation of electronic equipment and/or communication protocols.

The main idea behind user profiling is that the past behavior of a user can be accumulated in order to construct a profile, or a "user dictionary", or a "user signature» of what might be the expected values of a user's behavior.

This profile is a vector that contains single numerical summaries of some aspect of behavior or some kind of multivariate behavioral pattern. Future behavior of the user can then be compared with his profile in order to examine the consistency with it (normal behavior) or any deviation from his profile, which may imply fraudulent activity.

The data that can be used to monitor the usage of a telecommunications network are contained in the Call Detail Record (CDR) of any Private Branch Exchange (PBX). The CDR contains data such as: the caller ID, the chargeable duration of the call, the called party ID, the date and the time of the call. In order to test the ability of each profile to discriminate between legitimate usage and fraud, neural networks were used as classifiers.

Neural network classifiers performed very well on the problem, giving measures of TP rates better that 80% with FP less than 2%. However, from the data analysis point of view, neural networks work like black boxes so they do not reveal the nature of the discriminating characteristics. Hence other approaches are also used to show different aspects of the same problem. From the decision trees based approach it is concluded that with the daily representation of the user behavior only the 65% of the cases is classified correctly. This is in

contrast with the weekly user representation where 85% of the cases were classified correctly (Hilas, 2012).

| Researcher | DM techniques | Data Types | No of attribute | Total samples | Tools | Accuracy |
|---|---|---|---|---|---|---|
| Jember, (2005) | Neural networks | CDR | 10 | 900 | Matlab | 89% |
| Gebremesk(2006) | Neural networks | CDR | 9 | 29,463 | SPSS | 94% |
| Abidogun, (2005) | Neural networks | CDR | 6 | 227,318 | SPSS | 90% |
| Ojuka (2009) | J48 | CDR | 7 | 999 | Weka | 99% |
| (Hilas (2012) | Neural networks | CDR | 6 | | weka | 85% |

Table 3-1 Summary of related work

Both Jember and Gebremesk have tried to assess the possible application of data mining technology to support mobile fraud detection on Ethio-Mobile Services. Matlab and SPSS tools were used for model building, and both have been used only artificial neural network algorithm as data mining technique. The sample they took was two and three month.The performance both the local researchers scored accuracy of 89% and 94%. Since 2005 and 2006 the number of mobile subscribers about half million (Geremeskel, 2006), however, today mobile subscribers in Ethiopia reached more than 18 million (Ethiopian ICT, 2012).

In this study the researcher initiated with aim of develop predicting model for telecommunication frauds using data mining techniques in Ethiopia mobile communication services provision. The researcher of this study took six month of data to develop the models. In addition the researcher of this study, analysis was performed using WEKA environment for the research process. Furthermore, the researcher has been used four different classification algorithms for this study. Namely, J48, PART, random forest and Multilayer perceptron ANN have been used for experimentation of this study and the researcher compared the four algorithms and identified the best one. Finally the researcher achieved better performance from the global and local researches.

# CHAPTER FOUR

# 4. BUSINESS AND DATA UNDERSTANDING

## 4.1. Understanding of the problem domain

Telecommunication Company worldwide suffers from customers who use the provided services without paying. The estimated losses amount to several billions of dollars in uncollectible debt per day. Even though this is a small percentage comparing to the telecom operators revenue, it is still a significant loss. The mobile telecommunication industry stores and generates tremendous amounts of raw and heterogeneous data that provides rich fields for analysis.

To understand the telecommunication, the researcher observed Ethio-telecom head office and some area of Ethio telecom branches. As initial step the researcher involves working closely with domain experts to define the problem and determine the research goals. Domain experts were consulted to have brief understanding on the problem area. In order to have detail understanding and knowledge about the problem domain, the researcher discussed based on the discussion points shown in appendix 10 concerning the telecom fraud.

According to the domain experts of Ethio telecom, the Ethiopian Government has decided to transform the telecommunication infrastructure and services to world class standard. Thus, Ethio telecom is born from this ambition in order to bring about a paradigm shift in the development of the telecom sector to support the steady growth of our country.

The vision the company is to be a world-class telecommunications service provider. So that the following points are the mission of the company

- Connect every Ethiopian through Information Communication Technology.
- Provide telecommunication services and products that enhance the development of our Nation.
- Build reputable brand known for its customers' consideration.

- Build its managerial capability that enables Ethio telecom to operate an international standard.

In line with its ambitious mission, Ethio Telecom has ambitious goals:

- being a customer centric company
- offering the best quality of services
- meeting world-class standards
- building a financially sound company

Generally the researcher identifies key people at this domain who are from Ethio telecom and also expertise for information network security agency (INSA) and focuses on understanding the research objective and requirements from a business perspective, then converting this knowledge into a DM problem and a preliminary plan designed to achieve the objective. And then learns about current solutions to the problem domain.

## 4.2. Understanding of the data

Next to identifying the problem and building a simple plan for solving the problem, the researcher proceeded with the central item in data mining process which is data understanding. This includes listing out attributes with their respective values and evaluation of their importance for this research and careful analysis of the data and its structure is done together with domain experts by evaluating the relationships of the data with the problem at hand and the particular DM tasks to be performed. Finally, the researcher verifies the usefulness of the data with respect to the DM goals.

### 4.2.1. Data collection

The source of data for this research has been collected from Ethio-telecom database six month records of September 2012-February 2013. The database was manipulated by using oracle software system. The first and the major source of data was the researcher identified the Call detail records.

When a call is to be found on a telecommunications network, descriptive information about the call is saved as a call detail record. Call detail records include sufficient information to describe

the important characteristics of each call. The Call Detail Record has descriptive information about each call is saved as a call detail record.

This means that the CDR contains each call information related to telephone call, such as billing number, Time of call initiation, duration of call in second, mobile number initiating call, mobile number receiving the call, online charge system, recharge ID number, call type (local or international), the amount charged or to be charged for the call duration, subscriber ID number used for billing and subscriber line status are some of the details of the call. In the application of data mining technology and in developing a model that can support fraud detection, the goal of this research is to developing model so as discover the presence of illegitimate calling activity of telecommunications customers.

The illegitimate calling activity may not be able to be observed directly, but they are reflected in the calling behavior. The calling behavior is collectively described by the call detail record, which in turn can be observed. Therefore, it is reasonable to use call detail record to apply data mining technology and to formulate model using training and testing dataset and evaluate the accuracy of the model using the weka software.

Originally the CDR data were around 1.8 terabytes records in six month of September 2012, November 2012, October 2012, December 2012, January 2013 and February 2013. An average of each month is about 300 GB.

The researcher have been used the purposive sampling techniques. The fires reason why the researcher selected the purposive sampling technique was the size of the CDR data were very huge as indicated above. The second reasons were the behavior of the telecom subscription fraudsters. The researcher in collaborating with domain exports identified one of the major behaviors of subscription fraudsters are using high duration (talk long time) and disappeared from the network.

So that the researcher selected the high duration users that means if the user monthly call duration greater than 200000 second that is about 556 hours each month, picking for the first

sample. The data is still very huge and then needs other sampling techniques and the researcher selected similar sampling technique. Then the researcher selected high duration users that means if the user call duration greater than 2000 second at one call. That is about 34 minutes for each a single call. Finally the researcher was used purposive sampling technique. Therefore, in ordered to manage and use the huge CDR data, the researcher have been selected the purposive sampling techniques. To manage the huge CDR data was one of the challenged times of the researcher.

### 4.2.2. Description of the collected data

Description of the data is very important in data mining process in order to clear understand the data. Without such an understanding, useful application cannot be developed. As indicated before the research is collected the data form Ethio-telecom pre paid mobile CDR in six month records. That is very huge record and difficult to manage. In this section, from the source described above, the attributes with their data types and descriptions are shown in the following table 4.1. The tables bellow show that the initially and derived attributes of pre paid mobile subscriber.

| No | Attribute Name | Data Type | Description | Attribute remark |
|---|---|---|---|---|
| 1 | BILLING_NBR | String | Billing number (Mobile number to be charged) | Original |
| 2 | START_TIME | Date | Time of call initiation (calling time) or The time when the call is begins | Original |
| 3 | DURATION | Numeric | Total calling time or the duration of call in second(Call duration in seconds) | Original |
| 4 | CHARGE | Numeric | Amount of birr paid ( to be paid for the case of post-paid mobiles) | Original |
| 5 | CALLING_NBR | String | Mobile number initiating or originating the call | Original |
| 6 | CALLED_NBR | String | Mobile number receiving the call | Original |
| 7 | CELL_A | String | Mobile BTS cell sector A number(BTS-ID) where the call is originated | Original |
| 8 | FILE_ID | String | File number used for cross checking in case bill complain. This file ID is given by the department sending the CDR file. | Original |
| 9 | RECORD_SEQ | Numeric | Record sequence number give when the call is made | Original |
| 10 | FEE | Numeric | Amount paid or to be paid | Original |
| 11 | OCS_RE_ID | Numeric | Online Charge System recharge ID number | Original |
| 12 | Calling & called number | Nominal | Cheek the validity of the calling and called number | Derived |
| 13 | Charge & duration status | Nominal | The balance between the duration of call and the amount birr paid if it is normal situation | Derived |
| 14 | payment status | Nominal | To cheek whether or not pay per call in normal situation. | Derived |
| 15 | Call Cost | Numeric | Calculate current cost per call | Derived |
| 16 | Calling Time | Nominal | To cheek whether or the calling time is on peak or off-peak hours | Derived |
| 17 | Calling Day | Nominal | To cheek the off-peak hours on Sunday or night | Derived |
| 18 | Call Type | Nominal | To identify the call is local or international | Derived |
| 19 | calling zone | Nominal | Identify the calling is to/from witch zone. According to Ethio-telecom there are three different tariff zones. Zone 1 (Asia, Europe, Middle East and North America) and Zone 2 (Africa, Oceania and South America) or local call. | Derived |

Table 4-1Attribute with their description of the pre paid mobile

## 4.3. Preparation of the data

Preprocessing of the data in preparation for classification and prediction can involve data cleaning to reduce noise or handle missing values, relevance analysis to remove irrelevant or redundant attributes, and data transformation, such as generalizing the data to higher-level concepts or normalizing the data (Witten, & Frank, 2005).

The purpose of data preprocessing is to clean selected data for better quality. Some selected data may have different formats because the data is very huge and they stored the data in damp files. Then, in ordered to use the data it needs to convert in to suitable format. Because purpose of the preprocessing stage is to cleanse the data as much as possible and to put it into a form that is suitable for use in later stages. Starting from the data extracted from the source.

### 4.3.1. Data selection

This phase uses to create a target dataset. The whole target dataset may not be taken for the DM task. Irrelevant or unnecessary data are eliminated from the DM database before starting the actual DM function. Originally there were around 1.8 terabytes records in six month of September 2012 up to February 2013. An average of each month is about 300 GB. The researcher uses purposive sampling technique. So that the researcher selected the high duration users that means if the user monthly call duration greater than 200000 second or about greater than 56 hours each month picking for the first sampling technique.

The data was still very huge and then needs other sampling techniques. In the second the researcher sued similar sampling technique to select the subscribers. Then the researcher selected high duration users that means if the user call duration greater than 2000 second or about greater than 34 minutes each a single call for second sampling technique. The data was still very huge and difficult to use and manage in the application software; finally the researcher has been used purposive sampling technique. And then the researcher selected 25000 records from the huge CDR data.

Since this dataset contains irrelevant and unnecessary data, all are not used for training. Those irrelevant records for this study are CDMA, SMS, fixed telephone and data services etc are eliminating from the record. So, after eliminating irrelevant and unnecessary data only a total of 21367 datasets are used for the purpose of conducting this study.

### 4.3.2. Data cleaning

This phase is used for making sure that the data is free from different errors. Otherwise, different operations like removing or reducing noise by applying smoothing techniques, correcting missing values by replacing with the most commonly occurring value for that attribute (Witten, & Frank, 2005).

The researcher was cleaned the data, by removing the records that had incomplete (invalid) data and/or missing values under each column. Removing of such records was done as the records with this nature are few and their removal does not affect the entire dataset. The researcher makes use of the MS Access 2007 and MS-Excel 2007 application for cleaning the data.

Accordingly, the PREFIX, BILLING_NBR, BILLING_IMSI, MSRN, LAC_A, CELL_A, LAC_B, CELL_B, PLAN_NBR, OFFER_NBR, THIRD_NBR, PART_ID, BASIC_CHARGE, BENEFIT_CHARGE, BILLING_CYCLE_ID, RECORD_SEQ, EVENT_INST_ID, RE_ID, MSC, FEE2 FEE3 and FEE4 attributes was not important for this research. Then the researcher deleted the whole column since it is meaningless to use those attributes. Although it was a very important variable, the CALLING_AREA, CALLED_AREA, TRUNK_IN, TRUNK_OUT and CDR_TYPE attribute is also deleted because almost all of its values are missed.

### 4.3.3. Data integration

The data integration process was done before deriving the attributes. As described before the dataset which was discussed above were available in different excel files. Data integration method for retrieving important fields from different files and tables was done in the effort to prepare the data ready for the DM techniques to be undertaken in this research.

First the Oracle databases were used to carry out the data extraction process. By using the Oracle databases process the huge damp CDR files and save in different excel files for each month. These data integration process took a lot of time of the research.

Because of the reason that the data was very huge and saved in different damp files and after processed by Oracle databases also saved as a Comma Delimited in indifferent files.

Then the CSV File Splitter 1.1version software was used to split file data. By using this software the data was splitted until the size of MS Access 2007 and MS-Excel 2007 application can read. Till understanding and solving the problem lots of time was lost and finally the data is integrated and put together into a single excel file.

### 4.3.4. Data formatting

The datasets provided to WEKA were prepared in a format that is acceptable by WEKA. It accepts records whose attribute values are separated by commas and saved in an ARFF (Attribute Relation File Format).
The excel file was first changed into a comma delimited (CSV) file format. Next the file was saved with ARFF (Attribute Relation File Format) file extension. Now the dataset, which is in ARFF format, is ready to be used in the WEKA software.

### 4.3.5. Attribute selection

Data mining methods, such as attribute selection and attribute relevance ranking, may help identify important factors and eliminate irrelevant ones. Most machine learning algorithms are designed to learn in which are the most appropriate attributes to use for making their decisions. For example, decision tree methods choose the most promising attribute to split on at each point and should—in theory—never select irrelevant or unhelpful attributes (Witten, & Frank, 2005).

The importance of reducing the number of attributes not only speed up the learning process, but also prevents most of the learning algorithms from getting fooled into generating an inferior

model by the presence of many irrelevant or redundant attributes. So that very limited numbers of attributes that are most important for the study at hand are selected because of this reason.
In order to select the best attributes from this initial collected dataset, the researcher evaluates the information content of the attributes with helping the domain expert.

Hence, as discussed before at the data cleaning stage, the researcher together with the domain expert removes those attributes, which have less important for this research and the remain as shown in the table 4.2 are the final list of attributes that have been used in this study.

| No. | Attribute Name | Data Type | Description | Attributes remark |
|---|---|---|---|---|
| 1 | START_TIME | Date | Time of call initiation (calling time) or The time when the call is begins | Original |
| 2 | DURATION | Numeric | Total calling time or the duration of call in second(Call duration in seconds) | Original |
| 3 | CHARGE | Numeric | Amount of birr paid ( to be paid for the case of post-paid mobiles) | Original |
| 4 | CALLING_NBR | String | Mobile number initiating or originating the call | Original |
| 5 | CALLED_NBR | String | Mobile number receiving the call | Original |
| 6 | FEE | Numeric | Amount paid or to be paid | Original |
| 7 | Calling & called number | Nominal | Cheek the validity of the calling and called number | Derived |
| 8 | Charge & duration status | Nominal | The balance between the duration of call and the amount birr paid if it is normal situation | Derived |
| 9 | payment status | Nominal | To cheek whether or not pay per call in normal situation. | Derived |
| 10 | Call Cost | Numeric | Calculate current cost per call | Derived |
| 11 | Calling Time | Nominal | To cheek whether or the calling time is on peak or off-peak hours | Derived |
| 12 | Calling Day | Nominal | To cheek the off-peak hours on Sunday or night | Derived |
| 13 | Call Type | Nominal | To identify the call is local or international | Derived |
| 14 | calling zone | Nominal | Identify the calling is to/from witch zone. According to Ethio-telecom there are three different tariff zones. Zone 1 (Asia, Europe, Middle East and North America) and Zone 2 (Africa, Oceania and South America) or local call. | Derived |

Table 4-2 Final list of attributes used in the study for per paid mobile subscriber

# CHAPTER FIVE

# 5. EXPERIMENTATION

## 5.1. Model Building

Modeling is one of the major tasks which undertaken under the phase of data mining in hybrid methodology. The motivation why this model is selected in this study, because of hybrid of process model is both academic and industrial and already discussed in chapter one section 1.4.1 and in the second chapter section 2.1.9. In this phase, different techniques can be employed for the data mining problems. Some of the tasks include: - selecting the modeling technique, experimental setup, building a model and evaluating the model.

The output of series of experiments of classification models are analyzed and evaluated in terms of the details of the confusion matrix of the model. Furthermore, models with different algorithms were compared with respect to their speed and accuracy. This

### 5.1.1. Selecting Modeling Technique

In this research, the supervised classification techniques are adopted. Selecting appropriate model depends on data mining goals. Consequently, to attain the objectives of this research, four classification techniques have been selected for model building. The analysis was performed using WEKA environment. Among the different available classification algorithms in WEKA, the selected algorithms are J48, PART, Random forest and Multilayer perceptron are used for experimentation of this study. In this work an attempts were done to build models using selected algorithms for classification of frauds. The models were evaluated based on confusion matrix, performance measures such that Precision Recall F-measure and accuracy etc. finally the performance of the models were compared.

### 5.1.2. WEKA Interfaces

WEKA (Waikato Environment for Knowledge Analysis) is a machine learning and data mining software tool written in Java and distributed under the GNU Public License. The goal of the

WEKA project is to build a state-of-the-art facility for developing machine learning techniques and to apply them to real-world data mining problems. It contains several standard data mining techniques, including data preprocessing, classification, regression, clustering, and association. Although most users of WEKA are researchers and industrial scientists, it is also widely used for academic purposes (Baumgartner, & Serpen, 2009).


Figure 5.1 WEKA interface

WEKA has four interfaces, which start from the main GUI Chooser window, as shown in figure 5.1 each interface has a specific purpose and utility. Whereas the Explorer and Knowledge Flow are tailored to beginning users, the experimenter and simple CLI target more advanced users. (This discussion is based on WEKA version 3.7.9).The buttons can be used to start the following applications:

- Explorer: An environment for exploring data with WEKA (the rest of this documentation deals with this application in more detail).
- Experimenter: An environment for performing experiments and conducting statistical tests between learning schemes.
- Knowledge Flow: This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning.
- Simple CLI Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

The Explorer is possibly the first interface that new users will run simulations in. It allows for data visualization and preprocessing. In this study the researcher uses Explorer environment to conduct the experiment.

## 5.2. Experiment design

The model is built based on the default 66% percentage and 10-fold cross validation.
The default ratio is 66% for training and 34% for testing, and default parameter for the model training and testing which is 10-fold cross validation.

In 10-fold cross validation, the initial data are randomly partitioned into 10 mutually exclusive subsets or folds, 1,2,3,4 ...10, each approximately equal size. The training and testing is performed 10 times. In the first iteration, the first fold is reserved as a test set, and the remaining 9 folds are collectively used to train the classifier.
Train classifier on folds: 2 3 4 5 6 7 8 9 10; Test against fold: 1
Train classifier on folds: 1 3 4 5 6 7 8 9 10; Test against fold: 2
Train classifier on folds: 1 2 4 5 6 7 8 9 10; Test against fold: 3
Train classifier on folds: 1 2 3 5 6 7 8 9 10; Test against fold: 4
Train classifier on folds: 1 2 3 4 6 7 8 9 10; Test against fold: 5
Train classifier on folds: 1 2 3 4 5 7 8 9 10; Test against fold: 6
Train classifier on folds: 1 2 3 4 5 6 8 9 10; Test against fold: 7
Train classifier on folds: 1 2 3 4 5 6 7 9 10; Test against fold: 8
Train classifier on folds: 1 2 3 4 5 6 7 8 10; Test against fold: 9
Train classifier on folds: 1 2 3 4 5 6 7 8 9; Test against fold: 10
The classifier of the second iteration is trained on folds 1, 3, 4... 10 and tested on the 2nd fold etc. The accuracy estimate is the overall number of correct classifications from the 10 iterations divided by the total number of samples in the initial dataset (Han and Kamber, 2001).

Generally a procedure or mechanism was used to test the model's quality and validity is needed to be set before the model is actually built. In order to perform the model building process of this study, the researcher used 21367 dataset with 14 attributes to investigate the model. From the total data 3091was used as training dataset and the remains 18276 used as testing dataset to

evaluate the prediction performance of the classification model developed. The training and testing dataset was prepared by purposive sampling technique from the original dataset. The original dataset is presented in appendix 1 and the graphical representation of the sample training and test dataset are show in appendix 2 and appendix 3 respectively

## 5.3. J48 Decision tree model building

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value

Generally, this research is more interested in generating rules that best predict the mobile subscription fraud and to come to an understanding of the most important factors (variables) affecting the mobile telecom to be fraudulent.

As shown in the figure 5.2, J48 is one of the most common decision tree algorithms that are used today to implement classification technique using WEKA. This algorithm is implemented by modifying parameters such as confidence factor, pruning and unpruning, hanging the generalized binary split decision classification and other option available in table 5.1. Therefore, it is very crucial to understand the available options to implement the algorithms, as it can make a significance difference in the quality of the result. In many cases, the default setting will prove adequate, but to compare results /models and attain the research objectives other options are considered (Han and Kamber, 2006).

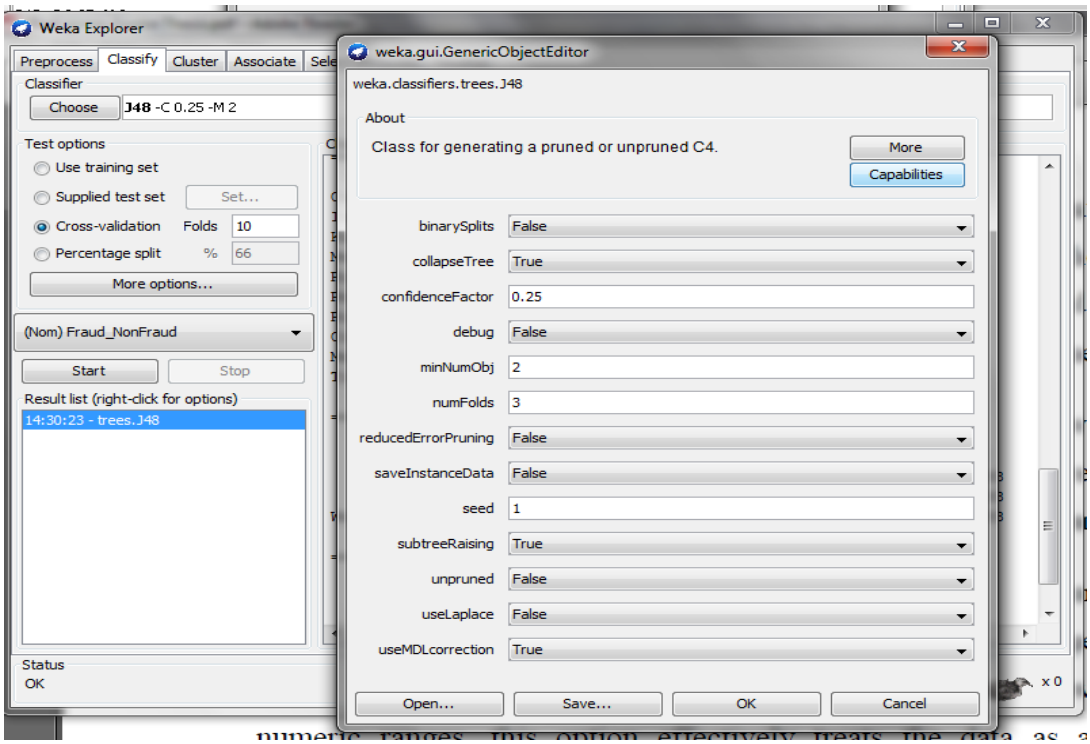



Figure 5.2 The snapshot of J48 algorithms setting

| Name | Default value | Possible values | Description |
|---|---|---|---|
| Confidence Factor | 0.25 | 10 - --0.5 | The confidence factor used for pruning (smaller values incur more pruning) |
| minNumObj | 2 | 1,2 | The minimum number of instances per leaf |
| Unpruned | False | True/ False | Use un pruned tree (the default value 'no' means that the tree is pruned). |
| Sub tree raising | True | True/ False | Whether to consider the sub tree raising operation in post pruning. |
| B-use binary splits | True | True/ False | Whether to use binary splits on nominal attributes when building the tree. |

Table 5-1 Some of the J48 algorithm parameters and their default values

Experiment 1

The first experimentation is performed with the default parameters. The default 10-fold cross validation test option is employed for training and testing the classification model.

Table 5.2 shows the resulting confusion matrix of J48 algorithm with 10-fold cross validation of the model.

| Confusion Matrix | | |
|---|---|---|
| a | b | classified as |
| 3843 | 13 | a = fraud |
| 7 | 17504 | b = notFraud |

Table 5-2 Confusion matrix output of the J48 algorithm with 10-fold cross validation

As shown from the resulting confusion matrix, the J48 learning algorithm scored an accuracy of 99.9064%. This result shows that out of the total training datasets 21347 (99.9064 %) records are correctly classified instances, while only 20 (0.0936 %) of the records are incorrectly classified. The snapshot running information with 10-fold validation technique provided on appendix 5:

Experiment 2

This experiment is performed, by changing the 10-fold cross validation to the default value of the percentage split (66%). In this learning scheme a percentage split is used to partition the dataset into training and testing data. With 66% of the data used for training and 34% used for testing.

The purpose of using this parameter was to assess the performance of the learning scheme by changing the 10-fold cross validation to the default value of the percentage split (66%) in order to testing dataset if it could achieved a better classification accuracy than the first experimentation. The result of this learning scheme is summarized and presented in table 5.3

| Confusion Matrix | | |
|---|---|---|
| a | b | classified as |
| 1266 | 9 | a = fraud |
| 1 | 5989 | b = notFraud |

Table 5-3 Confusion matrix result of J48 algorithm with default values (66%)

In this experiment out of the 21367 total records 14102 (66%) of the records are used for training purpose while 7265 (34%) of the records are used for testing purpose. As shown on the confusion matrix of the model developed with this proportion, out of the 7255 testing records

99.8624% of them are correctly classified instances. Only 10 (0.1376 %) records are incorrectly classified instances.

In this experiment when the testing data is increased the performance of the algorithm for predicting the newly coming instances also shows slightly increment until 80% of the training dataset but not better than the first experiment. The experiment is conducted by varying the value of the training and the testing datasets, the accuracy of the algorithm for predicting new instances in their respective class couldn't be improved. This shows that the previous experiment conducted with the default 10-fold cross validation, is better than this experiment.

Generally, from the two experiments conducted before, the model developed with the 10-fold cross validation test option gives a better classification accuracy of predicting newly arriving subscription fraud in their respective class category. Therefore, among the different decision tree models built in the foregoing experimentations, the first model, with the 10-fold cross validation, has been chosen due to its better classification accuracy.

## 5.4. PART Rule Induction model building

The second data mining algorithm applied in this research was PART Rule induction algorithm. The researcher selected PART for the reason that PART has the ability and to produce accurate and easily interpretable rules that helps to achieve the research objectives and have the advantages of simpler and has been found to give sufficiently strong rules
(Krishnaveni & Hemalatha, 2011).

To build the Rule induction model, 21367 dataset was used as an input to the system. The test options used for the experiment are 10-fold cross validation and default value of the percentage split (66%). The confusion matrix result of PART Rule induction algorithm is shown in table 5.4.

Experiment 3

| Confusion Matrix | | |
|---|---|---|
| a | b | classified as |
| 3844 | 12 | a = fraud |
| 5 | 17506 | b = notFraud |

Table 5-4 Confusion matrix result of PART algorithm with 10-fold cross validation

As shown from the resulting confusion matrix, of the PART Rule induction algorithm with 10-fold cross validation scored an accuracy of 99.9204 %).

This result shows that out of the total training datasets 21350 (99.9204 %) records are correctly classified instances, while only 17 (0.0796 %) of the records are incorrectly classified instances. The snapshot running information with 10-fold validation technique provided on appendix 6.

Experiment 4

| Confusion Matrix | | |
|---|---|---|
| a | b | classified as |
| 1265 | 10 | a = fraud |
| 2 | 5988 | b = notFraud |

Table 5-5 Confusion matrix result of PART Rule induction with default values (66%)

As shown from the resulting confusion matrix, of the PART Rule induction algorithms with default values (66%) scored an accuracy of (99.8348 %)This result shows that out of the total training datasets 7253 (99.8348 %) records are correctly classified instances, while only 12 (0.1652 %) of the records are incorrectly classified instances. This shows that the previous experiment conducted with the default 10-fold cross validation, is better than this experiment.

Generally, when the researcher compared the two experiments conducted before, the PART Rule induction algorithms with 10-fold cross and the PART with default values (66%), the model developed with the 10-fold cross validation test option gives a better classification accuracy of

predicting newly arriving telecom fraud. So that the first model, of 10-fold cross validation has been chosen due to its better classification accuracy.

## 5.5. Random Forest model building

In this experiment random forest algorithm is explored. It builds an ensemble of several models instead of building a single one, and prediction of the ensemble model is made as a consensus of predictions made by all its individual members. The model is built based on the default values of 66% percentage split and 10-fold cross validation.

Experiment 5

| Confusion Matrix | | |
|---|---|---|
| a | b | classified as |
| 3850 | 6 | a = fraud |
| 10 | 17501 | b = notFraud |

Table 5-6 Confusion matrix result Random Forest algorithms with 10-fold cross validation

As shown from table 5-6 Random Forest algorithm with 10-fold cross validation achieved an accuracy of 99.9251 %. This result shows that out of the total training datasets 21351 (99.9251 %) records are correctly classified instances, while only 16 (0.0749 %) of the records are incorrectly classified instances.

Experiment 6

| Confusion Matrix | | |
|---|---|---|
| a | b | classified as |
| 1265 | 10 | a = fraud |
| 1 | 5989 | b = notFraud |

Table 5-7 Confusion matrix results of Random Forest algorithms with default of values 66%

As shown from the resulting confusion matrix, of the Random Forest algorithms with default of values 66% scored an accuracy of (99.8348 %). This result shows that out of the total training datasets 7254 (99.8486 %) records are correctly classified instances, while only 11 (0.1514 %) of the records are incorrectly classified instances.

Generally, when we compared from the two experiments conducted, that are the Random Forest algorithms with 10-fold cross and the Random Forest algorithms with default values (66%), the model developed with the 10-fold cross validation test option gives a better classification accuracy of predicting. So that the first model, of 10-fold cross validation has been chosen due to its better classification accuracy as before.

## 5.6. Artificial Neural Network model building

Neural Network is a field in Artificial Intelligence (AI) where we use data structures and algorithms for learning and classification of data, by inspiration from the human brain.

Many tasks that humans perform naturally fast, such as the recognition of a familiar face, proves to be a very complicated task for a computer when conventional programming methods are used. By applying Neural Network techniques, a program can learn by examples, and create an internal structure of rules to classify different inputs, such as recognizing images (Nielsen, 2001).

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

The data is first normalized to the range [-1, 1] which is suitable for the neural network algorithm. ANN model learns very fast if the attribute values are normalized to the range [-1, 1]. The ANN model is built for categorical label classes, which are derived from the application prediction of unknown dataset. In order to normalize the values of the attributes, the researcher used WEKA's pre-processing facilities so that all the variables fall in the range [-1, 1].

Most of the values of the variables of the dataset of this research were categorical or nominal. Thus the values should be changed into numeric values for normalization. Hence the distinct values of the categorical attributes have been assigned numeric values as shown in the following table 5.8.

| **Attributes** | **Represented as** |
|---|---|
| Calling&called_number | different=2 and Similar=1 |
| charge&duration_status | Normal=1 and abnormal=2 |
| payment_status | Normal=1 and abnormal=2 |
| CallingTime | OFFPEAK= 2 and PEAK =1 |
| CallingDay | notSunday=1 and Sunday=2 |
| CallType | Local=1 and international=2 |
| calling_zone | Zone1=2, zone2=3 and local=1 |
| Fraud_NonFraud | notFraud =2 and fraud=1 |

Table 5-8 Representing the nominal values of the attributes by numeric values

After mapping these nominal values into the above stated numerical values, the WEKA's explorer normalize pre-processing facility has been used to normalize all values to fall in the range [-1, 1]. The attributes used in the other model are used in the neural network Multilayer perceptron algorithm as well. The sample of normalized data for Multilayer perceptron algorithm is presented on appendix 8.

After all necessary pre-processing is done; the neural network model experimentation is tested by changing the values of the hidden layer of Multilayer perceptron algorithm with 10-fold cross and the Multilayer perceptron algorithm with default values (66%). Table 5.9 shows the resulting confusion matrix of Multilayer perceptron algorithm with 10-fold cross validation of the model.
Multilayer perceptron ANN is further used to compare its performance against other methods. The ANN model is tested by passing different parameters and the result of the classification accuracy is reported and its performance is also compared in classifying new instances of records.

Experiment 7

| Confusion Matrix | | |
|---|---|---|
| a | b | classified as |
| 3763 | 93 | a = fraud |
| 4 | 17507 | b = notFraud |

Table 5-9 Confusion matrix result of Multilayer perceptron with 10-fold cross validation

As observed from the resulting of the above confusion matrix, the Multilayer perceptron algorithms with 10-fold cross validation scored an accuracy of 99.546%. This result shows that out of the total training datasets 21270 (99.546%) records are correctly classified instances, while only 97 (0.454%) of the records are incorrectly classified instances.: The running information of ANN with 10-fold validation technique is shown appendix 9.

Experiment 8

| Confusion Matrix | | |
|---|---|---|
| a | b | classified as |
| 1233 | 42 | a = fraud |
| 1 | 5989 | b = notFraud |

Table 5-10 Confusion matrix result of Multilayer perceptron with default values (66%).

As we observed from the resulting confusion matrix presented above, the Multilayer perceptron algorithms with default values (66%) scored an accuracy of 99.4081 %. This result shows that out of the total training datasets 7222 (99.4081 %records are correctly classified instances, while only 43 (0.5919 %) of the records are incorrectly classified instances.

Generally, when the researcher compared from the two experiments conducted, that are the Multilayer perceptron algorithms with 10-fold cross validation and the Multilayer perceptron algorithms with default values (66%), the model developed with the 10-fold cross validation test option gives a better classification accuracy of predicting. So that the first model, of 10-fold cross validation has been chosen due to its better classification accuracy as before.

## 5.7. Evaluation and Comparison of J48, PART, RF and MLPs ANN Models

Selecting a better classification technique for building a model, which performs best in handling the prediction and identifying the telecommunication frauds are one of the aims of this study. For that reason, four classification models are selected. The J48 and Multilayer perceptron ANN recommended by the many researchers on the field of telecom frauds and PART and Random forest are interest to test on this research. For each algorithm the best performance accuracy are listed in table 5.11 and table 5.12 below.

| Types of algorithms | Performance registered (%) | Time taken(sec.) | Correctly classified | Misclassified |
|---|---|---|---|---|
| J48 | 99.906 | 1.62 | 21347 | 20 |
| MLPs | 99.546 | 170.61 | 21270 | 97 |
| PART | 99.9204 | 1.2 | 21350 | 17 |
| Random forest | 99.9251 | 2.34 | 21351 | 16 |

Table 5-11 Comparison of the result of the selected models with 10 fold cross validation

| Types of algorithms | Performance registered (%) | Time taken(sec.) | Correctly classified | Misclassified |
|---|---|---|---|---|
| J48 | 99.86 | 1.49 | 7255 | 10 |
| MLPs | 99.4081 | 153.66 | 7222 | 43 |
| PART | 99.834 | 1.65 | 7253 | 12 |
| Random forest | 99.8486 | 2.47 | 7254 | 11 |

Table 5-12 Comparison of the result of the selected models with default of values 66%

As described above one of the basic aims of data mining is to compare different models and to select the better classification accuracy accordingly. Therefore, detailed experimentation for different models has been conducted. As a result, the best classification algorithm which is appropriate for this problem domain has been selected.

The researcher in collaboration with the domain experts selected the best model from the four algorithms and different options like default values (66%), 10-fold cross validation and by changing different parameter options.

The experts explained that more attention should be given to high duration user, high total number of call times and low payers or not paying customers because the major behaviors of subscription fraudsters are identified by not paying or low level of paying, talking long time and calling so many times in a day and finally may be the subscriber or the customer is in active or disappear for the network.

The results of the overall classification, the mobile telecommunication fraudsters and non fraudsters for both classification models are described in detailed as follows.

The 10-fold cross validation and default of values 66% are used to compare these four models, where the default parameter is taken for both cases. The decision tree model with 10-fold cross validation has a better classification accuracy of 21347 (99.9064 %) which correctly classified and 20 (0.0936 %) that are wrongly classified.

Again the decision tree model with default of values 66% has 7255 testing records 99.8624% of them are correctly classified instances and only 10 (0.1376 %) records are incorrectly classified instances. Then the maximum accuracy of J48 model is 99.9064 % with 10-fold cross validation
The PART Rule induction algorithms with 10-fold cross validation scored an accuracy of 99.9204 % out of the total training datasets 21350 (99.9204 %) records are correctly classified instances, while only 17 (0.0796 %) of the records are incorrectly classified instances.

The second experiment of the PART Rule induction algorithms with default values (66%). From the total training datasets 7253 (99.8348%) records are correctly classified instances, while 12 (0.1652 %) of the records are incorrectly classified instances.

This shows that the first experiment conducted with the default 10-fold cross validation, is better than the default values (66%) experiment. And then the maximum accuracy of PART Rule induction algorithms model is 99.9204% with 10-fold cross validation.

The third experiment of the Random Forest (RF) algorithms with 10-fold cross validation scored an accuracy of 99.9251% from the total training datasets 21351 (99.9251%) records are correctly classified instances, while only 16 (0.0749 %) of the records are incorrectly classified instances.

The second of the Random Forest algorithm use the default of values 66% get the accuracy of 99.8348% This result shows that out of the total training datasets 7254 (99.8486 %) records are correctly classified instances, while only 11 (0.1514%) of the records are incorrectly classified instances. Generally, when we compared the Random Forest algorithms with 10-fold cross and with default values (66%), the results both algorithms 99.9251% and 99.8348%, 10-fold cross validation and default values (66%), respectively. So that the accuracy model developed with the 10-fold cross validation is better than the default values (66%) accuracy. Then the maximum accuracy of Random Forest algorithm is 99.9251 % with 10-fold cross validation.

The final experiment on the algorithm is the Multilayer perceptron algorithms with 10-fold cross validation scored an accuracy of 99.546% from the that out of the total training datasets 21270 (99.546%) records are correctly classified instances, while only 97 (0.454%) of the records are incorrectly classified instances. The second experiment in this algorithm is applied with default values (66%) and scored an accuracy of 99.4081 % from the total training datasets 7222 (99.4081 %) records are correctly classified instances, while only 43 (0.5919 %) of the records are incorrectly classified instances.

Generally, when we compared from the two experiments conducted, that are the Multilayer perceptron algorithms with 10-fold cross validation with default values (66%), as usual the 10-fold cross validation is better than the other. The overall accuracy of the Multilayer perceptron algorithms the model scored 99.546%.

From the above results, the overall performance of the 10-fold cross validation experiment was better than the other options. And when we compare the different algorithms, the PART Rule induction algorithm and Random Forest (RF) algorithm with 10-fold cross validation scored accuracy 99.9204% and 99.9251 % each respectively. Those two algorithms are similar but Random Forest (RF) algorithm is slightly better than the PART Rule induction algorithm.

The remains of two algorithms, J48 decision tree and Multilayer perceptron algorithm are also good accuracy of 99.906% and 99.546 % each respectively. But the J48 decision tree is better than the Multilayer perceptron algorithm. Hence Random Forest (RF) algorithm with 10-fold cross validation is the model of this study the snapshot of the random forest algorithm is shown in appendix 7.

## 5.8. Discussion of the result with domain experts

This step includes understanding the results, checking whether the new information is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Usually, real world databases contain incomplete, noisy and inconsistent data and such unclean data may cause confusion for the data mining process (Han and Kamber, 2006).

Besides, data preparation task is a vital so that data mining tools can understand the data. Thus, data cleaning and analysis methods and outlier mining methods becomes a must to undertake such gap before discovered knowledge so as to enhance the efficiency and effectiveness of the data mining technique. In this effort, the researcher tried to assess such problems which require due emphasis while making the data ready for applying data integration and other different data reduction techniques.

As discussed in the chapter four in section 4.2.1 the source of data for this research has been collected from Ethio telecom database six month records. Due to the fact that the CDR data is very huge and requires more space, the Ethio telecom servers stored no more than six month data. Therefore, the researcher took six month data of the telecommunication company.

Originally the CDR data were around 1.8 terabytes records in six month of September 2012, November 2012, October 2012, December 2012, January 2013 and February 2013 and have more than 35 fields in per paid mobile services. An average of each month is about 300 GB. The researcher has used the purposive sampling to extract the dada. In ordered to extract the huge damp file from the Ethio-telecom database, the researcher used oracle database software. And in ordered to manage the data in application software (in MS-Access and MS-Excel), the researcher has used file splinter software.

After the necessary process is done the researcher took 21367 datasets have been used for the purpose of conducting this study. And from 35 field or initial attributes, the researcher evaluates the information content of the attributes in collaborated the domain experts 14 most appropriate attributes have been used for conducting this research.

To extract and to manage the huge CDR from the telecommunication server and to integration process of the data have been the challenged times of the researcher. Therefore, the researcher took a lot of time for this process. As discussed in chapter two and five, in selection of model building, four classification models was selected and developed for reason explained below. Firstly, the J48 decision tree algorithm is chosen for its simplicity and easy interpretability of the result generated by it. Secondly PART is selected similarly as of J48 and simplicity and generated list of rules easy to interpret and associate the problem domain.

Third Random Forest (RF) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees and because of the following advantages no need for pruning trees, Accuracy and variable importance generated automatically, Overfitting is not a problem, Not very sensitive to outliers in training data and easy to set parameters (Breiman, 2001). Finally, the applying artificial Neural Network

techniques because of many researchers used this algorithm in different areas of researches and so that the final algorithm is the Multilayer perceptron of artificial Neural Network techniques.

One of the basic targets of data mining is to compare different models and to select the better classification accuracy accordingly. Therefore, detailed experimentation for different models has been conducted. Accordingly, the best classification algorithm which is appropriate for this problem domain has been selected. The classification model used in this research to predict pre paid mobile telecommunication frauds.

The finding interpreted in chapter five, about the telecom subscription of pre paid mobile fraud have domain expert accepted a criteria. That are: Calling date and time, the amount of time talking, the birr to be paid, call initiating telephone number, call received telephone number, and the amount of birr currently paid, types of calling local or international, considering the tariff zone and peak or peak off hour tariff. Some of these attributes are also important for domain expert practically practicing to protect the telecommunication frauds.

The analysis, which was closely undertaken with domain experts are, achieved a good result. Among the four models, the Random Forest algorithm registers better than others and its accuracy of 99.9251%. This research will play an important role to controlling and preventive the current threat on the mobile communication of Ethio telecom.

Now a day the telecom fraud is very serious problem that's is why the ARTICLE 19, declared by the Ethiopia House of Peoples' Representatives passed the Proclamation on 11 July 2012, already discussed on chapter one in section 1.1.4. The Proclamation identifies "telecom fraud offences" as a "serious threat to national security beyond economic losses" so that this research will open new research areas in the field of subscription telecom fraud detection and preventive mechanism.

In order to analyze the natures and character of fraudsters in Ethio–telecom, the researcher in cooperation with domain experts have discussed the result of this study. The natures of the fraudsters in Ethio-telecom practicing telecom frauds are to achieve the following advantages.

The pattern shows that from the numerous reasons to commit telecommunications fraud exists, including:

- To avoid charges.
- To make money.
- To maintain anonymity.
- To demonstrate intellectual superiority.

The researcher discussed with INSA and etthio-telecom domain expert, then the most difficult aspect of fighting fraud is how can detect the subscription frauds in the Ethio- telecom. Because the major behavior the customers who made subscription frauds are calling more timing without intention to paying and disappearing from the network.

As discussed before from those experiments conducted in supervised approach, the J48 decision tree algorithm with the 10-fold cross validation and random forest tree algorithm model gives a better classification accuracy of predicting newly arriving class category. The rules indicate the possible conditions in which the CDR record could be classified in each of the fraud and non-fraud of telecom subscribers. Some of the prevailing rules are shown in the appendix 11.

The researcher faced different challenges while conducting this study. The first challenge was the dataset obtained from the Ethio-telecom, Because of the researcher used six month data. An average of each month is about 300 GB. As the amount of data increases it is really challenged task because of the memory size problem and to access the huge damp files.

The preprocessing task of this study was also another challenge. Especially, selecting important attributes for this study, integrating the different fields to build appropriate model of data mining tool and driving classification rules manually from the decision tree was not simple task.

The third challenge was related to understanding the domain area. The researcher had tried to work together with the domain experts. Accordingly the researcher took much time in data collection and prepossessing. Finally, the results obtained from this research indicate that data mining is useful in bringing relevant information to the service providers as well as decision makers.

# CHAPTER SIX

# 6. CONCLUSION AND RECOMMENDATIONS

## 6.1 Conclusion

Mobile Technology makes use of wireless communication devices that can be carried anywhere, as they require no physical connection to any external wires to work. A new era of mobile multimedia applications and services has been brought about by the rapid growth of Wireless Communication and access. However, mobile technology is not without its own problems.

Fraud is prevalent in both fixed and mobile networks of all technologies. Telecommunication Companies worldwide suffer from customers who use the provided services without paying. Fraud can be defined as criminal deception or the theft of telecommunication service. It is the use of false representation to gain an unjust advantage.

Subscription fraud is the starting point for many other telecoms fraud scams and such is recognized as the most damaging of non-technical fraud types. Subscription fraud is still the most prevalent fraud type faced by telecoms operators today.
Data mining is an effective method for detecting various types of fraud including mobile telecommunication systems.

The objective of this research was to develop model for detecting and predicting of telecom frauds using data mining techniques in Ethiopia mobile communication services provision.

To achieve this objective the researcher used call Detail Record (CDR) data. Because of the CDR data includes sufficient information about subscriber.

The researcher selected around 25000 subscribers from the huge CDR data. After eliminating irrelevant and unnecessary data only a total of 21367 datasets are used for the purpose of conducting this study. Fifteen attributes are selected from 35 initial attributes or the fields. It has been preprocessed and prepared in a format suitable for the DM tasks

Hybrid process model was used while undertaking the experimentation. The study was conducted using WEKA software version 3.7.9 and four data mining algorithms for classification techniques was used, namely J48, PART, Random Forest and Multilayer perceptron of Arterial neural network.

The Random Forest algorithm registered better performance of 99.9251% accuracy running with 10-fold cross validation using 14 attribute than any experimentation done for this research. On the other hand J48, PART, and Multilayer perceptron of arterial neural network algorithms were registered encouraging performance of classification accuracy. The result shows 99.906%, 99.9204% and 99.546% respectively with 10 fold cross validation.

It is described how data mining can be used to uncover useful information hidden within these data sets. Several data mining applications are described and together they demonstrate that data mining can be used to identify telecommunication fraud. In general, in this study only few experiments are carried out on the pre-paid fraud mobile communication using the algorithms of J48, PART, and Random Forest and Multilayer perceptron with few parameter setting. PART, Random Forest is shown better accuracy than the others.

The researcher has faced different challenges in conducting this study. The first challenge was to obtained dataset and the memory size problem. Second challenged task was to access the huge damp files. The third challenge was to preprocessing the data and selecting those important attributes for this study. Finally, the challenge was related to understanding the domain area. The researcher has tried to work together with the domain experts. Consequently the researcher took much time in collecting, accessing and prepossessing of the CDR data.

## 6.2 Recommendations

This research is mainly conducted for an academic purpose. This research has proven the applicability different DM classification techniques namely, J48, Random Forest, PART and Multilayer perceptron of artificial Neural Network algorithms which automatically discover hidden knowledge that are interesting and accepted by domain expert. The researcher gives the following recommendations based on the investigations of the study.

- In this research the researcher use only the CDR data; however further investigation is needed by including the other types of telecommunication data.
- This research is limited on pre paid subscription of telecommunication fraud. Research could be done on post paid subscription of telecommunication fraud

- As described above this research attempted to assess the potential application of DM techniques in detecting fraudulent subscription of telecommunication fraud by, however so many types of fraud are available today. Then detail research could be needed on other than the subscription of telecommunication fraud.

- Integration of the discovered classification rules with knowledge based system

- On the CDR data there are many missing values are available and some of the fields are empty. Then the company could complete the empty fields and the missing fields.

- Currently INSA and Ethio-telecom are working on telecommunication fraud detection. However, these organizations are performing their duties independently of each other. So the researcher recommends that both organizations should work in collaboration in ordered to enhance the detection and prevention mechanisms of telecommunication fraud

# REFERENCE


Abidogun, O. A. (2005). Data mining, fraud detection and mobile telecommunications: Call pattern analysis with unsupervised neural networks (Doctoral dissertation, University of the Western Cape).

Azevedo, A. I. R. L. (2008). KDD, SEMMA and CRISP-DM: a parallel overview.

Baumgartner, D., & Serpen, G. (2009). Large experiment and evaluation tool for WEKA classifiers. In 5th international conference on data mining. Las Vegas (pp. 340-346).

Becker, R. A., & Wilks, A. R. (2011). Fraud Detection in Telecommunications: History and Lessons Learned.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brooks, John., Mohamad,M., Hussin,A., Soto,F., Jacob,V., Sinha,A. & Eisner,T.(May 2012). Fraud Classification Guide.TM Forum GB954

Brown, S. (2005). Telecommunication fraud management. waveroad security

Crano, W. D. and Brewer, M. B. (2002), Principles and Methods of Social Research, Lawrence Erlbaum Associates, Mahwah, NJ. 176

Cios, K. J., & Kurgan, L. A. (2005). Trends in data mining and knowledge discovery. Advanced techniques in knowledge discovery and data mining, 1-26.

Cios, K ,Witold, P., Roman, S., and Kurgan, A. (2007). Data Mining: A Knowledge Discovery Approach, New York, USA: Springer,

Cios, K., Swiniarski, R., Pedrycz, W., & Kurgan, L. (2007). The Knowledge Discovery Process. In Data Mining (pp. 9-24). Springer US.

Cohen, T., & Southwood, R. (2004). An overview of VoIP regulation in Africa: policy responses and proposals. *Commonwealth Telecommunications Organization (CTO)*

Communications Fraud Control Association. (2006). World-wide telecom fraud survey 2006

Communications Fraud Control Association. (2009). Global Fraud Loss Survey. *Press Release.*

Dunham, M. H. (2003). Data mining introductory and advanced topics. Upper Saddle River, NJ: Pearson Education, Inc.

Da Cunha, C., Agard, B., & Kusiak, A. (2006). Data mining for improvement of product quality. *International Journal of Production Research*, *44*(18-19), 4027-4041.

Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. *Data mining and knowledge discovery*, *1*(3), 291-316.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Fekadu, M.(2004). Application of data mining techniques to support customer relationship management at Ethiopian telecommunications corporation, Unpublished Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Gary, M. Weiss (2005). Data Mining in Telecommunications. In O. Maimon and L. Rokach (.eds), Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. Kluwer Academic Publishers, 1189-1201.

Gaur, P. (2012). Neural Networks in Data Mining. *International Journal of Electronics and Computer Science Engineering (IJECSE, ISSN: 2277-1956)*, *1*(03), 1449-1453.

Gobena, M. (2000).Flight Revenue Information Support System for Ethiopian Airlines, Unpublished Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Gebremeskel, G. (2006). Data mining application in supporting fraud detection: on Ethio-mobile services. Unpublished Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Govrenment Ethiopa. (2012). Ethiopia House of Peoples Representatives Proclamation on Telecom Fraud Offences. ARTICLE 19 , 22.

Ghosh, M. (2010). Telecoms fraud. *Computer Fraud & Security*, *2010*(7), 14-17.

Han, J., & Kamber, M. (2006). Data mining: concepts and techniques (Morgan-Kaufman Series of Data Management Systems), Academic Press, San Diego

Han, J. (1996, June). Data mining techniques. In *ACM SIGMOD Record*, international conference on Management of data (Vol. 25, No. 2, p. 545). ACM.

Hilas, C. S. (2012). Data mining approaches to fraud detection in telecommunications. 2nd Pan-Hellenic Conference on Electronics and Telecommunications - PACET'12, March 16-18, 2012, Thessaloniki, Greece

Ethiopian ICT. (2012. December 11). ICT Entrepreneurship Conference. Retrieved from federal democratic of republic Ethiopia ministray of forgien affres: http://www.mfa.gov.et/news/more.php?newsid=1471

Jember, G. (2005). Data mining application in supporting fraud detection on mobile communication: the case of Ethio mobile. Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *Knowledge Engineering Review*, *21*(1), 1-24.

Kumar, A. S. (2011). *Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains*. Igi Publishing.

Krishnaveni, S., & Hemalatha, M. (2011). A perspective analysis of traffic accident using data mining techniques. International Journal of Computer Applications, 23(7).

Larose, D. T. (2004). *Discovering knowledge in data: an introduction to data mining*. Wiley-Interscience.

Lloyd, D. (2003*). International Roaming Fraud*: Trends and Prevention Techniques. Fair Isaac Corporation.

Melaku, G. (2009). Applicability of Data Mining Techniques to Customer Relationship Management: The Case of Ethiopian Telecommunications Corporation (ETC) Code Division Multiple Access (CDMA Telephone Service). Unpublished Master of Science Thesis, School of Information Science, Addis Ababa University: Addis Ababa, Ethiopia.

Matignon, R., & SAS Institute. (2007). *Data mining using SAS enterprise miner* (Vol. 638). Wiley-Interscience.

Marco, R. & Gianluca, C. (2005). Data Mining Applied to Validation of Agent Based Models, Proceedings of Ninteenth European Conference on Modelling and Simulation, RIFA.

Mamcenko, J., & Beleviciute, I. (2007, May). Data mining for knowledge management in technology enhanced learning. In *Proceedings of the 6th conference on Applications of electrical engineering* (pp. 115-119). World Scientific and Engineering Academy and Society (WSEAS).

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

Ojuka, N. (2009). Detection of subscription fraud in telecommunications using decision tree learning.

Osmar, R. (1999). Principles of Knowledge Discovery in Databases CMPUT690

Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Springer.

Pete, C. (2000).Step by step Data mining guide. CRISP-DM consortium,.6: 60-67

Redl, S., Oliphant, M. W., Weber, M. K., & Weber, M. K. (1995). *An introduction to GSM*. Artech House, Inc.

Roiger, R. J. & Geatz, M. W. (2003). Data Mining: A Tutorial-Based Primer, Addison-Wesley, ISBN 0-201-74128-8, Boston.

Rokach, L. and Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C 35 (4): 476–487

Rejesus, R. M., Little, B. B., & Lovell, A. C. (2005). Using data mining to detect crop insurance fraud: is there a role for social scientists?. *Journal of Financial Crime*, *12*(1),

Rebahi, Y., Gouveia, F., Capsada, O., Nassar, M., Festor, O., Dagiuklas, T., & Wik Thorkildssen, H. (2010). SCAMSTOP: Scams and Fraud Detection in Voice over IP Networks FP7 contract number: 232458 D2. 1 Fraud in VoIP Networks: Size and Scope.

Singh, Y., & Chauhan, A. S. (2009). Neural networks in data mining. *Journal of Theoretical and Applied Information Technology*, *5*(6), 36-42.

Shawe-Taylor, J., Howker, K., & Burge, P. (1999). Detection of fraud in mobile telecommunications. *Information Security Technical Report*, *4*(1), 16-28.

Seifert, J. W. (2004). *Data mining: An overview*. *National security issues*, 201-217.

Stüber, G. L. (2011). *Principles of mobile communication*. Springer.

Taye, E. (2010). *telecommunication in Ethiopia. Geneva* : UNCTAD.

Triantaphyllou, E. (2010). *Data Mining and Knowledge Discovery via Logic-Based Methods: Theory, Algorithms, and Applications* (Vol. 43). Springer.

Two Crows Corporation. (1999). *Introduction to Data Mining and Knowledge Discovery*. Two Crows Corporation.

Van Heerden, J. H. (2005). *Detecting fraud in cellular telephone networks* (Doctoral dissertation, University of Stellenbosch).

Witten, I., and Frank, E. (2000). Data mining: Practical Machine Learning Tools and Techniques with Java Implementations, 2nd edn, Morgan Kaufmann publishers, San Francisco.

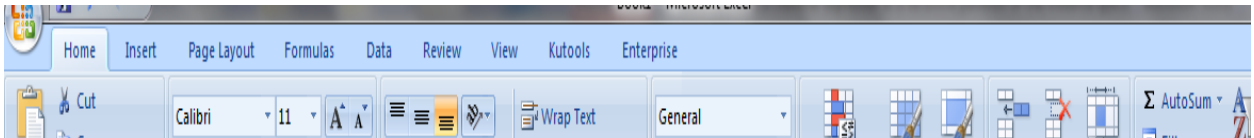
Williams, G. (2011). *Data Mining with Rattle and R: The art of excavating data for knowledge discovery*. Springer.

Witten, I. H., & Frank, E. 2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Weng, S., Chiu, R., Wang, B., & Su, S. (2006*). The study and verification of mathematical modeling for customer purchasing behavior*. *Journal of Computer Information Systems*, *47*(2), 46.

# APPENDICES

*Appendix 1: The original collected sample data*




M2 | 4

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | BILLING_NBR | START_TIME | DURATION | CHARGE | CALLING_NBR | CALLED_NBR | FILE_ID | EVENT_INST_ | RE_ID | MSC | FEE1 | PLAN_NBI | OFFER_NE | PART_ID | OCS_RE_II |
| 2 | 925519326 | 9/6/2012 22:12 | 25405 | 15246 | 251925519326 | 251922297401 | 1129349763 | 241928775 | 2461 | 251911299744 | 15246 | 2.42E+08 | 4 | 7 | 1008 |
| 3 | 924872420 | 9/9/2012 0:43 | 22560 | 13536 | 251924872420 | 251920075316 | 1131682818 | 264710136 | 2461 | 251911299744 | 13536 | 2.65E+08 | 8 | 9 | 1008 |
| 4 | 931720484 | 9/1/2012 23:22 | 20992 | 12600 | 251931720484 | 251928699682 | 1123837434 | 3597652653 | 2461 | 251911299744 | 12600 | 3.6E+09 | 2 | 2 | 1008 |
| 5 | 924872410 | 9/29/2012 23:04 | 20990 | 12594 | 251924872410 | 251923138852 | 1155992121 | 2327490612 | 2461 | 251911299744 | 12594 | 2.33E+09 | 2 | 30 | 1008 |
| 6 | 924872417 | 9/12/2012 1:20 | 17124 | 10278 | 251924872417 | 251921498870 | 1135716244 | 2110206795 | 2461 | 251911299744 | 10278 | 2.11E+09 | 7 | 12 | 1008 |
| 7 | 911407923 | 9/23/2012 22:03 | 16218 | 9732 | 251911407923 | 251926872125 | 1149250240 | 3520750365 | 2461 | 251911299706 | 9732 | 3.52E+09 | 5 | 24 | 1008 |
| 8 | 924872466 | 9/12/2012 21:36 | 15600 | 9360 | 251924872466 | 251921020694 | 1136824133 | 729513046 | 2461 | 251911299742 | 9360 | 7.3E+08 | 9 | 13 | 1008 |
| 9 | 913293602 | 9/4/2012 22:42 | 13620 | 8172 | 251913293602 | 251910516922 | 1127131465 | 3797748714 | 2461 | 251911299703 | 8172 | 3.8E+09 | 3 | 5 | 1008 |
| 10 | 926126111 | 9/12/2012 21:52 | 13126 | 7878 | 251926126111 | 251931319102 | 1136822761 | 729492530 | 2461 | 251911299742 | 7878 | 7.29E+08 | 9 | 13 | 1009 |
| 11 | 924154641 | 9/9/2012 23:56 | 12960 | 7776 | 251924154641 | 251925828075 | 1132918238 | 467851230 | 2461 | 251911299744 | 7776 | 4.68E+08 | 6 | 10 | 1008 |
| 12 | 926980702 | 9/15/2012 11:42 | 12618 | 17668 | 251926980702 | 251932289035 | 1139580906 | 257303685 | 2461 | 251911299744 | 17668 | 2.57E+08 | 15 | 15 | 1008 |
| 13 | 913293602 | 9/15/2012 23:15 | 12582 | 7554 | 251913293602 | 251910516922 | 1140118565 | 4049126103 | 2461 | 251911299703 | 7554 | 4.05E+09 | 1 | 16 | 1008 |
| 14 | 932100338 | 9/26/2012 9:56 | 12561 | 17598 | 251932100338 | 251913271394 | 1151951899 | 2524937901 | 2461 | 251911299706 | 17598 | 2.52E+09 | 7 | 26 | 1008 |
| 15 | 917827851 | 9/9/2012 22:40 | 12472 | 7488 | 251917827851 | 251917631252 | 1132913200 | 4146000827 | 2461 | 251911299740 | 7488 | 4.15E+09 | 4 | 10 | 1008 |
| 16 | 911798632 | 9/8/2012 23:21 | 12110 | 7266 | 251911798632 | 251927549148 | 1131636037 | 506031032 | 2461 | 251911299740 | 7266 | 5.06E+08 | 8 | 9 | 1008 |
| 17 | 913293602 | 9/11/2012 21:55 | 11280 | 6768 | 251913293602 | 251910516922 | 1135696518 | 3346464787 | 2461 | 251911299703 | 6768 | 3.35E+09 | 5 | 12 | 1008 |
| 18 | 913293602 | 9/20/2012 22:56 | 11150 | 6690 | 251913293602 | 251910516922 | 1145855822 | 941862149 | 2461 | 251911299703 | 6690 | 9.42E+08 | 9 | 21 | 1008 |
| 19 | 924144799 | 9/30/2012 16:16 | 10844 | 6510 | 251924144799 | 251912035777 | 1156784480 | 1536092029 | 2461 | 251911299744 | 6510 | 1.54E+09 | 12 | 30 | 1008 |
| 20 | 911844894 | 9/8/2012 21:42 | 10787 | 6474 | 251911844894 | 251914274622 | 1131625865 | 2264235333 | 2461 | 251911299706 | 6474 | 2.26E+09 | 7 | 9 | 1009 |

*Appendix 2: Sample from the training data set*

```
@relation telecom_pre_paid_Fraud
@attribute START_date date "dd/MM/yyyy"
@attribute duration numeric
@attribute charge numeric
@attribute calling_number numeric
@attribute called_number numeric
@attribute FEE1 numeric
@attribute Calling&called_number{similar,different}
@attribute charge&duration_status{normal,abnormal}
@attribute payment_status{normal,abnormal}
@attribute CallCost numeric
@attribute CallingTime{PEAK,OFFPEAK}
@attribute CallingDay{NSunday,Sunday}
@attribute CallType {local,international}
@attribute calling_zone{zone1,zone2,local}
@attribute Fraud_NonFraud{fraud,notFraud}
@data

06/09/2012,25405,15246,251922297401,251922297401,15246,similar,normal,normal,0.360070852,OFFPEAK,NSunday,local,local,fraud
09/09/2012,22560,13536,251920075316,251920075316,13536,similar,normal,normal,0.36,OFFPEAK,Sunday,local,local,fraud
01/09/2012,20992,12600,251928699682,251928699682,12600,similar,normal,normal,0.360137195,OFFPEAK,NSunday,local,local,fraud
29/09/2012,20990,12594,251923138852,251923138852,12594,similar,normal,normal,0.36,OFFPEAK,NSunday,local,local,fraud
12/09/2012,17124,0,251924872417,251921498870,-10278,different,abnormal,abnormal,0,OFFPEAK,NSunday,local,local,fraud
23/09/2012,16218,0,251911407923,251926872125,-9732,different,abnormal,abnormal,0,OFFPEAK,Sunday,local,local,fraud
12/09/2012,15600,0,251924872466,251921020694,-9360,different,abnormal,abnormal,0,OFFPEAK,NSunday,local,local,fraud
04/09/2012,13620,0,251913293602,251910516922,-8172,different,abnormal,abnormal,0,OFFPEAK,NSunday,local,local,fraud
12/09/2012,13126,0,251926126111,251931319102,-7878,different,abnormal,abnormal,0,OFFPEAK,NSunday,local,local,fraud
09/09/2012,12960,7776,251924154641,251925828075,777,different,normal,abnormal,0.36,OFFPEAK,Sunday,local,local,fraud
15/09/2012,12618,17668,251926980702,251932289035,1766,different,normal,abnormal,0.840133143,PEAK,NSunday,local,local,fraud
15/09/2012,12582,7554,251913293602,251910516922,755,different,normal,abnormal,0.360228898,OFFPEAK,NSunday,local,local,fraud
26/09/2012,12561,17598,251932100338,251913271394,1759,different,normal,abnormal,0.840601863,PEAK,NSunday,local,local,fraud
09/09/2012,12472,7488,251917827851,251917631252,748,different,normal,abnormal,0.360230917,OFFPEAK,Sunday,local,local,fraud
08/09/2012,12110,7266,251911798632,251927549148,7266,different,normal,normal,0.2,OFFPEAK,NSunday,local,local,fraud
11/09/2012,11280,6768,251913293602,251910516922,6768,different,normal,normal,0.22,OFFPEAK,NSunday,local,local,fraud
20/09/2012,11150,6600,251912202602,251910516922,6600,different,normal,normal,0.2,OFFPEAK,NSunday,local,local,fraud
```

*Appendix 3: Sample from the testing data set*

```
@relation telecom_pre_paid_?
@attribute START_date date "dd/MM/yyyy"
@attribute duration numeric
@attribute charge numeric
@attribute calling_number numeric
@attribute called_number numeric
@attribute FEE1 numeric
@attribute Calling&called_number{similar,different}
@attribute charge&duration_status{normal,abnormal}
@attribute payment_status{normal,abnormal}
@attribute CallCost numeric
@attribute CallingTime{PEAK,OFFPEAK}
@attribute CallingDay{NSunday,Sunday}
@attribute CallType {local,international}
@attribute calling_zone{zone1,zone2,local}
@attribute Fraud_NonFraud{fraud,notFraud}
@data

22/09/2012,2042,2870,251932515540,251911711917,2870,different,normal,normal,0.843290891,PEAK,NSunday,local,local,?
08/09/2012,2042,1230,251932100263,251912978696,1230,different,normal,normal,0.361410382,OFFPEAK,NSunday,local,local,?
06/09/2012,2042,2870,251925727746,251911396141,2870,different,normal,normal,0.843290891,PEAK,NSunday,local,local,?
19/09/2012,2042,2870,251931719538,251910310030,2870,different,normal,normal,0.843290891,PEAK,NSunday,local,local,?
12/09/2012,2042,2870,251925739551,251911114787,2870,different,normal,normal,0.843290891,PEAK,NSunday,local,local,?
11/09/2012,2042,1230,251932100260,251911108618,1230,different,normal,normal,0.361410382,OFFPEAK,NSunday,local,local,?
12/09/2012,2042,2870,251932100224,251911160364,2870,different,normal,normal,0.843290891,PEAK,NSunday,local,local,?
06/09/2012,2042,1230,251932100270,251911400178,1230,different,normal,normal,0.361410382,OFFPEAK,NSunday,local,local,?
01/09/2012,2042,2870,251926193372,251910045483,2870,different,normal,normal,0.843290891,PEAK,NSunday,local,local,?
06/09/2012,2042,1230,251932100248,251911108618,1230,different,normal,normal,0.361410382,OFFPEAK,NSunday,local,local,?
13/09/2012,2042,2870,251932100248,251911340595,2870,different,normal,normal,0.843290891,PEAK,NSunday,local,local,?
08/09/2012,2042,2870,251932100036,251911621226,2870,different,normal,normal,0.843290891,PEAK,NSunday,local,local,?
11/09/2012,2042,1230,251932100395,251910096545,1230,different,normal,normal,0.361410382,OFFPEAK,NSunday,local,local,?
14/09/2012,2042,2870,251925563151,251910863180,2870,different,normal,normal,0.843290891,PEAK,NSunday,local,local,?
05/09/2012,2042,1230,251932100271,251911432126,1230,different,normal,normal,0.361410382,OFFPEAK,NSunday,local,local,?
08/09/2012,2042,1230,251932100271,251922051696,1230,different,normal,normal,0.361410382,OFFPEAK,NSunday,local,local,?
07/09/2012,2042,2870,251932100210,251911024112,2870,different,normal,normal,0.843290891,PEAK,NSunday,local,local,?
```

\

*Appendix 4: Initial list of original and derived attributes with their description*

| No | Field name | Description | Attributes remark |
|---|---|---|---|
| 1 | BILLING_NBR | Billing number (Mobile number to be charged) | Original |
| 2 | START_TIME | Time of call initiation (calling time) or The time when the call is begins | Original |
| 3 | DURATION | Total calling time or the duration of call in second(Call duration in seconds) | Original |
| 4 | CHARGE | Amount of birr paid ( to be paid for the case of post-paid mobiles) | Original |
| 5 | CALLING_NBR | Mobile number initiating or originating the call | Original |
| 6 | CALLED_NBR | Mobile number receiving the call | Original |
| 7 | CELL_A | Mobile BTS cell sector A number(BTS-ID) where the call is originated | Original |
| 8 | FILE_ID | File number used for cross checking in case bill complain. This file ID is given by the department sending the CDR file. | Original |
| 9 | RECORD_SEQ | Record sequence number give when the call is made | Original |
| 10 | FEE | Amount paid or to be paid | Original |
| 11 | OCS_RE_ID | Online Charge System recharge ID number | Original |
| 12 | Calling & called number | Cheek the validity of the calling and called number | Derived |
| 13 | Charge & duration status | The balance between the duration of call and the amount birr paid if it is normal situation | Derived |
| 14 | payment status | To cheek whether or not pay per call in normal situation. | Derived |
| 15 | Call Cost | Calculate current cost per call | Derived |
| 16 | Calling Time | To cheek whether or the calling time is on peak or off-peak hours | Derived |
| 17 | Calling Day | To cheek the off-peak hours on Sunday or night | Derived |
| 18 | Call Type | To identify the call is local or international | Derived |
| 19 | calling zone | Identify the calling is to/from witch zone. According to Ethio-telecom there are three different tariff zones. Zone 1 (Asia, Europe, Middle East and North America) and Zone 2 (Africa, Oceania and South America) or local call. | Derived |

*Appendix 5: The snapshot running information of J48 with 10-fold validation technique*

```
Scheme:     weka.classifiers.trees.J48 -C 0.25 -M 2
Instances:  21367
Test mode:  10-fold cross-validation
 Classifier model (full training set) ==
J48 pruned tree
Number of Leaves:       12
Size of the tree:       22
Time taken to build model: 0.92 seconds
Correctly Classified Instances     21347          99.9064 %
Incorrectly Classified Instances      20           0.0936 %
Detailed Accuracy By Class
TP Rate FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
 0.997   0.000   0.998     0.997   0.997      0.997   0.999     0.997     fraud
 1.000   0.003   0.999     1.000   0.999      0.997   0.999     1.000     notFraud
0.999    0.003   0.999     0.999   0.999      0.997   0.999     0.999     Weighted Avg.
== Confusion Matrix ==
   a      b   <-- classified as

 3843   13 |    a = fraud

    7  17504 |    b = notFraud
```

*Appendix 6: The snapshot running information of PART with 10-fold validation technique*

```
Scheme:     weak. Classifiers. Rules. PART -M 2 -C 0.25 -Q 1
Instances:  21367
Attributes:  11
Test mode:  10-fold cross-validation
Classifier model (full training set)
PART decision list
Number of Rules  :    9
Time taken to build model: 1.08 seconds
Correctly Classified Instances     21350          99.9204 %
Detailed Accuracy By Class
TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
0.997    0.000    0.999     0.997   0.998      0.997   0.999     0.996     fraud
1.000    0.003    0.999     1.000   1.000      0.997   0.999     1.000     notFraud
0.999    0.003    0.999     0.999   0.999      0.997   0.999     0.999     Weighted Avg.

== Confusion Matrix ==
   a     b   <-- classified as
 3844    12 |    a = fraud
    5 17506 |    b = notFraud
```

*Appendix 7: The snapshot of RF with 10-fold validation technique*

Scheme: weka. classifiers. trees. Random Forest -I 10 -K 0 -S 1 -num-slots 1
Relation: telecom_pre_paid_Fraud-weka.filters.unsupervised.attribute.Remove-R1,4-6
Instances: 21367
Attributes: 11
Test mode: 10-fold cross-validation
Classifier model (full training set)
Random forest of 10 trees, each constructed while considering 4 random features.
Out of bag error: 0.0096
Time taken to build model: 1.7 seconds

| | | |
|---|---|---|
| Correctly Classified Instances | 21351 | 99.9251 % |
| Incorrectly Classified Instances | 16 | 0.0749 % |
| Total Number of Instances | 21367 | |

| TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|
| 0.998 | 0.001 | 0.997 | 0.998 | 0.998 | 0.997 | 1.000 | 0.999 | fraud |
| 0.999 | 0.002 | 1.000 | 0.999 | 1.000 | 0.997 | 1.000 | 1.000 | notFraud |
| 0.999 | 0.001 | 0.999 | 0.999 | 0.999 | 0.997 | 1.000 | 1.000 | Weighted Avg. |

Confusion Matrix

| a | b | <-- classified as |
|---|---|---|
| 3850 | 6 | a = fraud |
| 10 | 17501 | b = notFraud |

*Appendix 8: Sample of the normalized data for Multilayer perceptron algorithm.*

```
06/09/2012,25405,15246,251922297401,251922297401,15246,1,1,1,0.360070852,2,1,1,1,1
09/09/2012,22560,13536,251920075316,251920075316,13536,1,1,1,0.36,2,2,1,1,1
01/09/2012,20992,12600,251928699682,251928699682,12600,1,1,1,0.360137195,2,1,1,1,1
29/09/2012,20990,12594,251923138852,251923138852,12594,1,1,1,0.36,2,1,1,1,1
12/09/2012,17124,0,251924872417,251921498870,-10278,2,2,2,0,2,1,1,1,1
23/09/2012,16218,0,251911407923,251926872125,-9732,2,2,2,0,2,2,1,1,1
12/09/2012,15600,0,251924872466,251921020694,-9360,2,2,2,0,2,1,1,1,1
04/09/2012,13620,0,251913293602,251910516922,-8172,2,2,2,0,2,1,1,1,1
12/09/2012,13126,0,251926126111,251931319102,-7878,2,2,2,0,2,1,1,1,1
09/09/2012,12960,7776,251924154641,251925828075,777,2,1,2,0.36,2,2,1,1,1
15/09/2012,12618,17668,251926980702,251932289035,1766,2,1,2,0.840133143,1,1,1,1,1
15/09/2012,12582,7554,251913293602,251910516922,755,2,1,2,0.360228898,2,1,1,1,1
26/09/2012,12561,17598,251932100338,251913271394,1759,2,1,2,0.840601863,1,1,1,1,1
09/09/2012,12472,7488,251917827851,251917631252,748,2,1,2,0.360230917,2,2,1,1,1
08/09/2012,12110,7266,251911798632,251927549148,7266,2,1,1,0.2,2,1,1,1,1
11/09/2012,11280,6768,251913293602,251910516922,6768,2,1,1,0.22,2,1,1,1,1
20/09/2012,11150,6690,251913293602,251910516922,6690,2,1,1,0.3,2,1,1,1,1
30/09/2012,10844,6510,251924144799,251912035777,6510,2,1,1,0.33,1,2,1,1,1
08/09/2012,10787,6474,251911844894,251914274622,6474,2,1,1,0,2,1,1,1,1
01/09/2012,10783,6474,251913293602,251910516922,6474,2,1,1,0.360233701,2,1,1,1,2
16/09/2012,10617,6372,251924872445,251913980898,6372,2,1,1,0.360101724,1,2,1,1,2
17/09/2012,10313,6192,251911997393,251920325300,6192,2,1,1,0.360244352,2,1,1,1,2
29/09/2012,10200,6120,251912263500,251911172884,6120,2,1,1,0.36,2,1,1,1,2
16/09/2012,9780,5868,251911673216,251922435115,5868,2,1,1,0.3,2,2,1,1,1
27/09/2012,9574,5748,251911798632,251927549148,5748,2,1,1,0.360225611,2,1,1,1,2
17/09/2012,9529,5718,251911407923,251926872125,5718,2,1,1,0.360037779,2,1,1,1,2
21/09/2012,9417,13188,251931720484,251923295499,13188,2,1,1,0.7,1,1,1,1,1
02/09/2012,9395,5640,251911997393,251926351744,5640,2,1,1,0.3,2,2,1,1,1
07/09/2012,9384,5634,251913293602,251910516922,5634,2,1,1,0.360230179,2,1,1,1,2
19/09/2012,9288,5574,251913293602,251910516922,5574,2,1,1,0.360077519,2,1,1,1,2
02/09/2012,9240,5544,251913303234,251928325166,5544,2,1,1,0.35,2,2,1,1,1
08/09/2012,9223,5538,251911756833,251922083404,5538,2,1,1,0.36027323,2,1,1,1,2
18/09/2012,9190,12866,251927818435,251923054645,12866,2,1,1,0.7,1,1,1,1,1
01/09/2012,9156,5496,251931720444,251926296563,5496,2,1,1,0.360157274,2,1,1,1,2
23/09/2012,8809,5286,251911844894,251914274622,5286,2,1,1,0.360040867,2,2,1,1,2
```

*Appendix 9: The snapshot of ANN with 10-fold validation technique*

```
Scheme:       weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V
0 -S 0 -E 20 -H a
Instances:    21367
Attributes:   11
Test mode:    10-fold cross-validation
Classifier model (full training set)
Time taken to build model: 107.69 seconds
Correctly Classified Instances       21270               99.546  %
Incorrectly Classified Instances        97                0.454  %
Total Number of Instances            21367
Detailed Accuracy By Class
TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
0.976    0.000    0.999      0.976   0.987      0.985    0.998     0.997     1
1.000    0.024    0.995      1.000   0.997      0.985    0.998     0.997     2
0.995    0.020    0.995      0.995   0.995      0.985    0.998     0.997     Weighted Avg.

=== Confusion Matrix ===
    a     b   <-- classified as
 3763    93 |     a = 1
    4 17507 |     b = 2
```

*Appendix 10: Discussion points with domain experts*

In order to have detail understanding and knowledge about the problem domain, the researcher discussed based on the following discussion points with the experts concerning the telecom fraud.

1. How do you explain fraud in telecom sector?
2. Who and why people commit telecommunication fraud?
3. Which attributes are fraud indicators from the fields of CDR data?
4. How can we identify and prevent telecommunication fraud?
5. What is the current practice to prevent telecommunication fraud?
6. What are your responses on the predictive of model of this research?

## *Appendix 11: Prevailing rule*

Rule 1: If the calling number is equal to the called number then it automatically considered as fraud otherwise it goes to farther investigation.

Rule 2: If duration of the call greater than zero and the charge equals zero then considered as abnormal relation between duration and charge otherwise it goes to farther investigation.

Rule 3: If the charge to be paid and greater than the paid then considered as abnormal payments status otherwise it goes to farther investigation.

Rule 4: let x is charge divided by 100 and y is duration divided by 60.

If x divided by y is below 0.35 cents then it automatically considered as fraud. Otherwise it goes to the next rule.

Rule 5: let the call is local and off-peak, if the cost per call less than 0.35 cents it automatically considered as fraud. Otherwise it goes to the next rule.

Rule 6: let the call is local and peak hours, if the cost per call less than 0.83cents it automatically considered as fraud. Otherwise it goes to the farther investigation.

Rule 7: let the call is international and zone 1, if the cost per call less than 8.63 birr it automatically considered as fraud. Otherwise it goes to the farther investigation.

Rule 8: let the call is international and zone 2, if the cost per call less than 10.29 birr it automatically considered as fraud. Otherwise it goes to the farther investigation.

## DECLARATION

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

Signature ---------- Date ----------

Tesfay Haddish

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

Signature ---------- Date ----------

Workshet Lamenew