ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

PREDICTIVE MODELING FOR FRAUD DETECTION IN TELECOMMUNICATIONS: THE CASE OF ETHIO TELECOM

YESHINEGUS GETANEH

June 2013

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

PREDICTIVE MODELING FOR FRAUD DETECTION IN TELECOMMUNICATIONS: THE CASE OF ETHIO TELECOM

YESHINEGUS GETANEH

June, 2013

Addis Ababa

Ethiopia

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

Predictive Modeling for Fraud Detection in Telecommunications: The Case of ethio telecom

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

By

Yeshinegus Getaneh

June 2013

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

Predictive Modeling for Fraud Detection in Telecommunications: The Case of ethio telecom

By

Yeshinegus Getaneh

June 2013

Name and signature of Member of the Examining Board

Name	Title	Signature	Date
_____	Chairperson	_____	_____
_____	Advisor	_____	_____
_____	Examiner	_____	_____

Acknowledgment

Above all, my gratitude goes to the almighty GOD who is in control of the existing and the coming world. He has been with me in all bad and good times and will always be.

I am also thankful to Dr. Dereje Teferi (PhD) who helped me to reach this level and provided all the support I needed in all situations. He also gave me the chance to use all my effort and welcomed my request for advice, given his tight schedule. Without his support this research would not have been successful.

I am also indebted to my friend and instructor Tibebe Beshah for his patience about my suddenly triggering questions about this paper during tea time. While, he is busy in his PhD paper and many other things, he was more than happy to help me. His support and follow up was one of the reasons for the success of this research.

My beloved wife Adanech Haile and son Nati boy, I love you so much. I can't wait to give you the time you deserve as a family. I am grateful for your patience when I was busy and not able to give the time and attention you deserve. My special thanks also goes to my mom Wro. Zewdnesh Eshete and my brother Seyoum Getaneh. Though, she is not educated, her belief in education and the price she paid for all her children is the reason for my success today.

Encouragement, advice, valuable comments and unfailing supports from friends in and outside office were another reason for my success. I am also indebted to Wro. Saba W/Mariam, Ato Getenet Tesfaye, Ato Derese Agonafir, Ato Addis Ashagrie, Wro. Zelalem Worku, Ato Markos Alemu, Ato Wondwesen Demsie, Ato Bekalu Mamo, Ato Birhanu Zebrie, Ato Naod Tedla, people from ethio telecom, classmates and others. I can't finish mentioning what you did for me with one page but at least acknowledge your names, with many thanks.

Dedication

This research work is dedicated to my mother Wro. Zewdnesh Eshete, my beloved wife Wro. Adanech Haile and my son Natan Yeshinegsu.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Appendixes

Acronyms

BSC: Base Station Controller

BTS: Base Transceiver Station

CCB: Customer Care and Billing

CDR: Call Detail Record

CDMA: Code Division Multiple Access

CEO: Chief Executive Officer

CTIT: College of Telecommunications and Information Technology

GPRS: General Packet Radio Service

GSM: Global Stations for Mobile communications

HLR: Home Location Register

IMEI: International Mobile Equipment Identity number

IN: Intelligent Network

MSC: Mobile Switching Center

OCS: Online Charging System

SIM: Subscriber Identity Module

SMS: Short Message Service

UNMS: Unified Network Management System

VLR: Variable Location Register

Abstract

Telecom fraud is a major concern for telecom operators as well as for governments all over the world. This is mainly because of security threats and economic impacts. These facts can be substantiated by the rules and regulations put in place by different countries.

In this study an effort has been made to predict fraudulent calls made using SIM-boxes to terminate international calls. Such frauds greatly affect the revenue of telephone operators.

Classification methods of data mining are applied using J48, PART and multilayer perceptron algorithms on data collected from ethio telecom. WEKA data mining tool has been used to come up with a model for predicting fraudulent activities. For this study pre-paid sampled voice CDR data has been used along with SMS, GPRS and other data such as pre-paid wallet recharge log from OCS and CCB data warehouse in ethio-telecom.

The experimentation result showed that the model from the PART algorithm exhibited 100% accuracy level followed by J48 algorithm with 99.98%. The rules generated from PART and J48 algorithms enable telecom operators in general and ethio telecom in particular to locate the whereabouts of SIM-boxes as well as other critical information. Moreover, an effort has been made to show the impact of SIM-boxes on telecom operators' revenue.

Chapter One

1. Background

Ethio telecom is government owned and the sole telecom operator in Ethiopia. The name ethio telecom is coined in 2010 after France telecom took the management of Ethiopian Telecommunication Corporation due to government transformation plan. The introduction of telecommunications in Ethiopia dated back to 1894. The operator has passed through different names (brand names) and logos, by different governments that came in power, since the beginning. Fixed telephone (both wired and wireless), Internet (dialup and broadband), mobile (pre-paid and post-paid), CDMA (voice, internet and data), and other value-added services are among the major telecom services provided by the corporation (Yigzaw, Hill, Banser, & Lessa, 2010).

Mobile service in Ethiopia has existed since 1999 and at that time the network coverage was limited to Addis Ababa with a network capacity of not more than 60, 000 subscribers. The company placed mobile service division in its structure beginning form 1996. After the launch of mobile service in Addis Ababa in April 1999 network expansion was a must not only because of the demand from the subscribers but also due to government policy that is in place in the country (Gebremeskal, 2006).

Telecommunication fraud occurs whenever a person committing the fraud uses deception to receive telephony services free of charge or at a reduced rate. It is a worldwide problem with substantial annual revenue losses for many companies. Globally, telecommunications fraud is estimated at about 55 billion US dollars. In the United States of America, telecommunication frauds have a share of 2% of network operators' revenue. However, it is difficult to provide precise estimates since some fraud may never be detected, and the operators are not willing to reveal figures on fraud losses. Sometimes they may not have the evidence and the technique to stop the fraud but they have only the

information from different sources. The situation can significantly be worse for mobile operators in Africa for, as a result of fraud, they become liable for large hard currency payments to foreign network operators. Thus, telecommunication fraud is a significant problem which needs to be addressed, detected and prevented in the strongest possible manner. Popular examples of fraud in the telecommunication industry include subscription fraud, identity theft, voice over the internet protocol (VoIP) fraud, cellular cloning, billing and payment fraud on telecom accounts, prepay and postpaid frauds and PBX fraud(Abidogun, 2005).

Among the revenue sources of ethio telecom, international traffic takes the lion share of it. As Asfaw (2006) indicated 40% of Ethiopian Telecommunications Corporation revenue is from international traffic. SIM box fraud or gateway fraud is one of the fraud types that attack the revenue from international traffic. The researcher understood from domain experts working in international traffic area that the international incoming call termination tariff is currently 0.19 USD per minute. The fraudsters involved in SIM box fraud paying only 0.83 birr (for peak hour) and 0.35 birr (for off-peak hour) for international calls coming through the SIM box. In this regard ethio telecom is losing the difference per minute. In addition, it has negative effect on telecom security and quality of service.

SIM box fraud is affecting not only ethio telecom but also telecom operators in Africa like Ghana. It is a system by which fraudsters re-route international calls by using SIM box device and local SIM cards. It is also one of the reasons for telecom operators for losing millions of dollars every year (Adu-Boafo, 2013).

In this regard, fraud detection helps the company increases revenue by minimizing loss through fraudulent practices in the country. Both security division and revenue assurance section of the company are working to protect and assure the revenue as well as existence of ethio telecom by minimizing

loop-holes for revenue leakage. In this regard this work will have its own contribution for the company as well as for the country in general.

1.1 Statement of the problem

Telecommunications in general produce a large amount of data every minute. Due to the nature and size of the data, it is almost impossible to analyze the data manually. A very large amount of data is generated from different network elements and telecom applications like OCS, HLR/VLR, CDR, UNMS, network alarms that are generated from each network devices, Z-smart customer service application system for sales and collection, Z-smart Trouble Ticket application system, log files of different services that the company provide for customers and many more. All these data are stored or pushed to different servers. In here, we see the power of data mining approaches to extract different useful knowledge, pattern and prediction ability from these data. It is almost impossible to identify or predict fraudulent calls among the calls made every second (Weiss, 2005).

Fraud is costly to a network carrier both in terms of lost income and wasted capacity. The various types of telecommunication fraud can be classified broadly into two categories that are subscription fraud and superimposed fraud. Subscription fraud is a fraud to a service, often with false identity details, with no intention of paying. Cases of bad debt are also included in this category. On the other hand superimposed fraud occurs from using a service without having the necessary authority detected by the appearance of unknown calls on a bill. This fraud includes several ways, for example, mobile phone cloning, ghosting (the technology that tricks the network in order to obtain free calls), insider fraud, tumbling (rolling fake serial numbers are used on cloned handsets so that successive calls are attributed to different legitimate phones), etc (Estévez, Held, & Perez, 2006; Melaku, 2009; Yigzaw et al., 2010). Interconnection bypass through GSM gateways by using SIM boxes to terminate international calls falls under super imposed fraud category.

Among the calls you received, one of them could be international call but the displayed number is a local mobile number. You might have asked yourself as to how this could happen? If you find it as missed call, you might have tried to dial it but it didn't work. This kind of fraud is done using SIM boxes that are used by passing telecom operators normal international route. Such kind of activity leaks the revenue of network carriers that they get by serving international calls. According to a research in Bangladesh, in 2007 the global traffic in international voice calls was nearly 350 billion minutes, accounting for global revenues of USD 78 billion. The high cost of international calls creates arbitrage opportunities for illegal network operators who can provide termination of international voice traffic in domestic network. It is estimated that illegal bypass can account for as much as 30-60% of international call volumes in countries with International Gateway (IGW) monopolies. Such bypass results in reduced voice call quality, security issues and also reduced revenues for the government. International termination rate varies from country to country (like from USD 0.09 to 0.125 per minute) depending on the agreement with telecom operators that have international gateways(Choudhary & Aftab, 2011).

These calls are coming through internet (data network) using broad band connection. The international calls are terminated as local calls using SIM box technology which is used for this purpose. The SIM boxes are located in the country where the calls are terminated. In other words these SIM boxes take or use local SIM cards and as a result ethio telecom will not have a chance to identify these international calls are terminating in its network. In this study, the models enable to show on how to identify or predict such fraudulent numbers that are used for this purpose. It also enables to predict the location of where this fraud activity is taking place or where the SIM box is located. For this purpose, the researcher used HLR data, Call Detail Records, SMS, GPRS and OCS data. By integrating the results obtained by analyzing these data it enables to predict the fraudulent calls. As a result it enables ethio-telecom to

identify revenue leakage spots and take appropriate action to maximize its profit.

SIM box fraud is very difficult to detect for telecom operators as it is coming through the internet and appears like local call. The calls are international and telecom operators' loss some amount of money, the difference of 0.19 and 0.0466 USD per minute, by not identifying these calls. In addition, this kind of fraud is a recent phenomenon that is challenging ethio telecom in particular and Africa in general.

Identifying the location of SIM boxes is a challenging task for telecom operators. Because SIM box fraud doesn't need an office or license from government like ethio telecom or telecommunications agency. You simply buy the device, connect it to the internet, insert the SIM cards and let it work. You can put it anywhere like in drawer or celling where there is network. This makes it for telecom operators like ethio telecom difficult in identifying the where about of the SIM boxes.

Moreover, ethio telecom or other operators can't identify the mobile numbers used for SIM box purpose manually or by simple observation. These mobile numbers are operating like the non-fraudulent mobiles. Mobile users also have no way of identifying these numbers till they hear the voice of their friend or family living in the other part of the world.

Currently, not only ethio telecom but also other telecom operators in Africa are suffering such kind of fraud for different reasons due to the rising profile of the continent. This SIM box fraud is practiced in order to avoid the high cost of international call termination tariff (Adu-Boafo, 2013).

Ethio telecom collects millions of dollars each month, for instance in March, 2012 it served 67,969,403 minutes of incoming call. This figure will amount to 12,914,195 USD, calculated using the current call termination rate of 0.19 USD per minute (Ethio-telecom, 2012a). Summary of incoming international

calls terminated from 2003 to 2012 is presented in table 1.1 to show the trend and effect of international incoming calls. As Abidogun (2005) indicated the fraud lost by operators is indicated to be 2% of their revenue and if we calculate it only form the international traffic, it is estimated to be 258,283.9 USD for the month of March. As we can understand it is a large amount of hard currency that ethio telecom is losing. According to Asfaw (2006) about 40% share of revenue for ETC, now ethio telecom, is obtained from international traffic. We can also consider the impact for the country in terms of hard currency needed for expanding infrastructure for the development or betterment of the country.

Years	Incoming traffic (min)
2003	87,858,482
2004	171,906,120
2005	227,847,117
2006	274,396,612
2007	355,760,645
2008	400,393,029
2009	481,637,558
2010	541,382,467
2011	669,461,088
2012	795,863,707

Table 1.1 Summary of International Incoming Calls from 2003 to 2012 (source: Internal Reports)

By taking the CDR data along with location and OCS data, data mining can be applied to see the pattern of fraudulent calls. A predictive model can also be derived by applying different data mining techniques and algorithms.

This is the point that data mining comes into play where the different tools, techniques and algorithms are applied to solve the problem of ethio telecom in relation to SIM box fraud. By applying data mining for this specific fraudulent problem, it is possible to give due solution. Different data mining techniques are checked for optimal solution at this point. The data used for this research is more of quantitative and bulky by its very nature. It is now at the mercy of

data mining. Data mining provides an optimal solution for such kind of problems and data types.

1.2 Scope and limitation of the study

Scope

There are many fraudulent activities and large amount of data in telecom industries that waits for the mercy of data mining to be of use for the company's advantage and to maximize ethio-telecom's benefits.

This study limits itself to illegal incoming international telephone calls that are coming using SIM boxes. International calls are made via Voice Over-IP (VOIP) telephone through internet. These calls are made by diverting the route of proper international call routes. This study only analyzes HLR, OCS and CDR data that help to predict these fraudulent calls.

The fraud detection is limited to pre-paid mobile customers only. Similar frauds could be found on post-paid mobiles as well which could be conducted by fellow researchers.

Limitation

The researcher initially intended to get CDR data from switch and to identify the SIM box devices using IMEI number of the device. Unfortunately, it was not possible to get the data from the switch (MSC) for security reason but the CDR from CCB database is used for this study. In addition, the contribution of domain experts was very helpful and vital. It would have been impossible to finish this research without their help and support. But the participation was limited or with some reservation in order to maintain their business secret.

1.3 Objective of the study

1.3.1 General objective

The major objective of this study is to investigate and come up with a predictive model for detecting incoming international calls that are terminated using local mobile numbers. The derived model can be integrated with the existing system to detect frauds in telecommunication companies, specifically in ethio telecom. This is done by implementing derived models from data mining tools, techniques and algorithms.

1.3.2 Specific objectives

The specific objectives of this research are:

- To explore and understand the domain and fraudulent cases by reviewing literatures
- To select data to predict fraudulent international incoming calls
- To conduct experiments and build predictive models to evaluate and interpret the results.
- To propose the best model that can be implemented or integrated with the existing system.
- Report the result of the study and recommend future research works

1.4 Justification of the Study

There are few researches that are made on Ethiopian Telecommunications Corporation (ETC), currently ethio telecom, both in Addis Ababa University and CTIT. The trend of telecom fraud is increasing from time to time in Ethiopia and in Africa in general. Data mining can help in solving this problem by using the data in the hand of telecom operators. Moreover, many researches need to be done in telecom sector not only on mobile service but also in other services

like data, internet, AAA (Authentication, Authentication and Accounting) and others.

The telecom fraud is increasing and wide-spreading from time to time. In addition, SIM box fraud is affecting mainly the revenue from international traffic, which is largest or 40% share of Ethiopian telecom revenue. As you can see from table 1.1, the number of calls (minutes) from international calls is increasing from time to time. This also implies that the loss from this sector is increasing. For instance, in 2010 the figure was a little more than 540 million minutes but in 2012 it reached 795 million (Ethio-telecom, 2012a; Negarit, 2012). The amount of loss is clearly seen if 2% loss is calculated from international incoming calls only using the current TAR or international call termination rate 0.19 USD per minute. It is about 2.05 and 3.02 million USD from 2010 and 2012 respectively. This figure is calculated from international incoming calls but the loss has to be calculated from annual revenue. That makes the figure much bigger and the case more critical since the profit of ethio telecom is in hundreds of billions of birr.

According to the Negarit Gazeta (2012), telecom fraud is a burning problem, in addition to the increase and wide-spread of telecom frauds, it is a serious threat for national security beyond economic loss.

Few researches are made on Ethiopian telecom regarding fraud but no research is conducted to detect fraudulent international calls that are using SIM boxes. In addition, the fraudulent activity using SIM boxes is a newly emerging one and this makes this research new and it also fills the gap that the previous researchers did not address.

On the other hand, this research tries to identify mobile numbers that are used for this purpose and also approximate locations of SIM boxes. These locations can be identified by using the cell id of the BTS's as well as IMEI (International Mobile Equipment Identity) numbers (Willassen, 2003) and sometimes referred

as International Mobile Service Equipment Identity (Kivi, 2009). In this study, different data sources that have never been used by previous researchers like OCS and HLR data are used. In addition, to protect the privacy of customers' due care is given by providing different sequential numbers or codes for each mobile number by discussing with domain experts at ethio telecom.

This research is different with previously made researches on fraud detection in many aspects. The major difference is in the fraud type that is SIM box fraud. The previous researchers were made on fraud detection in general like Jember (2005) and Gebremeskel (2006). In addition, SIM box fraud is a recent phenomenon. The second difference is on the type and size of data. In this study in addition to voice CDR additional data like SMS, GPRS, OCS and other derived attributes are used based on domain experts' recommendation. The other difference is on data mining tool used. They used MATLAB and Brain Maker data mining tool respectively and neural network algorithm only. But in this research J48 and PART algorithm are used in addition to neural network. Moreover, by the time when the last research was made on fraud detection by Gebermeskal (2006) the number of mobile customers were less than one million but currently it is 18.28 million, as reported by the CEO on performance related press conference (Ethio-telecom, 2012b).

1.5 Significance of the Research

At the end of the study, the outcome of the research enables ethio telecom to detect frauds in relation to international incoming call. And provide relevant information to take necessary action and maximize its revenue, by closing the back doors of revenue loss. It also has positive impact to maximize hard currency revenue for the country.

As indicated in section 1.5, the amount of hard currency that ethio telecom lose is increasing from year to year and the effect of SIM box fraud is not tolerable by telecom operators in general and ethio telecom in particular. Due

to SIM box fraud, millions of dollars are lost from telecom operators every year and ethio telecom is not an exception.

It also indicates and provides the suitable algorithm and model that can predict international incoming telephone frauds with best accuracy. The accuracy level is determined by the type of algorithm and techniques used during the study.

It indicates SIM cards that are used for SIM boxes and locates the approximate location of the SIM boxes. It helps to take appropriate measure to terminate international call via illegal route.

It indicates or approximates the amount of revenue loss by calculating the total call durations of those SIM cards.

This research also fills the gap of previously conducted researches in telecommunication fraud detection.

1.6 The Way the SIM Box Functions

SIM box is a hardware which is used to bypass the legal or normal route for international incoming call. It is used for unauthorized bypassing of lawful billing systems to gain personal advantage (Augustin et al., 2012). According to Adu-Boafo (2013), SIM box fraud is a system to re-route international calls and by inserting local SIM cards in the box to make it appear a like local call. This SIM box fraud is also known by the name SIM gateway fraud and is found in countries where total accounting rate (TAR) or international call termination rate is relatively high. The fraudsters enjoy some portion of the difference between the international termination rate and local tariff. Countries, especially developing countries, in Africa suffer this loss due to high incoming traffic of international call for different reasons (Lokanathan & Samarajiva, 2012).

Whole sellers in other part of the globe route the international call via cheaper route in order to minimize the payment made for international call termination. This route uses mostly the internet and in the destination country the SIM box device is connected to the internet having many SIM cards in it. This device is intelligent enough to identify voice traffic packets and free SIM cards in order to use them to make outgoing calls (CxB-Limited, 2013). In order to identify the SIM box numbers or SIM cards that are used to terminate international calls the following characteristics are advised by different researchers and domain experts. Among the factors charged IMSI (International Mobile Subscriber Identity), First Cell ID, Chargeable Duration, B type of number, non-charged party and charging start time are relevant measures for fraud detection (Burge et al., 1997). The following figure 1.1 shows how the SIM box works and the routes for both legal and illegal ones.

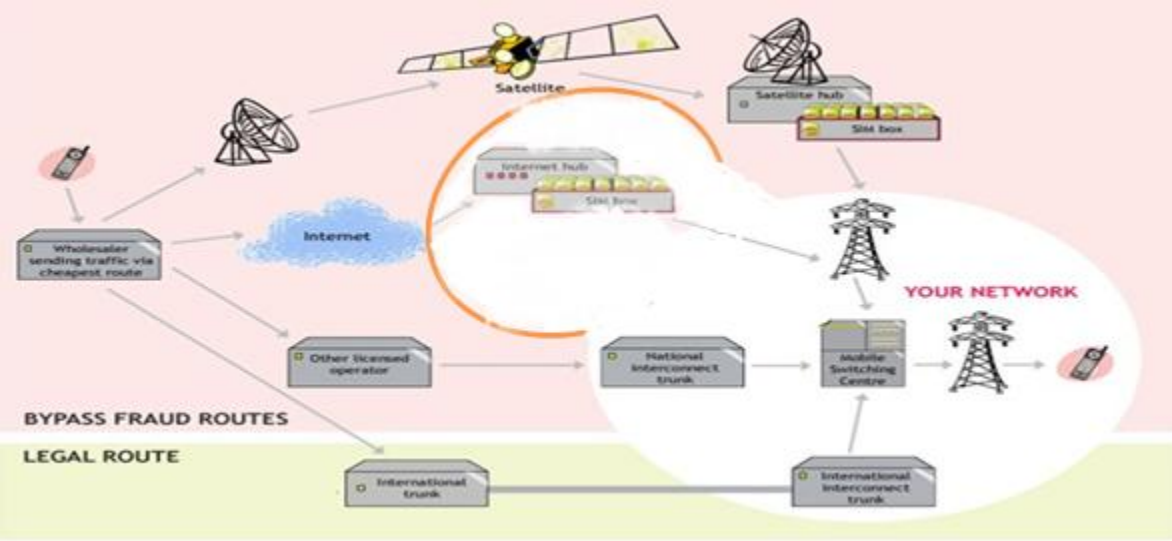


Figure 1.1 How SIM box works **(source: CxB Limited, 2013)**

Current Fraud Handling Practice of the Company

In ethio telecom there is a section called fraud management under security division. The fraud management section receives different letters from different stake holders, requesting to identify list of telephone/mobile numbers. In addition internal sources, employees, also communicate such kind of fraud via

e-mail. By incorporating the section effort, an effort is made to detect SIM box frauds, as domain experts responded to the researcher using questionnaire.

Currently, there is a group organized in quality circle to study about the problem with SIM box fraud. The group members are from fraud management section, NNOC (National Network Operation Center) and other departments. This group sends report for CEO and information and communication minister. Since the trend is increasing from time to time, they are expected to propose a solution.

1.7 Methodology

1.7.1 General Approach

In this study CRISP-DM (Cross Industry Standard Process for Data Mining) model is used and each steps are strictly followed to reach to the desired knowledge. The researcher used classification methods like J48 and PART from decision tree and multilayer perceptron from artificial neural network for prediction purpose. These algorithms are selected based on previously made researches recommendations on fraud detection and to select best performing algorithm (Adu-Boafo, 2013; Augustin et al., 2012). By applying the above algorithms a predictive model that can detect fraudulent international calls using SIM boxes is developed. Depending on the nature of the data and the purpose of the study, the mentioned techniques and algorithms are selected. The data mining tool selected for this study is WEKA that stands for Waikato Environment for Knowledge Analysis. It is developed at the University of Waikato in New Zealand, written in java (object oriented programming language) and tested under different operating systems (Witten & Frank, 2000). In addition applications tools like WEKA, MS Excel, MS Access and MySQL

Database are used for data analysis and experimentation, preparation and storage purposes respectively.

1.7.1.1 Literature Review

Literatures such as books, journals, magazines, internet sources and others are consulted regarding the concept and the researches made on telecom related frauds. Not only these locally made researches but also other papers that have direct or indirect relation to this work are reviewed.

1.7.1.2 Data Collection

In order to get the data from ethio telecom, the researcher got a letter from Addis Ababa University and delivered it to CEO of the company. Finally, this letter was directed to human resource division, security division, IT operation division and other concerned departments and sections. Then CDR, GPRS, SMS and OCS data are collected from ethio telecom, IT operation division. Telecom operators generate a large amount of data every second and the nature of the data is bulky. There is a need to take sample of the data so that the data size could be manageable. As indicated above, the data is collected from different servers in ethio telecom. These servers are the billing server that provides the researcher the CDR and the HLR data that stores location data when a customer makes calls. The GPRS and SMS CDR are stored in a separate table in CCB database. The OCS database stores customers recharge history. By integrating the data from these sources analysis is made to deliver a predictive model that can detect fraudulent calls using mobile numbers to terminate international calls (incoming).

1.7.1.3 Business Understanding, Data Understanding and Preprocessing

Business understanding is the basic starting point for any research that use data mining processes. As the researcher is working in telecom industry and this opportunity gave a chance to understand the business. Experts from IT/IS security department and management/ technical audit department are

communicated to provide their technical advice on the problem to be addressed in this study. Domain experts are consulted in order to understand the data and to know more about the business. Their advice was needed for identifying additional data needed for detecting SIM box frauds. The preprocessing phase helped to clean the data by avoiding empty columns, dealing with missing values, data reduction, data integration and transformation and other activities.

1.7.1.4 Modeling and Experimental Techniques

It is necessary that the modeling of the data mining process is done using CRISP-DM process model. This model is selected because it is an industry standard and telecom is one of them. Additionally, different researches on fraud prediction use this standard (Tariku, 2011). In addition to this descriptive and predictive modeling are used by allowing classification methods. From classification decision tree and neural network are applied in this study. The classification method showed the rules for predicting calls from the data which is fraudulent and which is not (Hornick, Marcadé, & Venkayala, 2007). For this study WEKA data mining tool is used to analyze the data and different algorithms are compared to select the best one.

1.7.1.5 Evaluation Techniques

Evaluation is appropriate and crucial in order to assess the output of the study. The result for this research is evaluated in different ways. The first technique is using the test data that WEKA set aside. In addition, among the alternative techniques that WEKA provides, the model that results in higher accuracy is selected. Additionally, the accuracy levels, the result from the confusion metrics, time taken to build the model and detail accuracy by class are considered to assess the classification result. The other way of evaluating the models is by discussing on the output with domain experts in fraud management section, audit division and others at ethio telecom.

1.8 Organization of the Thesis

This paper is organized in a manner assuming the sequential activities of the study. Following this introductory chapter, the principles and concepts of data mining, data mining methodology and related works are discussed under literature review in chapter two. Then chapter three covered the discussion on the data mining methods that are used in this study. Chapter four is about data preparation, which deals with the data to make it ready for experimentation and analysis. The fifth chapter focuses on experimentation, analysis and modeling. The last chapter covered conclusions and recommendations from this study.

Chapter Two

Review of Literature and Related Works

2.1 Fraud in Telecom Industry

Telecom industries are still being challenged by fraud scenarios. The dynamicity of fraud types and misuse of technological advancement made the operation of telecom industries more challenging. The industry is not simply watching what fraudsters are doing rather doing its best to secure their customers and minimize revenue leakages. This piece of work is intended to help the industry, specifically ethio-telecom, in the effort to protect customers and minimize revenue loss.

This chapter discusses the fraud in telecom industry, related works and data mining concepts. Fraud in telecom industry, sub topic will start by defining fraud, then follows fraud types, the effect of fraud on telecom and fraud detection techniques in telecom industries. The second sub topic discusses data mining concepts and finally related works conclude the chapter.

2.1.1 Definitions

In legislation, the term fraud is used broadly to mean misuse, dishonest intention or improper conduct without implying any legal consequences (Abidogun, 2005).

The term fraud can also be referred to as the abuse of a profit organization's system without necessarily leading to direct legal consequences (Phua, Lee, Smith, & Gayler, 2010).

Fraud covers a wide range of illicit practices and illegal acts that is intentional deception or misrepresentation. It is defined as any illegal act characterized by deceit, concealment, or violation of trust. Frauds are usually committed, by individuals and/or organizations, to secure personal or business advantage

through unlawful act to obtain money, property, services or to avoid payment or loss of services (Tariku, 2011).

The above definitions mainly focus on fraud in general. Strictly speaking the definitions of frauds in telecom sector have to be customized to its context. In this regard, different scholars defined fraud in different perspectives. Fraud is defined as any transmission of voice or data across a telecommunications network where the intent of the user is to avoid or reduce legitimate call charges. It is also defined as obtaining un-billable services and undeserved fees. Additionally, telecommunication fraud occurs whenever a person committing the fraud uses deception to receive telephony services free of charge or at a reduced rate. Fraudsters see themselves as entrepreneurs, admittedly utilizing illegal methods, but motivated and directed by essentially the same issues of cost, marketing, pricing, network design and operations as any legitimate network operator (Abidogun, 2005).

In general telecommunications fraud can be simply described as obtaining telecommunication service with no intention of paying for the service. The major characteristic that makes telecommunications fraud more attractive to fraudsters is that the danger of localization is minimal. This is because all actions are performed from a distance which makes the process of localization time-consuming and expensive. The simple knowledge of an access code, which can be acquired even with methods of social engineering and advancement of technological progress make the implementation of fraud feasible. Finally, it is possible to say that the product of telecommunications fraud can easily be converted to cash (Hilas & Mastorocostas, 2008).

2.1.2 Fraud Types

There are many fraud types that exist in the telecom sector. Different scholars list and categorize the fraud types into different manner. According to (Estévez et al., 2006), there are about six fraud scenarios. These are: subscription fraud,

PABX fraud, free phone call fraud, premium rate fraud, handset theft and roaming fraud. These fraud types are explained with some detail in the referred literature by Estévez, Held, & Perez (2006). Similarly (Hilas & Mastorocostas, 2008) categorized fraud as technical fraud, the contractual fraud, the hacking fraud, and the procedural fraud. There is no distinct figure for fraud types or we can't be exhaustive by enumerating fraud types, due to its dynamic feature and for they can be performed by combining them.

In telecommunication, these frauds are broadly categorized as subscription and super imposed frauds. On top of these, categories ghosting (technology that tracks the network in order to obtain free calls) and insider fraud are labeled under other fraud types. Insider fraud is simply selling of information by telecom employees to fraudsters that can be explained for fraudulent gain (Akhter & Ahamad, 2012). According to (Phua et al., 2010), fraud can also be grouped as internal and external fraud. The internal fraud involves employees of the business which incorporates both management and non-management employees. The external fraudsters can either be prospective/ existing customer (consumer) or supplier. They can also be profiled as average offender, criminal offender and organized crime offender. Except the average offender the other two are risky.

a) Subscription Fraud

In such type of fraud, fraudsters get an account without intention to pay the bill. This is to mean that the cheater accesses the services without being subscribed. Then the account is characterized by abnormal usage and fraudulent calls or transactions. This account could be used for call selling or intensive self-usage (Akhter & Ahamad, 2012).

Subscription fraud is the most common type of fraud encountered on the GSM network. A person subscribes for a service by using false identification. Then the fraudster may use the service either for personal use or for profit making.

In the first scenario, he/she may use it for personal use or he passes the phone to someone else. The second is for real profit. Here the fraudster claims to be a small business to obtain a number of handsets for Direct Call Selling purposes. The fraudster, who has no intention of paying his bill, now sells the airtime, to people wishing to make cheap long distance calls (Shawe-Taylor, Howker, & Burge, 1999).

b) Superimposed Fraud

Unlike subscription fraud, a legitimate account will be used in superimposed fraud. In this case, abnormal usage is superimposed on the normal usage of legitimate customers. Mobile phone cloning and obtaining calling card authorization are among the several ways to carry out this fraud type. Examples of such cases include cellular cloning, calling card theft and cellular handset theft (Akhter & Ahamad, 2012).

Superimposed fraud is the most common fraud scenario in private networks. This is the case of an employee, the fraudster, who uses another employee's authorization code to access outgoing trunks and costly services (Hilas & Mastorocostas, 2008).

In addition, fraudsters make use of the legitimate account for an illegitimate use by different means. In such cases, abnormal usage is observed and it is somewhat challenging to detect(Bella, Olivier, & Eloff, 2005). The fraud type to be addressed in this research is of this type. They subscribe for both internet and mobile service from the service provider, in this case ethio-telecom. There is also a device called SIM box, sometimes called gateway, which takes many SIM cards at a time. The international call (voice traffic) comes through the internet and the gateway forward the call to one of the idle mobile numbers. Finally, it takes that number when the call reaches to the called number.

2.1.3 The Effect of Fraud on Telecom

According to Estévez, et al., (2006) fraud is one of the major revenue leakage sources for telecom industry. Globally, telecommunications lose tens of billions of dollars per year due to fraud. In addition telecom fraud has a negative impact in terms of quality of service, lost income and wasted capacity. These frauds have either direct or indirect loss of money for a service provider. The direct loss is when resources are consumed and the service provider does not receive payment. If a user succeeds in damaging the reputation or market value of the service provider, then we call this indirect loss (Jonsson, Lundin, & Kvarnström, 2000; Kou, Lu, Sirwongwattana, & Huang, 2004).

According to Akhter & Ahamad (2012), in addition to multi billions of dollars of revenue loss, fraud can also affect the credibility and performance of telecom companies. It involves theft of services and deliberate abuse of voice and data networks. The intent of fraudsters is to avoid or at least reduce the charges for using the service. The negative impact of fraud on the telephone company is described in four ways such as financial, marketing, customer relations and shareholders perceptions.

2.1.4 Fraud Detection Techniques in Telecom Industries

Since the 90's different approaches have been used by telecommunications based on statistical analysis and heuristics methods to help them detect and categorize fraud situations. It is recently that they adopted to use and explore data mining and knowledge discovery techniques for this task (Bella et al., 2005).

Among the techniques available for managing and detecting telephone fraud include manual review of data, conventional analysis using rule based expert system and advanced flexible techniques using data mining (advanced data analysis).

In manual review of data, the problem with this technique is the bulkiness of the data that makes almost impossible for a team to filter the fraudulent calls manually. Especially telecom companies will have millions of call detail records generated by their customers for a single month within a specific region. As a result this makes it a time consuming and laborious technique for detecting fraud (Akhter & Ahamad, 2012).

The second technique is conventional analysis using a fixed rule based expert system together with statistical analysis. A rule based system is a set of rules that take into account the normal calling hours, the called destinations as well as the normal duration of the call etc (Akhter & Ahamad, 2012). Rule-based is described as something which is very difficult to manage because of the proper configuration of rules requires precise, laborious, and time consuming programming for each imaginable fraud possibility. The dynamicity of new fraud types requires constantly updating the rules to adapt to the existing, emerging and future fraud options. This will introduce a major obstacle for scalability. There will also be a drastic performance downfall of the system when more data is processed by the system (Kou et al., 2004).

The third technique, according to Akhter & Ahamad (2012), is adaptive flexible techniques using advanced data analysis like artificial neural networks (ANNs). Neural network can quickly learn to pick up patterns of unusual variations that may suggest instances of fraud on a particular account by feeding the raw data (Akhter & Ahamad, 2012). Supervised and unsupervised neural networks are the two main forms. Unsupervised learning neural network is one future system that will reduce the processing load for both rule based system and supervised neural based system (Burge et al., 1997).

2.2 Data Mining

The paragraphs that follow discuss relevant topics in data mining that have direct and indirect relation to this specific research. Due to the scarcity of time and other resources it is found to be reasonable to mention few points on data mining in a precise manner.

The technological progresses in digital data acquisition and storage have resulted in the growth of huge databases. This is true in all areas of human endeavor, from the very ordinary (such as CDR), government statistics, credit card usage and supermarket transaction data) to the more exotic (such as images of astronomical bodies, molecular database, and medical records) (Hand, Mannila, & Smyth, 2001).

The fast growing, tremendous amount of data, collected and stored in large and numerous databases, has far exceeded our human ability for comprehension without powerful tools. Data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research. The widening gap between data and information calls for systematic development of data mining tools that will turn “data tombs”, data archives that are seldom visited, into “golden nuggets” of knowledge (Jiawei & Kamber, 2001).

Bearing in mind that defining a scientific discipline is controversial and accepting that others might disagree about the details, data mining is defined as: “ The analysis of (often large) observational data sets to find unsupervised relationships and to summarize the data in novel ways that are both understandable and useful to the data owner”(Hand et al., 2001). Similarly, Jiawei & Kamber (2001) viewed data mining as a result of the natural revolution of information technology. An evolutionary path has witnessed in database industry in developing functionalities like data collection and

database creation, data management, and data analysis and understanding (involving data warehousing and data mining).

2.2.1 Overview

Data mining is becoming a mainstream technology used in business intelligence applications supporting industries such as financial services, retail, healthcare, telecommunications, and higher education, and lines of business such as marketing, manufacturing, customer experiences, customer service, and sales. Now a day, it is becoming a common practice among business analysts, scientists and researchers to apply data mining on seemingly random data points. Hence data mining is widely used to solve business problems across industries (Hornick et al., 2007).

As mentioned in the previous paragraphs data mining is being applied in different areas. Comparison of data mining and other statistical tools will be crucial to know the real importance and difference of data mining. In addition, the review on data mining technologies such as data mining models, tasks and techniques also help us to recall about the technology. Application of data mining in the telecom industry will narrow down the focus area and help us to pave the way for this research.

Among the giants, telecommunication industry was an early adopter of data mining technology to its needs. Marketing customer profiling, fraud detection, churn management and network fault isolation are few application areas in telecom industry. Given the issues like privacy and legal restrictions in telecom industries, they are using data mining techniques to improve their services and solve their business problems (Umayaparvathi & Iyakutti, 2011).

2.2.2 Data Mining and the KDD Process

Data mining has different definitions according to different authors from different disciplines as well as in the same discipline. The term KDD is widely used to denote the overall process of extracting high-level knowledge from low-level data. Others also use the terms data mining and KDD interchangeably. The multitudes of names used for KDD are many but to mention few: data or information harvesting, data archaeology, functional dependency analysis, knowledge extraction, and data pattern analysis (Sumathi & Sivanandam, 2006).

According to Hornick & et al., (2007), data mining is the process of finding patterns and relationships in data. It consists of developing a model from historical data and applying that model on a new data. It is also defined as the application of specific algorithms for extracting patterns from data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996b). Data mining is sometimes known as "secondary" data analysis because it deals with data that are collected for some purpose other than the data mining analysis, unlike statistics. It also refers to extraction or "mining" knowledge from large amounts of data. The term is actually a misnomer. Different writers are arguing that it should have been more appropriate to call it "knowledge mining from data," or in short "knowledge mining". Knowledge mining from databases, knowledge extraction, data or pattern analysis, data archaeology, and data dredging are some of the terms which have similar or slightly different meanings with data mining (Jiawei & Kamber, 2001).

Data mining can also be seen as a combination of tools, techniques and processes in knowledge discovery. In other words, it uses a variety of tools ranging from classical statistical methods to neural networks and other new techniques originating from machine learning and artificial intelligence in improving database promotion and process optimization (Tesema, Abraham, & Grosan, 2005).

According to Sumathi & Sivanandam, (2006), data mining can be broadly divided into two as verification driven and discovery driven data mining. Verification-driven data mining extracts information in the process of validating a hypothesis postulated by a user. It uses techniques such as statistical and multidimensional analysis. Discovery-driven data mining applies tools such as symbolic and neural clustering, association discovery, and supervised induction to automatically extract information. Data mining applications derived from the above two forms should consider three things to be effective. First, it must have access to organization-wide views of data instead of department specifications. Second, the data mining application must be applied on the data warehouse. Third, it must provide the mined information in a way that support or ease decision making.

Data mining broadest scope is referred as KDD. However, data mining is generally thought of as a particular activity of KDD, a single step in KDD, that applies a specific algorithm to extract patterns that help convert data into knowledge. Data mining has been defined as the process of sifting through large amounts of data to spot patterns and trends that can be used to improve business functions. It combines techniques from statistics, databases, machine learning, and pattern recognition to extract (mine) concepts, concept interrelations, and interesting patterns automatically from large business databases. These techniques explain the multidisciplinary nature of data mining in general (Sumathi & Sivanandam, 2006).

On the other hand knowledge discovery in databases (KDD) is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." Whereas, data mining is one of the steps in the KDD process(Fayyad, Piatetsky-Shapiro, & Smyth, 1996a). Additionally, the term "KDD" is used to refer to the overall process of discovering useful knowledge from data (Fayyad et al., 1996b).

2.2.2.1 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 2.1 below, source: (Fayyad et al., 1996b).

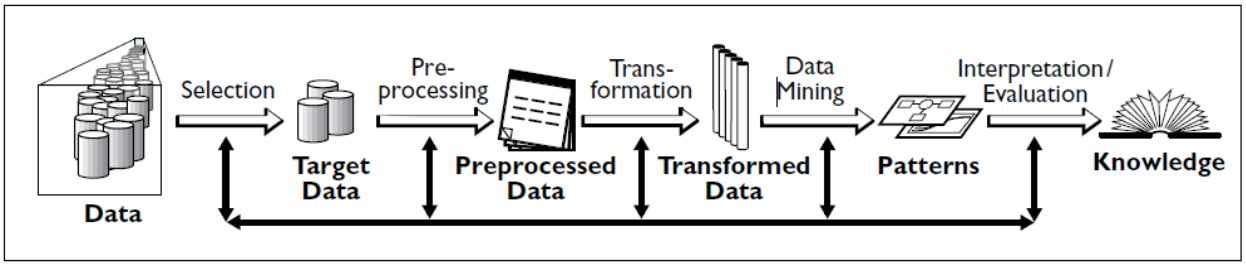


Figure 2. 1 The KDD process

1. **Defining the business problem**

Understanding the data and the business area is crucial and mandatory to knowledge discovery. Algorithms alone will not solve the problem without having clear statement of the objective and understanding.

2. **Creating a target dataset**

This process includes selecting a dataset or focusing on a subset of variables or data samples which are going to be used for discovery.

3. **Data cleaning and pre-processing**

Tasks like 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.

4. **Data Reduction and Projection**

It includes tasks such as identifying useful features to represent the data and reducing the effective number of variables under consideration or to find invariant representations for the data.

5. **Choosing the Functions of Data Mining**

In this particular step activities including deciding the purpose of the model derived by the data mining algorithm are defined. These purposes could be summarization, classification, regression and clustering.

6. **Choosing the Data Mining Algorithms**

Selecting method(s) to be used for searching patterns in data and matching a particular data mining method with the overall criteria of the KDD process are the major activities in this step.

7. **Data Mining**

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.

8. **Interpretation**

In interpreting the discovered patterns and returning to any of the previous steps is a possibility. In addition, visualization of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable by users are part and parcel of this step.

9. **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 (Fayyad et al., 1996b). 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 (Han, j. & Kamber, M. 2001).

2.2.2.2 The Data Mining Process

Data mining process has four major steps. These are: data selection, data transformation, data mining and result interpretation. The processes are depicted in Figure 2.2 (source: Sumathi & Sivanandam (2006)) and followed by the discussion of the steps.

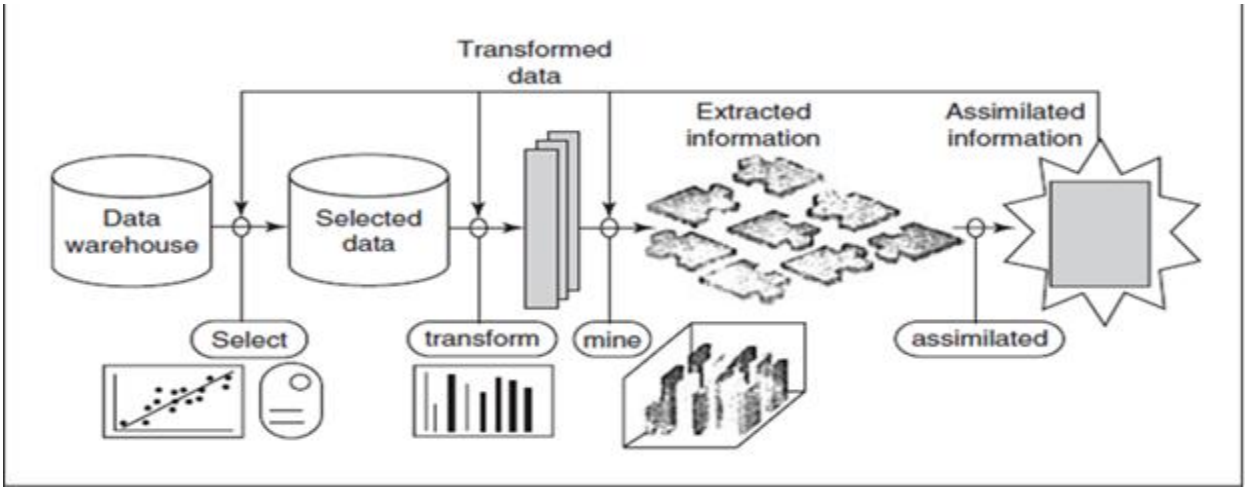


Figure 2. 2 The data mining process

Data Selection

All the data in the data warehouse is not useful to solve a given problem at hand or to achieve a data mining goal. Therefore, preparing a target data is the

first step in the data mining process. It is also important and less expensive to take a sample data to mine.

1. **Data Transformation**

In this step three things should be considered to perform data transformation. These are: the task (fraud detection, for example), the data mining operations (such as predictive modeling), and the data mining technique (such as decision tree) involved. The transformation methods include organizing data in the desired ways (organization of individual consumer data by household), converting one type of data to another (changing nominal values into numeric ones), the definition of new attributes (derived attributes), applying mathematical or logical operators on the values of one or more database attributes – for example, by defining the ratio of two attributes. Hence, it is to be noted that the transformation methods are dictated by the task, data mining operation and the technique used.

2. **Data Mining**

At this point, the desired information is extracted by using one or more techniques on the transformed data. For instance, rule induction could be applied to automatically create a classification model in addition to clustering to predict whether magazine subscribers will renew their subscription or not. Clustering will help to segment the subscriber database before applying rule induction.

3. **Result Interpretation**

According to his or her goals and decision-support task, the researcher must finally analyze the mined information. This analysis helps to identify the best of the information. During this step, the user must also determine how best to present the selected mining-operation results and to clearly show the effect to

the decision maker, who will apply them in taking specific actions (Sumathi & Sivanandam, 2006).

2.2.3 Data Mining Technologies

2.2.3.1 Data Mining Models

In data mining there are several models, examples include, the Six Step Cios Model, the KDD process model, the CRISP-DM model, SEMMA (Sample Explore Modify Model Assess) and others. Since KDD is discussed previously with data mining, the remaining models are discussed as follows:

a) The Six Step Cios Model

This model is developed by customizing the CRISP-DM model for the needs of the academic research community. It consists of six steps, as the name implies, namely understanding the problem domain, understanding the data, preparation of the data, data mining, evaluation of the discovered knowledge and using the discovered knowledge.

1. **Understanding the Problem Domain**

In order to understand the problem domain, defining the problem and determining the research goals, identifying key people, learning current solutions to the problem and domain terminology are necessary. To get this knowledge, there is a need to work closely with domain experts. The research goals then need to be translated into the DM goals and selection of data mining methods is also required from this phase (Kurgan & Musilek, 2006; Tariku, 2011).

2. **Understanding the Data**

This step includes collection of sample data, and deciding which data to use including its format and size. If background knowledge does exist some attributes may be ranked in order of importance. In addition, we need to verify usefulness of the data, completeness, redundancy, missing values, and

plausibility of attribute values with respect to the data mining goals (Kurgan & Musilek, 2006; Tariku, 2011).

3. Preparation of the Data

The success of the knowledge discovery process relies upon this phase and it is assumed to consume half of the research time. The output of this phase is the input for the data mining phase. This step involves sampling, and data cleaning (removing or correcting noise). Feature selection and extraction algorithms are used to do further cleaning on the cleaned data. This will result in a new data record meeting specific input requirements for the planned data mining method (Tariku, 2011). Additionally, Kurgan & Musilek (2006) indicated the need for correlation and significance test, derivation of new attribute and data summarization.

4. Data Mining

It is a major step in the knowledge discovery process. The selected data mining method is applied on the prepared data and then testing the generated knowledge are the core activities in this phase. The phase could result in different models but the one that scored best on test data will finally be selected. The test procedure and result will determine the best model (Kurgan & Musilek, 2006; Tariku, 2011).

5. Evaluation of the Discovered Knowledge

At this point understanding the result, checking for novelty and interestingness of the discovered knowledge are major tasks in this step. Consulting domain experts is necessary in the interpretation of the result, checking the impact of the discovered knowledge and to retain only the approved model. It is also important to revisit the data mining process to pin point the alternative actions that could have been taken to improve the results.

6. Using the Discovered Knowledge

This step is characterized by deployment of the discovered knowledge, planning to monitor the discovered knowledge, documenting the project and extending the application area to other domains (Kurgan & Musilek, 2006; Tariku, 2011).

b) CRISP-Data Mining Model

The Cross Industry Standard Process for Data Mining (CRISP-DM) process model was first established in 1990s by four companies. These companies are Integral Solutions Ltd (commercial data mining solutions provider), NCR (Data base provider), Daimler Chrysler (automobile manufacturer) and OHRA (insurance company). CRISP-DM process model is characterized by six phases and in each phase an iterative process is made to come up with the desired outcome. The reference model for data mining provides an overview of the life cycle of a data mining project. The sequence of the phases is not rigid (Tariku, 2011; Wirth & Hipp, 2000). It is useful for planning, communication within and outside the project team, and documentation. The generic check-lists are helpful for novice users of the model as well as for experienced ones (Wirth & Hipp, 2000). The CRISP-DM process model is depicted in figure. 2.3 and discussion on each phases of the model are presented following the figure.

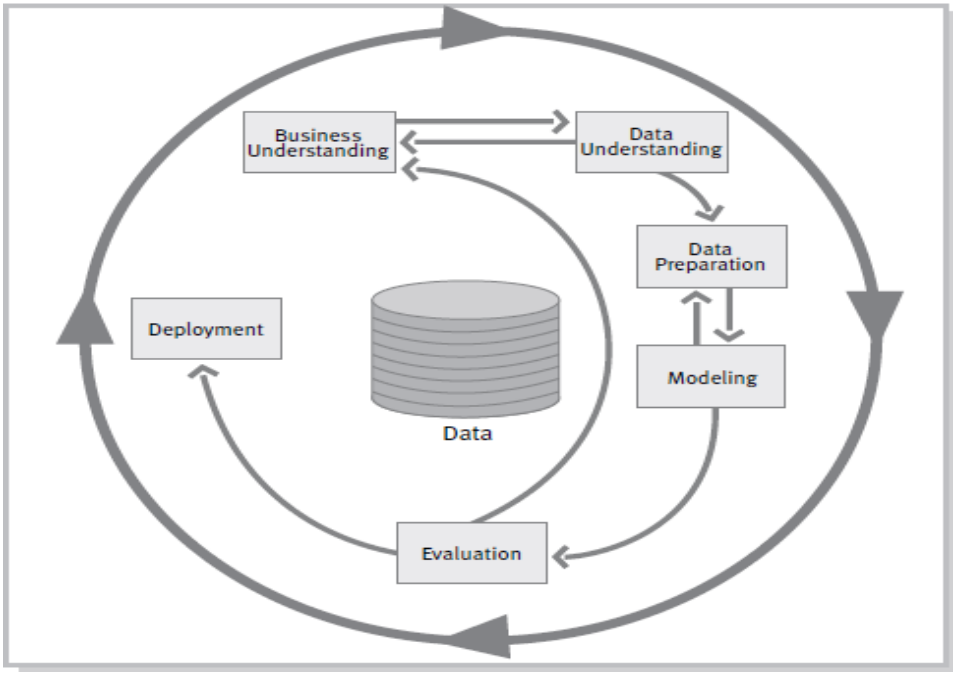


Figure 2. 3 Phases of CRISP DM Reference Model

1. **Business Understanding**

Here, the main objective of this phase is a thorough understanding of what the customer really needs to accomplish. The goal of this phase is to uncover important factors that influence the success of the project. As an output data mining goals and project plan are expected from this phase. Therefore, the major tasks in this phase are determining business objectives, assess situations, determine data mining goals and produce project plan (Chapman et al., 2000).

2. **Data Understanding**

This phase is accompanied by four major tasks which are collecting initial data, describing data, exploring data and verifying data quality. If multiple data sources are used or multiple data are collected, integration will be an additional task.

3. Data Preparation

Selecting, cleaning, constructing, integrating and formatting the data are the tasks in this phase. Regarding constructing the data, it is the process of producing of derived attributes or entire new records, or transformed values for existing attributes. Integration of data also refers to merging two or more tables/tuples to come up with one table/tuple. Merging could also mean aggregating or computing summarized information from multiple records or tables. Dataset(s) are the output of data preparation phase that will be used for modeling or the major analysis work of the project.

4. Modeling

In this phase, selecting modeling techniques, generating test design, model building and assessing the model are the tasks undertaken. In model building sub task running the modeling tool on the dataset will result in parameter setting (With any modeling tool, there are often a large number of parameters that can be adjusted), models (actual models produced by the modeling tool) and model description (report on the interpretation of the models and document any difficulties encountered with their meanings). As a sub task model assessment also the researcher or data mining engineer interprets the models according to his domain knowledge, the data mining success criteria, and the desired test design. Finally it results in model assessment (listing qualities of generated models and ranking them) and revised parameter setting (revising parameters and tuning is necessary till the best model is obtained). (Chapman et al., 2000).

5. Evaluation

This phase is characterized by evaluating results, reviewing the process and determining the next steps. Here evaluating the results is made by assessing the model against the business objective and checking if there is any business reason that make the model deficient. The model will be approved if it meets all

the business objectives. In reviewing the process, a thorough review of the data mining engagement is done to make sure that important tasks are not overlooked. As a result, it enables us to highlight activities that have been missed and those that should be repeated. In determining the next step, decisions will be made based on the result of the assessment and process review. At this point, the decision is finishing the project and moving to deployment, initiating further iterations, or setting up new project. In doing so, analysis of resources and budget are given due consideration.

6. Deployment

This phase has four tasks namely: planning deployment, planning monitoring and maintenance, producing final report and reviewing project. Planning deployment considers the evaluation result and summarized deployment strategy will be determined including the necessary steps and how to perform them. In planning monitoring and maintenance a detailed monitoring process plan and careful preparation of maintenance strategy needs to be prepared including the necessary steps and how to perform them. At last, final report will be produced including deliverables by summarizing and organizing the results.

c) SEMMA

According to the SAS process for data mining, SEMMA is to mean Sample, Explore, Modify, Model and Access. SEMMA is developed as a set of functional tools for SAS's Enterprise Miner software. Therefore, those who use this specific software for data mining task are more likely to adopt this methodology(Harding, Shahbaz, & Kusiak, 2006). The forth coming paragraphs discuss the five steps of SEMMA in a precise manner.

1. **Sampling**

After identifying the input dataset, there is a need to sample the data when the data source is a large database. This will help to reduce the model training time. The tasks, in this phase, are sampling from a larger data set, partitioning data set into training (for preliminary model fitting), validation (to monitor and tune the model weight as well as model assessment), and test data sets (for additional model assessment).

2. **Explore**

Under explore node, there are different nodes like distribution explorer, multi-plot, insight, association, variable selection and link analysis.

The distribution explorer node enables to explore large volume of data in multidimensional histograms. To exclude extreme values for interval variables, it is possible to set a range cutoff. The node also generates simple descriptive statistics for the interval variables. Similarly, multi-plot enables to view the large volume of data graphically by creating bar chart and scatter plots automatically for input and target variables. On the other hand, insight node enables to analyze univariate distributions, investigate multivariate distributions, and fit explanatory models by using generalized linear models. Additionally, the association node enables to identify association relationships within the data and sequence discovery if a time-stamp variable (a sequence variable) is present in the data set. The variable selection node enables to evaluate the importance of input variable for classifying or predicting the target variable by using either an R-square or a Chi-square selection (tree-based) criterion. At last, link analysis node enables to transform data into a data model that can be graphed. The data model generates cluster scores that can be used for data reduction and segmentation.

3. Modify

Modify nodes is characterized by different nodes namely data set attribute node, transfer variable node, filter outliers node, replacement node, clustering node, SOM/Kohonen node, time series node and interactive grouping node. Each node is expected to complete its sub task before moving to the next node. Then the aggregate result of each node takes to the next phase that is modeling.

4. Model

Modeling node has different nodes with varying functionalities. The first one is regression node which enables to fit both linear and logistic regression model by accepting continuous and discrete variables as input. The second one is tree node which performs multi-way splitting of database, based on nominal, ordinal and continuous variables. The node supports interactive and automatic training. It implements a hybrid of CHAID, CART and C4.5 algorithms. The third one is the neural network node which enables to construct, train and validate multilayer feed-forward neural network. The default multilayer neural network is a multilayer network that has one hidden layer consisting of three neurons and each node is fully connected. It also supports many variations of this general form. The remaining nodes include princomp/Dmneural, user defined model, ensemble, memory based reasoning and two stage model nodes.

5. Assess

Under assess nodes there are two nodes namely assessment node and reporter node. Assessment node is characterized by providing a framework for comparing models and predictions from any of the modeling nodes. It also produces charts like lift, profit, return on investment, receiver operating curves, diagnostic charts, and threshold-based charts that help to describe the usefulness of the model. But, the reporter node assembles reports that can be viewed using web browser (SAS-Institute, 2003).

2.2.3.2 Data Mining Tasks

Depending on the objective that is analyzing the data, it is possible to categorize data mining based on tasks. According to (Hand et al., 2001), data mining tasks include exploratory data analysis (EDA), predictive modeling, descriptive modeling, discovering patterns and rules, and retrieval by content.

1. **Exploratory Data Analysis (EDA)**

As the name implies, the goal here is to simply explore the data without any clear ideas of what we are looking for. The techniques used here are interactive and visual. Further, there are many effective graphical display methods for relatively small and low-dimensional data sets. Pie chart is an example of EDA application (Hand et al., 2001).

2. **Predictive Modeling**

This task is used with the aim to build a model that will permit the value of one variable to be predicted from the known values of other variables. In classification, the variable being predicted is categorical, while in regression the variable is quantitative(Hand et al., 2001). There are a number of methods developed in statistics and machine learning to tackle predictive modeling problems, and work in this area has led to significant theoretical advances and improved understanding of deep issues of inference. These predictive models, including additive regression, decision trees, neural networks, support vector machines, and Bayesian networks, have attracted attention in data mining research and applications. As modern computing power allowed data miners to explore and come up with more complex models(Sumathi & Sivanandam, 2006).The difference between prediction and description is in target variable. Prediction has a single or unique variable as objective, while descriptive problems have no single variable, central to the model (Hand, J. and et al., 2001).

Predictive modeling is mostly considered as a high-level goal of data mining in practice. After outlining the predictive modeling problem, we focus on two classes of algorithm: decision tree methods and support vector machines. Input into predictive modeling algorithms is a data set of training records. The goal is building a model that predicts a designated attribute value from the values of the other attributes (Sumathi & Sivanandam, 2006).

3. Descriptive Modeling

In descriptive modeling the goal is simply describing all of the data (or the process generating the data). As an example of such description include models for the overall probability distribution of the data (density estimation), partitioning of p-dimensional space into groups (cluster analysis and segmentation), and models describing the relationship between variables (dependency modeling). Segmentation analysis is used to group together similar records. Here the goal is to split the records into homogenous groups. It is widely and successfully used in marketing to segment customers based on Sub_Age, income and other variables. This contrasts with cluster analysis, in which the aim is to discover 'natural groups in data', in scientific database. Descriptive modeling is used in many ways like, segmentation and cluster analysis (Hand et al., 2001).

4. Discovering Patterns and Rules

The previous tasks are concerned on model building whereas other data mining applications are concerned with pattern detection. Examples in this area are fraud detection, detection of unusual stars in astronomy and finding combination of items occurring frequently in transaction databases. Such kinds of problems are addressed using data mining and using algorithm techniques based on association rules. A fraudulent use of cellular telephones is estimated to cost the telephone industry several hundred million dollars per year in United States. Application of rule learning algorithms to discover

characteristics of fraudulent behavior resulted more accurate than existing hand-crafted methods of fraud detection.

5. Retrieval by Content

Here the user's interest and wish is to find similar patterns in the data set. It is widely applied in text and image data sets, in information retrieval. The web and QBIC (Query by Image Content) are examples for applications in retrieval system (Hand, J. and et al., 2001).

2.2.3.3 Data Mining Techniques

Data mining adopted different techniques from many research areas, including statistics, machine learning, database systems, neural networks, rough sets, and visualization. These techniques are used to solve different problems in different areas using data mining. It is common in data mining to use multiple methods to deal with different kinds of data, different data mining tasks, and different application areas.

It is the researcher's belief that discussing few techniques of data mining will suffice for this work. The following discussion on data mining techniques limits itself on decision tree and neural networks.

Decision Trees

Decision trees are produced by algorithms that identify various ways of splitting a dataset into branch-like segments. These decision trees are of two types. These are: classification and regression trees. Classification trees label records and assign them to the appropriate class, predict categorical variables. Regression trees estimate the value of a target variable that takes on numeric values, predict continuous variables (Tariku, 2011). In other words, decision trees used to predict continuous variables are called *regression trees (Two-Crows, 1999).* The segments, produced by the algorithms, form an inverted decision tree that originates with a root node at the top of the tree. The object

of analysis is reflected in this root node as a simple, one-dimensional display in the decision tree interface. The display of this node reflects all the data set records, fields, and field values that are found in the object of analysis. The discovery of the decision rule to form the branches or segments underneath the root node is based on a method that extracts the relationship between the object of analysis (that serves as the target field in the data) and one or more fields that serve as input fields to create the branches or segments. The values in the input field are used to estimate the likely value in the target field. The target field is also called an outcome, response, or dependent field or variable.

Once the relationship is extracted, then one or more decision rules can be derived that describe the relationships between inputs and targets. Rules can be selected and used to display the decision tree, which provides a means to visually examine and describe the tree-like network of relationships that characterize the input and target values. Decision rules can predict the values of new or unseen observations that contain values for the inputs, but might not contain values for the targets (De Ville, 2006).

Decision trees are especially attractive in data mining environments since the resulted model can easily be understood by human analysts. The construction of the trees does not require an analyst to provide input parameters and prior knowledge about the data. A record can be associated with a unique leaf node based on the splitting criterion, which evaluates a condition on the input records at the current node.

Decision tree adopt different algorithms for tree building (splitting) and pruning. Among the algorithms Chi-Squired Automatic Interaction Detection (CHAID), Classification and Regression Tree (CART), Quest, C4.5 and C5.0 are the commonly implemented ones (Bounsaythip & Rinta-Runsala, 2001; Two-Crows, 1999). The entire tree algorithms, mentioned above, suit for classification and only some of them are adaptable for regression. They are distinguished by target variables, split (binary or more than two splits), split

measures (criteria based on gain, gain ratio, GINI, chi-squired and entropy) and rule generation. Regarding, target variables most tree algorithms require target (dependent) variables to be categorical and continuous values to be binned (grouped) to be used by regression tree. Rule generation is all about generalizing the rules. Generalizing rules will help to remove redundancies; algorithms like C4.5 and C5.0 are cases in point of this type.

All algorithms have their own pros and cons. As an advantage decision tree algorithms are not affected by missing values. However, they impose restrictions on the data analyzed. Among the restrictions include, allowing only one dependent variable, and requiring continuous data to be grouped or categorized (Bounsaythip & Rinta-Runsala, 2001). Detail discussion is made in chapter three.

Artificial Neural Networks

ANN can be classified in two major classes, namely feed forward neural networks (FNNs) and recurrent neural networks (RNNs). In feed forward networks, activation is “piped” through the network from input units to output units. Sometimes they are also referred as static networks. FNNs contain no explicit feedback connections. Conventional FNNs are able to approximate any finite function as long as there are enough hidden nodes to accomplish this. RNNs on the other hand, are dynamical networks with cyclic path of synaptic connections which serve as the memory elements for handling time-dependent problems.

ANNs have the capability to learn from their environment through an iterative process of adjustments applied to its synaptic weight and bias level. They are also able to improve their performance through learning. There are many varieties of learning algorithms for the design of ANNs. They differ from each other in the way in which the adjustment to a synaptic weight of a neuron (node) is formulated. Learning algorithms can be described as a prescribed set of well-defined rules for the solution of a learning problem. Error-correction

learning, memory-based learning, Hebbian learning, competitive learning, and Boltzmann learning are among the learning algorithms for ANN.

ANN learning paradigm is either supervised (associative learning) or unsupervised (self-organizing). In the case of supervised, there is a need to train or teach the input and output pattern. But for the case of unsupervised neural network, it only requires input patterns from which it develops its own representation of the input stimuli.

1. **Feed Forward Neural Networks**

A feed forward network has a layered structure. Each layer consists of processing units (or neurons). The layers are arranged linearly with weighted connections between each layer. These layers are input, hidden and output layers. The input and output layers have no incoming and outgoing connections respectively. All connections point in the same direction – a previous layer feeds into the current layer and that layer feeds into the next layer. The hidden layers possess both types of connectivity. These layers contain hidden units which are able to extract higher-order statistics. This is particularly valuable when the size of the input layer is large.

In general, conventional feed forward neural networks (FFNNs) are static learning devices since they only have a very limited ability to deal with time-varying input. However, it is possible to adapt them to deal with temporal relationships (Abidogun, 2005).

2. **Recurrent Neural Networks**

Recurrent neural networks are grouped as Elman networks, Jordan networks and fully recurrent neural networks. These networks differ in the connection for feedback among the neurons. It is to be recalled that neurons are arranged in different layers. In Elman networks, feedback connections exist between hidden neurons and the hidden neurons are used to learn a representation of a

dynamical system's hidden states being modeled. In Jordan networks, feedback connections in the output layer are fed back into the hidden layer. In fully recurrent neural networks, connections exist among all the network's neurons. These feedback connections enable these networks to create a memory of past events that occurred before.

Among the algorithms that recurrent neural network (RNN) use is gradient descent learning which is the most commonly used one. The aim of gradient descent learning is to find the best possible set of weights with minimum margin of error. Learning in recurrent networks is accomplished by finding the minimum of an error function over all sequences. It is calculated by measuring the difference between desired target outputs and actual output.

The weight is then updated according to the learning rate. With a small learning rate, a network will take a considerable period of time to converge to the desired solution if one exists. Too large a learning rate may result in divergence. If the learning rate parameter is increased, the settling time of the network also increases which is the result of overshooting the solution. After the error signals have been calculated, they are added together and contribute to one big change for each weight. This is known as batch learning. An alternative approach is on-line learning which allows the weights to be updated after each pattern is presented to the network.

Potential applications of RNNs are time series prediction, time series production (like motor control in non-Markovian environments) and time series classification or labeling (e.g., rhythm detection in music and speech). The commonly used gradient descent based learning algorithms for recurrent networks are Back-propagation Through Time (BPTT) and Real-Time Recurrent Learning (RTRL) (Abidogun, 2005).

The major drawback on neural networks, either supervised or unsupervised one, is that the features used to reach to the desired performance is not clearly

known (Hilas & Mastorocostas, 2008). Neural networks are considered as "black boxes" due to their non-linear behavior and complexity than other methods. The output is not easily understood by the user compared to other methods or when the output is seen by decision tree tool. Therefore, it is difficult to identify the important characteristics that lead to a successful classification and yet they are applicable in a variety of business applications and save their users time and money in the process (Tesfaye, 2002).

2.2.4 Data Mining and Other Statistical Tools

Data mining is a multidisciplinary field that has interactions with other disciplines primarily with statistics, computer science, and information systems. Data mining explores the general areas of massive data set. Similarly, the issue of data storage and retrieval is addressed by information science. Whereas the algorithms that are used to extract information/data from the databases is the concern of computer science. On the other hand, statistics try to study the best way of analyzing the data and seek meaningful relationships among sets of variables(Tibebe, 2005). Although, data mining and statistical techniques have things in common, many scholars tried to show their difference in a precise manner.

Moss (2000) described that statistical techniques use numerical data whereas; data mining uses different types of data, including numerical data.

Statistics is mainly concerned on hypothesis testing but data mining is with the formulation of the process of generalization as a search through hypothesis testing. Regarding their similarity, both statistics and machine learning work on classification and regression. Almost at the same time when statisticians published a book on classification and regression techniques, the machine learning field came up with a system that works on classification trees. Additionally nearest-neighbor method is used for classification by both disciplines (Tibebe, 2005).

2.2.5 Data Warehousing and OLAP Technology for Data Mining

Data warehousing and OLAP (Online Analytical Processing) are the two technologies that are related to data mining. OLAP is becoming an important tool for decision making in corporations and other organizations. It is also one of the main focuses of the database industry. However, the functions and properties of decision support system are rather different from the traditional database application.

Here the data is collected and stored in the data warehouse (Sumathi & Sivanandam, 2006). Data warehouse is a recently emerged technology to store multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management decision making. In data warehouse technology data cleansing, data integration, and OLAP techniques are included. Figure 2.4 below illustrates the data warehouse system architecture. OLAP is also the analysis technique with functionalities such as summarization, consolidation, and aggregation, and the ability to view information from different angles as well (Jiawei & Kamber, 2001). Before it is stored, it is processed (cleaned and transformed) to meet the requirements of data warehouse. Data warehouse is the relational database management system. But it is specifically designed for a transaction processing system. These warehouses contain millions of pieces of information about customer's needs and distribution decisions. Data warehouse uses the data to analyze business needs and to make the decisions. They make amounts of data that span over many years. A data warehouse is not a transactional database. Transactional database uses data for day to day business operations to run fast and in an efficient manner while data in data warehouse run very slowly and not used for day to day operations directly.

OLAP is among the category of software tools that provide analysis of data stored in a database. OLAP provides the users with multidimensional database to generate on-line description, or it compares the "views" of data and other

analytic queries. OLAP gives the answers to multidimensional business questions quickly and easily. OLAP technology provides facts and efficient access to summarized data. It enables to give control over global views of the business. For example, OLAP provides time series and trend analysis views(Sumathi & Sivanandam, 2006). Even if OLAP tools support multidimensional analysis and decision making, additional tools like data classification, clustering and other techniques are required for in depth analysis (Jiawei & Kamber, 2001).

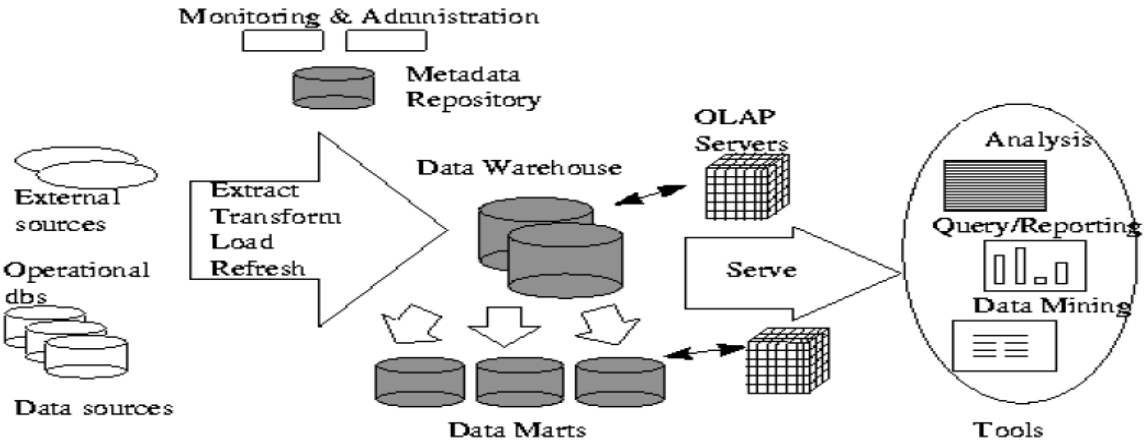


Figure 2.4 Architecture of typical data warehouse system

It is possible to apply OLAP technology in many areas. Some of these are: sales and marketing analysis, financial reporting, quality tracking, profitability analysis, manpower and pricing applications, our unique data discovery needs, and so on (Sumathi & Sivanandam, 2006).

2.2.6 Application of Data Mining Technologies

Data mining is applied in almost every discipline and every part of the world. Data mining is a blend of concepts and algorithms from machine learning, statistics, artificial intelligence, and data management. The emergence of data mining enabled researchers and practitioners to apply this technology on data from different areas such as telecommunications, banking, finance, retail, marketing, insurance, fraud detection, science, engineering, etc., to discover

any hidden knowledge, relationships or patterns. Data mining is, therefore, a rapidly expanding field with growing interests and importance and telecommunication is one potential application area where it can provide significant competitive advantage (Harding et al., 2006).

1. General Applications

In general, data mining is applied in different industries with different purposes. For instance, it is applied to attract new customers, to maximize revenue and to retain existing customers. Data mining helps to profile customers that have experiences with the business organization. It can be applied on customers that are good, that bought or did not buy their product and who used to be their customers. Profiling customers in such a manner will help companies to design appropriate marketing strategies to win new ones and retain existing customers.

A range of industries are gaining values from data mining. The leading industries that adopted data mining for fraud detection are telecommunication and credit card companies. Insurance and stock exchange markets apply data mining to minimize fraud. Data mining can also be applied in medical, financial, retail, security and pharmaceutical firms (Two-Crows, 1999).

2. Application of Data Mining in Telecommunications

The rapid expansion of telecommunication market and high competition created the demand for data mining. It helps to understand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of service. As mentioned earlier, telecommunications and credit card industries are among the leading industries by applying data mining.

Among the scenarios where data mining contribute to improve telecommunication services are many but only few of them are discussed below.

The first one is on multidimensional analysis of telecommunications data. Since telecommunication data are intrinsically with dimensions such as calling date and time, duration, location of caller, type of call, amount charged and the like. The multidimensional analysis of such data can be used to identify and compare the data traffic, system work load, resource usage, user group behavior, profit, and so on.

The second scenario is on fraud pattern analysis and identification of unusual patterns. Millions of dollars are lost every year due to fraudulent activity in telecom industry. The importance of detecting fraudulent users and their patterns is very high for the company as well as for customers (Sumathi & Sivanandam, 2006). In addition to the revenue leakage, there will also be a compromise on quality of service and security (Lokanathan & Samarajiva, 2012; Negarit, 2012).

The other scenario is the application of data mining on telecommunications churn analysis. This is a critical problem for telecom industries where there are multi venders. Customers easily switch from one operator to the other due to different reasons like lack of good service, pressure from peer group and others. Data mining still provides solution for such problems by predicting and understanding why particular customers churn (Sumathi & Sivanandam, 2006). Thus, telecommunication fraud is a significant problem, for the developed as well as for the developing nations, which needs to be addressed, detected and prevented in the strongest possible manner (Abidogun, 2005).

2.3 Related Works

Regarding fraud detection different research are conducted both in Addis Ababa University and CTIT. An effort has been made to review the papers that the researcher found as presented in the forthcoming paragraphs.

A research has been conducted on fraud detection with the title Fraud Detection in Telecommunication Networks using Self-Organizing Map (SOM): the case of Ethiopian Telecommunication Corporation (Berhanu, 2006). He used unsupervised feed-forward neural network model, which is SOM, it helps to analyze and visualize high dimensional data. SOM also enables clustering data without knowing the class membership of the input data, unlike neural network models based on supervised learning. Then the clustering capability of SOM is used to group similar call pattern behavior analysis. He used extended map model to identify suspicious calls and the result has shown that these calls are identified as fraudulent or suspicious call patterns. As a result domain experts and users confirmed the result; however, verification from fraud analysts is not done.

Another research on Ethiopian Telecommunication Corporation (ETC) is on challenges facing international telecom business and the way forward, ETC's perspectives by Asfaw (2006). The research is done by limiting the scope to assess the trends in international voice telephony business and gave attention for identifying challenges faced by ETC in the sector. The challenges are seen from three perspectives. These challenges are: technological, policy reform and pressure on total accounting rate (TAR). For this research, interview and document analysis were used as a methodology to pin point the challenges ahead. As a finding, technology, such as voice over internet protocol (VOIP), calling cards and home country directed services were identified as the major challenges. It is also indicated that developing countries are beneficiaries as long as the international traffic balance is in favor of them. When the incoming international traffic is higher than the outgoing traffic then the operator with

higher incoming traffic will get settlement income. Developing countries use this settlement income to finance their development programs. ETC is among those state owned operators benefiting for a long time. But according to the finding of the research, it is indicated that technology is among the major challenge for terminating international voice traffic, which affect the income of ETC. The research is concluded by emphasizing how developed nations use technological progress which can bypass the traditional bilateral route so as to narrow their deficit gap. As a result, the settlement income that developing countries earn diminishes and the carriers will be challenged if they mainly relay on it, like ETC.

Another study was also conducted under the title Using Data Mining to Combat Infrastructure Inefficiencies: The Case of Predicting Non-payment for Ethiopian Telecom. It is basically on customer relationship management (CRM) of ETC with regards to customer complaints or refusal to pay bills, and the resulting actions of cancellation and charging. They tried to rank Ethiopian Telecommunication customers likelihood of facing subscription cancelation or service termination due to bill nonpayment. They applied data mining classification techniques like decision trees, nave Bayes, and logistic regression to predict nonpayment. The applied data mining techniques showed that a change in call usage pattern has a strong relationship with nonpayment of bill in future. In other words, they found out that a change in behavior or bill consumption is a key indicator of future non-payment. They also proposed prediction and early prevention of default could save the corporations revenue loss (Yigzaw et al., 2010).

The other work referred here is also on CRM on ETC, specifically for CDMA (Code Division Multiple Access) telephone customers and applied data mining to make behavioral segmentation. It is made with the objective to enable the corporation to identify, create and maintain good relationship with customers. The researcher used classification and clustering data mining techniques on

customers' database and adopted CRISP-DM model. The applied data mining tools are K-means clustering, decision tree (J48) and artificial neural network (feed forward backward propagation). For his research he used the CDMA CDR, bill data and customers profile data from 'USHACOM' system of the corporation. As a result using decision tree he managed to get 98.97% accurately classified and 98.62% using neural network. The numbers of customers wrongly classified are 103 and 139 using decision tree and neural network respectively. For both high valued and low valued customers decision tree resulted in better accuracy than that of neural network (Melaku, 2009).

Still another study on CRM with the title Application of Data Mining to Support CRM on ETC is made with the objective to help the organization to maintain appropriate CRM for the purpose of transforming customer data into meaningful segments of customers based on underlining similarity. The researcher followed both qualitative and quantitative approaches. The scope was limited to postpaid mobile customer and their calling behavior for a month. He followed CRISP-DM process model and implemented clustering and classification techniques. As a result, K-means exhibited good result (more dissimilar clusters) when K is 6. Decision tree also resulted 94.93% overall accuracy level where 60% of the data was used for training, 30% for testing and 10% for validation. The researcher also reported that domain experts agreed on the result that customers are clustered based on their calling behavior and as per their long term value to the organization (Fekadu, 2004).

The other research with a title "Fraud detection on post-paid mobile customers of Ethiopian Telecommunications Corporation" is made by using CDR data, bill data and customer database of returned bills maintained by finance department for follow up purpose. The main focus was on subscription (accounting) fraud type only. He applied data mining steps assuming business problems are already identified. He used neural network model and MATLAB software for his study. As a result he found an accuracy level of 89% and

proposed further study on other possible sources of fraud based on pure CDR data which is automatically generated from the switch machine (Jember, 2005).

The other paper on fraud detection focused on detecting illegal calls from CDR switch machine of Ethiopian Telecommunication Corporation and enabling early detection of those calls. Neural network technique and Brain Maker Neural Network Software was used in the study. The data source for the study was CDR and the main focus was on pre-paid mobile phone of ETC. Finally, 88.46% of accuracy level for fraudulent call is achieved and 4.49% error rate for non-fraudulent calls. Finally, a recommendation is made to conduct further research by including other attributes of CDR, so as to build models with better performance and accuracy level (Gebremeskal, 2006).

Another research on principles of effective regulation to curb illegal bypass in international voice traffic is made by taking the case of Bangladesh. The problem with illegal bypass of international voice traffic is reduced voice call quality, security issue and reduced revenue for governments. It is a descriptive research made to show the impact of illegal bypass and provide recommendations for telecom operators in general. They made reviews of the laws of the Bangladesh government telecom policy and actions taken against illegal practices. They also assessed the impact and contributions of different organizations. These organizations include International Gateway (IGW), International Internet Gateways (IIG), Interconnection Exchange (ICX) and International Long Distance Telecommunication Service (ILDTS). In 2007, 80% of Bangladesh's incoming international traffic was routed by illegal VOIP business. Due to this fact, the government decided to take legal actions on those who are found doing the business. Additionally, ILDTS 2007 policy was implemented to restructure the telecom sector. This paper analyzed the structural flaws in ILDTS 2007. As a result, four recommendations were made to fill the gap created by the policy implementation. The recommendations are

categorized into two groups namely reduction of international call termination rates in Bangladesh to domestic termination rates, and by aligning the incentives and abilities of the various stakeholders in the new network topology. The first option is claimed to eliminate illegal bypass entirely but the second option is made by considering the government's need for revenue from international voice traffic (Lokanathan & Samarajiva, 2012).

As you can see from the above literatures reviewed no research is conducted to detect fraudulent international calls using SIM boxes. The research by Lokanathan & Samarajiva (2012) focused on policy measures taken by Bangladesh government and other international telecom organizations policy problems. But this study is not on policy issues but on developing a predictive model identifying fraudulent calls, their approximate locations and availing them for decision making.

The other study made by Gebremeskal (2006) is somewhat similar with this study but it is entirely different. Initially, he tried to identify fraudulent calls in general but this study tried to identify fraudulent calls used to terminate international calls. This one is specifically, the calls from SIM boxes. In addition, such frauds are very recent phenomenon. Moreover, he used neural network and Brain maker data mining tool, but in this study decision tree, and neural network are applied using WEKA data mining tool. The result from his experiment is 88.46% accuracy but the result from this study is expected to exceed.

Finally, an effort is made in this study to indicate the location of the fraudulent mobile from the location information found in the CDR. Additionally, in this study SMS, GPRS and voice CDR are used in addition to OCS data but in the case of Gebremeskal (2006), only CDR from switch is used. On top of this, due care is given to protect the privacy of customers by using different number other than mobile numbers. By the time Gebremeskal made the study on fraud detection the number of mobile customers were less than one million but

currently it is more than 18.28 million, as reported by the CEO on performance related press conference (Ethio-telecom, 2012b).

Chapter Three

Data Mining Methods

For this particular research, different data mining methods like decision tree and neural network are used. J48 and PART algorithms are used from decision tree. In addition, multilayer perceptron is used from neural network as other researchers proposed it for fraud detection (Adu-Boafo, 2013). The resulted models from the above algorithms are compared to propose the best one for this study. Further discussion on the mentioned techniques and algorithms is the focus of this chapter.

3.1 Classification techniques

3.1.1 Decision Tree

Decision trees are among the fundamental techniques used in data mining. It is used for classification, prediction and feature selection. Decision trees are easily interpretable and intuitive for humans, suitable for high dimensional applications, fast and produce high quality solutions, and its objectives are consistent with data mining and knowledge discovery.

Decision trees are described as universal approximators, like neural networks, because they map linear and nonlinear relationships. They require less training data compared to other universal approximators.

Decision tree consists of a root and internal nodes. The nodes are labeled with questions in order to get solution to the problem at hand. A decision tree, as indicated in figure 3.1 below, is a binary tree if each node is splited into two and non-binary tree (multi-branch) otherwise. If a node can't be splited any further, it is known as terminal node. When a terminal node is reached, its stored value (or content) is returned (M. W. Berry & Browne, 2006).

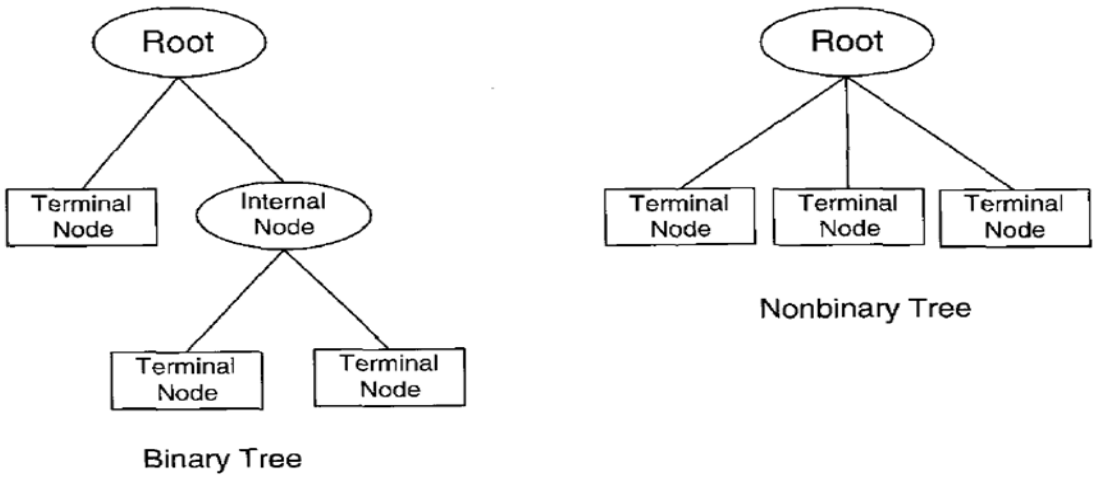


Figure 3. 1 Decision Tree Types

Decision trees are known for portioning data and identifying local structure in small as well as in large databases. It has two objectives namely producing accurate classifier and understanding the predictive structure of the problem. Among the two goals, accurate decision tree classification is the first one. The second goal is developing understandable pattern that can be interpreted as knowledge. The unique characteristics of decision tree made data mining to be preferred by experts in the domain. Among them are the clear depiction of relationship between input data and target output. They accept several types of variables (nominal, ordinal and interval) that can be implemented with little or no consideration for converting odd variables (e.g., opinions, biases, or risks) to more appealing types. They are robust with respect to missing values and distribution assumptions, and are well suited for high dimensional data.

Decision trees can make use of dynamic feature selection to produce fast nonlinear prediction methods. The number of features can also be reduced using approaches like principal component analysis and decision trees. Decision trees are helpful for feature selection particularly when there are large feature spaces. As a result time consuming prediction methods, like ANN, can

be applied on the reduced database. The database is reduced when some features are deleted after applying the mentioned approaches. Decision trees are easy to interpret, amenable to graphical display and intuitive for humans, given the size of the tree is minimal. Decision trees are used as a bench mark to evaluate the performance of other techniques (Berry & Browne, 2006).

Decision tree methods learn the decision trees by implementing a top-down approach. This approach starts by selecting an attribute that will be used for partitioning. It is determined by evaluating using information gain, to measure how well it classifies the training examples. This process will be repeated iteratively to determine the descendant nodes but it never back tracks to reconsider previous choices since decision tree follows greedy search. Decision tree is robust to noisy data and capable of learning (Ye, 2003).

3.1.2 Selection of splitting variable

A less complicated tree could be constructed by selection variables with best splits. Split measures are considered to select which variable to use for splitting a particular node. They base the splitting criteria on gain, gain ratio, GINI or chi-squared.

Stopping criteria

Manageable sizes of a decision tree are easy to interpret. But smaller trees do not describe the training data very well. So, trees should not be too small or too large to perform well on new data sets. Allowing the algorithm to make use of all the data may result in large trees, but it guarantees that all information has been captured (are included for the algorithm to make decision). Stopping rules are disadvantageous in three aspects namely in choice of statistical test, accommodations for multiple comparisons, and the choice of a threshold (M. W. Berry & Browne, 2006).

Tree pruning

As experts agreed, stopping rules cannot work due to the mentioned drawbacks in the previous paragraph. Rather, tree pruning technique will help to reduce the fully grown trees to manageable size. Tree pruning process evaluates sub trees rather than individual splits. Pruning is useful to avoid over-fitting the data. Error estimation techniques such as reduced error pruning, cost complexity pruning and pessimistic pruning algorithms play a major role in tree pruning (M. W. Berry & Browne, 2006).

Stability of decision trees

Tree pruning alone does not guarantee stable decision tree. Other techniques in data mining, such as neural networks, decision trees may face instability. To avoid such things from happening methods such as arcing, boosting and bagging are used to make decision trees more stable and provide accurate predictions.

The objective of all decision tree algorithms is to minimize the size of the tree by maximizing its classification accuracy. Decision tree algorithms choose the splitting attribute as well as decide on how many branches or what values to assign to that node.

Interactive Dichotomizer 3 (ID3), C4.5, C5.0, J48, PART, and random forest are classification and regression tree (CART) are the most commonly used decision tree algorithms. Among the mentioned decision tree algorithms J48 which is the improved version of J4.5 is used for this specific study in addition to PART.

C4.5 algorithm creates trees using basic inductive approach like that of ID3 but it is capable of classifying continuous values by grouping together discrete values of an attribute into subsets or ranges. The advantages of this algorithm are predicting values for data based on knowledge of relevant domain and providing ways for pruning (sub tree replacement and sub tree raising) without

significant decrease in accuracy. Here sub tree replacement is replacing a sub tree with leaf node and sub tree raising is replacing sub tree with a most frequently used one.

In both scenarios, replacement is accepted if the original tree undergoes minimal distortion. When the decision tree structure complexity can't be minimized effectively, then C4.5 algorithm will generate rules based on the choices associated with the path which is defined as set of branches connecting two nodes (M. W. Berry & Browne, 2006).

J48 is WEKA's implementation of this algorithm. This algorithm is helpful in generalizing rules associated with a tree and helps to remove redundancies. J48 is actually known as J4.5 revision 8. Since the researcher is using WEKA as a tool, J48 is applied on the data selected for this purpose.

In addition to J48, PART algorithm is also used in this study from decision tree. PART is a partial decision tree algorithm, which is the developed version of C4.5 and RIPPER algorithms. The main advantage of the PART algorithm is that it does not need to perform global optimization like C4.5 and RIPPER to produce the appropriate rules. However, decision trees are sometime more problematic due to the larger size of the tree which could be oversized and might perform badly for classification problems (Ali & Smith, 2006). PART is an indirect method for rule generation. Using separate-and-concur strategy, PART generates a pruned decision tree for each of the iterations. From the best tree, the leaves are translated into rules. PART adopts the divide-and-concur strategy in that it builds a rule, remove the instances it covers, and continue creating rules recursively for the remaining instances till none are left. It uses C4.5 statistical classifier to generate a pruned decision tree (Frank & Witten, 1998).

3.1.3 Advantages of decision tree

The pros and cons of every method used in data mining have to be known in advance before it is applied for any particular study. In the same manner, the following discussions focus on the advantages followed by disadvantages of using decision tree.

1. Human understandable representation: The decision tree can be transferred into a set of "if-then" rules to improve human readability and easy understanding.

2. Decision tree learning methods are robust to errors in the classification of the training set. The testing set and pruning algorithms are used to minimize the tree size and misclassification.

3. Some decision tree methods can be used even when some training examples have unknown values. ID3 does not provide a mechanism to handle the missing attribute values, but C4.5 and C5.0 do.

After decision tree is finalized, there is pruning to eliminate some of the branches to avoid lengthy decision tree. In this process, unnecessary branches are eliminated and the tree will have reasonable size. The basic algorithm of reduced error pruning works in the following manner. Initially, a validating set (like training set) is prepared, next the decision tree is built till over fitting occurs, then nodes are pruned iteratively from the leaf (bottom) by using validating set and pruning will stop when no more pruning is required (Ye, 2003).

The following are limitations and dangers for decision trees. Initially, the limitation is subject to ordinal or nominal data, and execution speed is an issue when continuous variables are used. The next problem is caused by splitting the data by single variables at a time. This can cause high error rates if classes are shaped like two circles but with overlapping projections on some

axes. This problem can be reduced by adding linear combinations of variables to the list of explanatory variables but it will affect the execution time greatly. The third major danger of continuous data is over fitting. When we are dealing with measured data with random variation, it is common for the measurement variability to be large. This creates a large number of unique values to split the training set on, which may not match up well with the test data set. What fits well for a training set may have much higher error rates for the test set. The last issue with this method is the size of the decision tree. As the number of generated nodes exceeds a certain limit, it greatly decreases the interpretability of the decision tree (Ye, 2003).

As indicated earlier, with these entire limitations decision tree is widely used in many data mining projects and researches. As other data mining algorithms are not free of drawbacks, decision tree is not an exception. It is also clear that on selecting an algorithm the type of data at hand, the purpose of the data mining task and other factors will determine the type of algorithm to use. By doing so, we can maximize the benefits that we get from different data mining techniques. Therefore, some of the results from different algorithms may depend on the researcher's effort on identifying which techniques best suits for the problem at hand. These efforts and considerations contribute for the benefits to out-weight the limitations. It is with this intention that the researcher is implementing decision tree.

3.2 Artificial Neural Networks

Neural networks are now popular in many areas including medical research, finance and marketing. This is due to their performance in predictive power compared to other statistical techniques. In addition to the discussion about artificial neural network under 2.2.3.3 the following are worth mentioning. Neural networks are broadly categorized as supervised and unsupervised neural networks based on their learning methods. Among the supervised neural networks multilayer perceptron (MLP) or radial basis functions falls

under this umbrella. In supervised neural network, a model is built using training and test data. The training data is used by the neural network to learn on how to predict the known output. On the other hand test data is used to validate the prediction accuracy (Cerny & Proximity, 2001).

3.2.1 How artificial neural networks work?

The basic building blocks of biological neurons receive inputs from external environments, process them in some fashion, nonlinear operation is performed on the result and finally output the result. The artificial neural network is designed to simulate the natural neurons and the basic representation is depicted in figure 3.2 as follows.

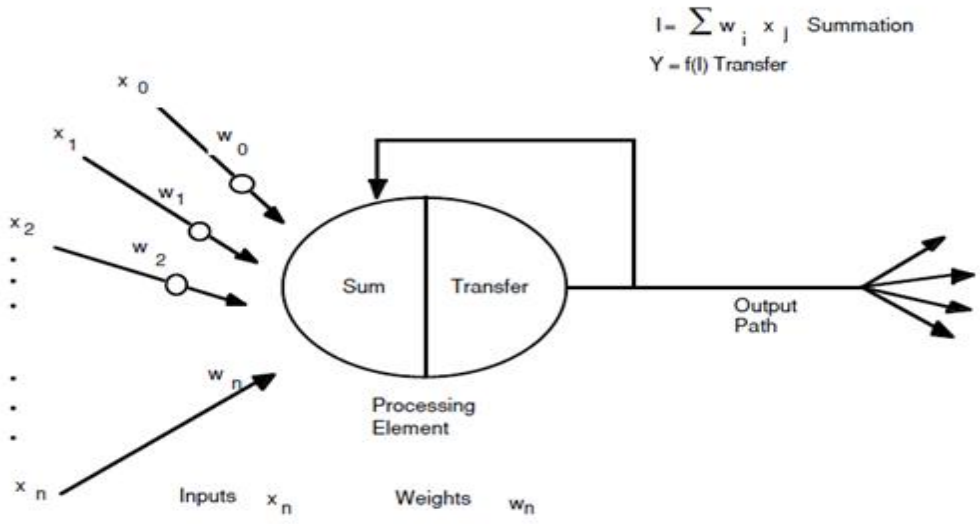


Figure 3. 2 Basic Artificial Neuron

In figure 3.2 the inputs are represented by X_n, each input (X_i) is multiplied by the weight W_i. For a simple case each input is multiplied with the corresponding weight and the products are summed to be fed to transfer function. Finally, the results are sent to output function for display or otherwise.

The two sections in the processing element, sum function and transfer function, process the input data using different mathematical computations and algorithms before sending the result to the output path. The transfer function, after accepting the output for the summation function as an input, turns this number into real output using some algorithms. This algorithm turns the input to a zero, one, minus one or some other numbers. Sigmoid, sine, hyperbolic tangents and others are the commonly supported transfer functions. This transfer function can also scale the output or control its value via thresholds. In most cases, the result of the transfer function is the direct output of the processing element.

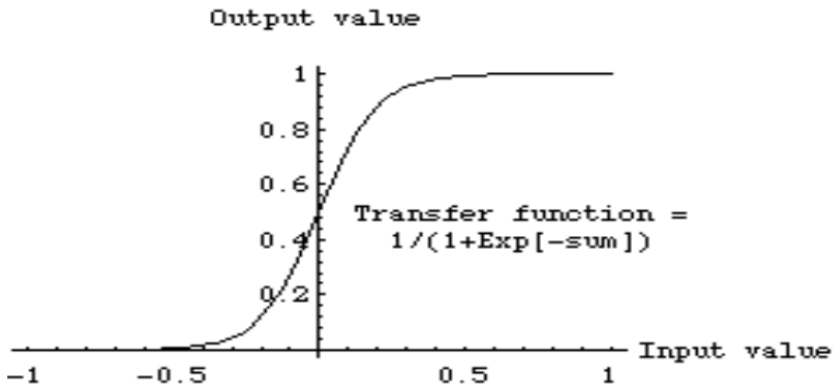


Figure 3.3 Sigmoid Transfer Function

The sigmoid transfer function, shown in figure 3.3, takes the value sum from summation function and turns into a value between zero and one. Lastly, the processing element is ready to output the result of its transfer function. This output is dictated by the structure of the network to be sent either as an input into other processing elements or to an outside connection (Anderson & McNeill, 1992).

In supervised learning, the patterns have to be a priori categorized as belonging to some class. During learning, the network tries to adapt its units in order to produce the correct label at its output for each training pattern. After

completing the training, the units are frozen and when a new pattern is presented, it is classified according to the output produced by the network. But in unsupervised learning the system will find meaningful patterns or clusters without the need to label the data (Burge et al., 1997).

3.2.2 Supervised learning

Multilayer perceptron is used for supervised learning purpose. They are arranged in layers namely input, hidden and output layers. The elementary units are called neurons (Burge et al., 1997). They are also known as processing elements (PEs) (Cerny & Proximity, 2001). A simple non-linear transformation of the inputs is produced by each neuron as an output depending on the value of the weights of the network, thus:-

$$y = \sigma\left(\sum_{i=1}^{n} w_i x_i + w_0\right), \quad \text{where } \sigma(z) = \tanh(z) \text{ or}$$

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Where x_i is the value on the i-th input line and W_i is the weight on that line.

The neurons are then arranged in a two-hidden-layer network (could be more than 2 neurons) with D inputs, H_1 hidden neurons or PEs in the first layer, H_2 hidden neurons or PEs in the second layer, and C outputs. The layers could be more than two depending on the problem type and data at hand to solve the problem. The outputs Z_m of the network is defined as:

$$h_{1k} = \sigma\left(\sum_{l=1}^{D} w_{kl} x_l + w_{k0}\right), h_{2_i} = \sigma\left(\sum_{k=1}^{H_1} v_{ik} h_{1k} + v_{i0}\right),$$

$$z_m = \sigma\left(\sum_{l=1}^{H_2} u_{lm} h_{2_l} + u_{l0}\right)$$

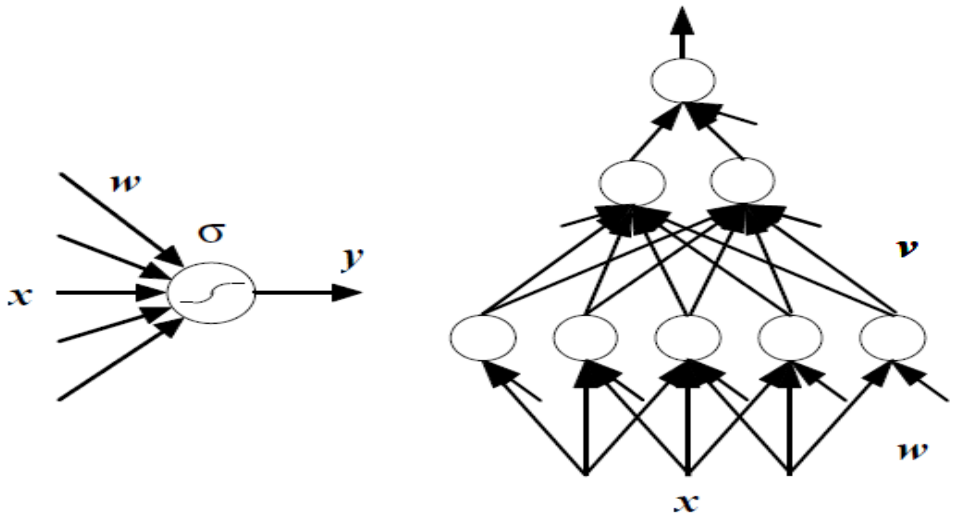


Figure 3. 4 Sigmoidal Neuron and Multi-layer Perceptron Architecture

The main property of multi-layer perceptron's, as figure 3.4 depicts the architecture, is that they can approximate any function of the input to an arbitrary degree of accuracy. This is true only if there are enough hidden neurons available. These hidden neurons are decided by the researcher in the case of supervised neural networks. They can achieve this approximation with a relatively small number of parameters.

For supervised learning, there is a need to organize the data available for design in a data set of labeled pairs $D = [(X_1, Y_1),\ldots,(X_K, Y_K)]$, where Y_K is the fraud label (Y_K= 0for normal behavior, Y_K= 1 for fraud) associated to the K_{th} pattern with features X_K extracted from the user profile (Burge et al., 1997).

The first step in training the neural network is choosing the number of layers and the number of neurons in each layer. Once we are done with the architecture, the output of the network is a function of its input X_K and of the parameters w (the weights) of the neural network. There is always a discrepancy between the output of the classifier $z(X_K, w)$ and the desired output Y_K. The training of the classifier consists of adapting the weights so as to

minimize this discrepancy. The measure of discrepancy is quadratic, where w is found such that E is a minimum, thus:-

$$E = \sum_{k=1}^{K} \left\| Y_k - z(X_k, w) \right\|^2$$

3.2.3 Applications of Multi-layer perceptron

Multi-layer perceptron has been applied in different areas and tasks. But the application areas can further be classified as prediction, function approximation, and pattern classification. Prediction refers to forecasting of future trends in time series data by providing previous and current data. On the other hand, function approximation deals with modeling the relationship between variables. The last application, pattern classification, involves classifying data into discrete classes.

Function approximation and prediction are similar in some ways. In applying multi-layer perceptron for prediction, there is a need to train the network to output the future value of a variable given the input vectors. These input vectors contain observations of earlier times. Multi-layer perceptron is known for approximating functions that are highly non-linear and without requiring prior knowledge of the nature of relationship. This is also mentioned as one of the benefits of multi-layer perceptron. The other benefit is its ability to perform well in noisy circumstances. Since the real world is characterized by nonlinear relationships, nonlinear regression is useful given non-linearity is consistent over the entire range of measurements (Anderson & McNeill, 1992).

Multi-layer perceptron is useful in many ways. First, it doesn't require prior assumptions regarding the distribution of training data. The other benefit is that no decision needs to be made about the relative importance of the input measurements. Because the most discriminating input measurements are made during training by adjusting the weights. Therefore, multi-layer

perceptron is considered to be superior to the traditional classification approaches due to the aforementioned benefits(Gardner & Dorling, 1998).

As Cerny & Proximity (2001) summerized the benefits and drawbacks of neural networks from forty studies that used neural network in their study. Among the benefits high accuracy, noise tolerance, independence from prior assumption, ease of maintenance,overcoming the drawbacks of other statistical methods, ability to be implemented in parallel hardware, minimized human intervention (highly automated), and suitability to be implemented in non-conservative domain are major ones. On the other hand, the limitations are poor transparency, trial and error design, data hangryness (requires large amount of data), over fitting, lack of explicit set of rules to select a suitable neural network, dependency on the quality and amount of data, lack of classical statistial properties (confidence interval and hypothesis testing) and the techniques are still evolving (not robust). Moreover, the black-box nature – difficulty of identifying the rules as to how the prediction or classification is made from the model derived by neural network, and sensitivity of neural networks to file or data formats are catagorized as the two major drawbacks of neural networks(Cerny & Proximity, 2001).

3.2.4 Feed Forward Neural Network

Feed forward neural network is used to calculate output values from input values, as indicated in figure 3.3. The network topology or structure used for feed forward neural network is the same as to that of prediction and classification. As mentioned earlier, the three network layers are input, hidden and output layers. In the first layer, the input layer, each unit in the input layer is connected to one source where the values are mapped ranging from -1 to 1. This layer is considered as an important part of neural network vocabulary. The input layer does not do anything except processing and mapping values to a reasonable range. It is also considered as a reminder of the very important aspect of using neural network successfully.

The next layer is hidden layer where each unit in the hidden layer is connected to all the units in the input layer. Then each unit in the hidden layer calculates its output by multiplying the value of each input unit with the corresponding weight, adding these up and applying transfer function. The hidden layer can have any number of hidden layers. Though, one is enough, the wider the layer the higher the capacity to recognize pattern. The hidden layer should not be too wide because the neural network opts to memorize rather than generalizing it. Therefore, we have to make sure that the number of hidden layers is optimal.

The last layer is known as output layer because it is connected to the output of the neural network. Output layer is fully connected to all units in the hidden layer. In most cases, a neural network is being used to calculate a single value and in such scenario the output layer will only have one unit. This time the output layer implements a simple linear transfer function and the output will be a weighted linear combination of inputs. This will also avoid the need for mapping the output. In other scenarios, it can have more than one output unit depending on the purpose the neural network is used. For this purpose we have different transfer functions like sigmoid and hyperbolic tangent function, besides the one mentioned above (M. J. Berry & Linoff, 2004).

3.2.5 Back Propagation

Back propagation, even though it is the old method, was the original method for training feed forward neural network. It is used for adjusting the weights by comparing the output of the feed forward neural network against the desired output. Training neural network, is the process of setting the best weight on the edges connecting all the units in the network. Back propagation provides an insight on how the training works. The following are the steps used in training the back propagation:

1. *The network gets a training example and, using the existing weights in the network, it calculates the output or outputs.*

2. *Back propagation then calculates the error by taking the difference between the calculated result and the expected (actual result).*
3. *The error is fed back through the network and the weights are adjusted to minimize the error—hence the name back propagation because the errors are sent back through the network (M. J. Berry & Linoff, 2004).*

The overall error of the network is obtained by comparing the output from the training set against the actual value using back propagation algorithm. It adjusts the weights to minimize the error but not to eliminate it. The process of error correction will not stop here but the blame goes back to the earlier nodes in order to reduce the overall error. In general, it can be said that back propagation uses complicated mathematical procedure that requires taking partial derivative of the activation function to accomplish its task.

As it is true in any of the training techniques, there is a problem of falling into something called a local optimal solution. This happens when the network performs best for the training set and changing the weights will no longer change the performance of the network. In such scenario, combinations of weights are used that yield much better solution for the network. It is also possible to use techniques like hill climbing. In addition, controlling the learning rate and momentum help us find the best solution *(M. J. Berry & Linoff, 2004).*

Chapter Four

Data Preparation

This chapter focuses on data preparation process starting from data understanding, initial data collection, data description, data preparation, data quality assurance, data integration and transformation up to data formatting. As mentioned in chapter one, CRISP-DM process model is followed in order to come up with the desired output.

4.1 Ethical Standard

As mentioned in section 1.7.3, the data for this research is obtained from different sections of ethio telecom by using a cooperation letter from Addis Ababa University. The letter was directed to concerned departments from chief executive officer (CEO) of the company.

4.2 Understanding of the Data

The data for this research is obtained from Information Technology (IT) operation division of ethio telecom. It is found in Z smart CCB database which is one of the projects implemented by ZTE Corporation (China Telecom Company) at the cost of 1.5 billion dollars. It is a huge database system that replaced the previous 'USHACOM' system for CCB (Melaku, 2009). This billing database system is used to process all billing data of the services that the company is providing. In addition, the data source sections for this specific research are Billing Operation Section for the CDR and OCS, VAS & ISP Section for the OCS data in IT operation division.

Both CDR and OCS data collected are used for this research. The voice CDR, as the name implies, holds each and every record of calls made by the customer with details like calling number, called number, date and time of call, duration, amount charged, MSC number, location of calling number and other

details. SMS and GPRS CDRs are also detail records for the mentioned services. Additionally, the OCS data holds information for pre-paid customers like balance/money recharged, time and date recharged, amount of money recharged and other information. The location data found on CDR enabled the researcher to get the HLR data without the need to request the data separately from the department concerned. This location data is found in the CDR with attribute field “CELL_A” or “LAC_A” that tells the BTS number. This BTS number (Cell number) indicates the place where the customer initiated the call and operators have detail information about each BTS.

The CCB database is implemented on Oracle 11g and users accessing the database can access customer profile and CDR data. CDR data are available for the past six months only. Data older than that are sent to data warehouse of the company. The users of this database are sales, collection, marketing, customer service and others. Except, the database users the remaining are accessing the database using web-based application interface called Z-smart CCB. The different modules in the application are provided to users according to the need for the department they are in. The other data is the OCS data which is found in the same department. There are six servers that are dedicated for this data. Each server is dedicated for a range of prepaid mobile and other service numbers. These servers are found in two different locations and they hold OCS data only for the past three months. From these servers the researcher found data only for the months January, February and March 2013. Regarding the CDR, data from December to March 2013 are taken for this research purpose.

The CDR data used for the research was so bulky. One month CDR is more than 280 GB (Giga byte) of data and storage was a big challenge before thinking about processing the data. Then samples are taken based on call duration. The detail is found under data reduction section (4.3.3).

4.2.1 Initial Data Collection

Due the size of the data, different query techniques are applied in order to reduce the size. The database specialists from the billing operation section queried the database based on the requirement given by the researcher. Their advice on limiting the size of the data was considered in the data collection process. From the CCB database four months data were taken which is more than eight million records. From these records 11,592 are taken for this research by carefully sampling the data.

4.2.2 Description of Data Collected

The CCB database has more than three hundred tables among these the table for prepaid mobile, SMS, GPRS and the OCS database recharging tables were selected for this research. The remaining tables are not selected for they have no direct relation with this study. The following table, table 4.1, describes the attributes selected with their corresponding data type and description. The prepaid mobile table has 34 attributes and only 10 of them are selected by consulting different articles (Abidogun, 2005; Burge et al., 1997; Ferreira, Alves, Belo, & Cortesão, 2006), based on the researcher's 12 years' experience in telecom sector and the information gained from fraud experts. These 10 attribute fields have 3 repeated fields with different attribute names like LAC_A and CELL_A. This implies only 7 attribute fields are taken from voice CDR table. Among the 34 attributes 13 of them have either no data or zero value (both for security reason and they are only reserved for future use), 2 other fields have missing values (254 values out of 1 million records), and the remaining fields have either one constant value only or other numbers that are not useful for this study. These fields are removed from the data based on domain experts' advice and literatures support about tolerable or acceptable level of missing values (Acuna & Rodriguez, 2004). The following table focuses on describing the selected and derived attributes only.

No	Field name	Data type	Description
1	BILLING_NBR	Number	Billing number (Mobile number to be charged). Mostly it is the same as calling number. For privacy reason this number is changed to four digit starting from 1010.
2	START_TIME	Date	Time of call initiation (calling time). This field value is changed to PH, OP and OP_WE.
3	DURATION	Number	Call duration in seconds
4	CHARGE	Number	Amount paid in cents
5	CALLING_NBR	Number	Mobile number initiating or originating the call. It is the same as billing number.
6	CALLED_NBR	Number	Mobile number receiving the call.
7	LAC_A	Number	Mobile BTS location sector_ A number (BTS-ID) where the call is originated (same as CELL_A).
8	CELL_A	Number	Mobile BTS cell sector_A number (BTS-ID) where the call is originated (same as LAC_A).
9	MSC	Number	Mobile Switching Center where the BTS's send the initiated call via BSC.
10	FEE1	Number	Amount paid in cents (same as CHARGE).
11	Sub_Age*	Number	Subscription age of customer. It ranges from 1 to 4. Because the data is taken for 4 months (Dec to Mar) only.
12	OCS*	Number	Online charging system (OCS). Derived from OCS CDR table.
13	SMS*	Number	Average number of SMS made by the service (mobile) number. Derived from SMS CDR table.
14	NBR_Of_GPRS*	Number	Average number of GPRS connections made by the service (mobile) number. Derived from GPRS CDR table
15	Avg_Monthly*	Number	Average number of calls made during the months. Derived from voice CDR table.
16	Calls_Per_Day*	Number	Average number of calls per day (avg_monthly/30). Derived from voice CDR table.
17	Call_Inter*	Number	Average number of international calls made. Derived from voice CDR table.
18	Call_ratio*	Number	Ratio of calls made (unique number of calls made over total number of calls made). Derived from voice CDR table.
19	C_type*	Char	Call type (fraudulent or non-fraudulent). FRD refers to fraudulent call type and NFR for non-fraudulent ones
NB: * refers to derived attributes or taken from other DB/Table other than voice CDR.			

Table 4. 1 Attribute Fields, Data Types and Description

4.2.3 Data Quality Assurance

The CDR data obtained from the information system office specifically customer care and billing section has almost no missing value. The missing

values are found in only 2 attribute fields and this is happening due to the nature of call. The value in these attributes, CDR type and third party number, are recorded when the calling number and the billing number (bill paying number) are different. Such kind of calls are rare, it is observed from the data that only 254 calls from 1 million have such type of record. Therefore, it is possible to say the data is complete.

In relation to the relevance of the data, all the CDR data were not relevant for this study. The researcher tried to filter the relevant data from the large database based on the total number of calls made and total call duration (talk time) during the month. For the mentioned four months a representative sample is taken for this study. In addition, the domain experts' contributions in determining the number of attributes were very helpful.

4.3 Data Preparation

Data is the reason for the need and importance of data mining. Without it there is no value to discuss and worry about data mining. Data preparation or data preprocessing is a crucial activity for the next step: - modeling. According to Han and Kamber (2000) preprocessing is useful for improving the quality of the data so as to improve efficiency, mining process and result of data mining. It is important in order to deal with incomplete, inconsistent and noisy data. Under preprocessing, tasks like data cleaning, data integration, transformation and data reduction techniques are addressed.

4.3.1 Data Cleaning

As mentioned earlier, the first data cleaning is made by way of reducing non mobile calls, calls made by fixed line numbers and CDMA (wireless) numbers. The next data cleaning is made by removing attributes having similar (fixed) values like zeros or ones and other constant values. In addition attribute fields containing not allowed values are removed and mobile numbers are also changed with other value to respect privacy right of customers. Redundant

value containing attributes, like billing number and calling number, only one of them is used. For instance, the billing number is 911640000 and the calling number is 251911640000. Due to this fact, only billing number is selected. In addition, non-relevant attributes based on information from related literatures and from domain experts are also removed. Due to this from the total of 34 attributes from CDR table only 7 attributes are selected. And additional 8 attributes are used, including the target variable, for this study. From the additional 8 attributes 5 of them are derived and the rest 3 attributes are summaries from OCS, SMS and GPRS tables. But the total numbers of attributes are 15 including derived attributes. The whole attributes are listed in Appendix X with a short explanation as to why the attribute is selected and with sample data.

4.3.2 Data Integration and Transformation

Data integration is made for start time, number of calls per day, SMS, GPRS, average total number of monthly calls and international calls. The start time is transformed as peak hour (PH), off peak hour (OP) and off peak hour week end (OP_WE). This is just by taking the peak and off-peak hour times of the company. The aggregate value for each attribute is average value of the four months. Mobile numbers with less than 4 months of subscription duration or Sub_Age are also given due attention in calculating average values.

4.3.3 Data Reduction

The size of the data was originally big requiring a server for processing it separately. One month compressed data was more than 280 GB and for this research at least 3 months data was needed to take representative sample and properly study the trend. By discussing with the database specialists, a script that reduced the size of the data is written. Query reformulation incorporates criteria's like call duration, usage rate (money recharged each month), number of calls made, GPRS and international calls made. By doing so manageable size

of data (7 megabyte) is obtained and further data reduction is made based on sampling.

The CDR obtained by querying the database to display CDRs with call duration greater than 100,000 seconds (talk time) per month was the initial step to reduce the data. This criterion incorporates both fraudulent and non-fraudulent ones. This is done with analysis on sample mobile numbers that the researcher identified and cross checked on Z-smart (web based application used to access CCB and OCS data) to see the trend in the system.

The data obtained using the above query is further reduced using excel filtering technique to exclude calls made to short numbers like 929, to exclude calls made by wireless prepaid services and calls with zero charge. After getting only mobile prepaid numbers sampling is made by dividing each month data into weeks, each weeks data is further divided as week days and weekends data. Each week days and weekends data is further divided as peak hour and off-peak hour. After all these classification the first 250 records are taken, 125 from peak hour and 125 from off-peak hours for weekends calls. Similarly, 500 records of calls are taken that are made during week days. The sampled data for one week is 750 records and 3000 records for a month. Total number of records taken as a sample for this study is 12,000. After further reduction of missing values, the remaining total record is 11,592. This reduction is not exactly missing values but some attributes like called number and CELL_A (location data) are not registered for special individuals or organizations. The main reason for this is security and special request from organizations. Therefore, these records are excluded from the sample.

Additional data like total number of SMS, GPRS, number of calls made and international calls are queried by providing the mobile numbers which fall in the sample. The result of the query return values for mobile numbers that used the indicated services. If the number is not found in the list, it implies that the subscriber did not use the service during that month.

4.3.4 Data Formatting

Before dealing with the data modeling the data set has to be formatted in a manner that suits the tool to be used for modeling. In this study WEKA 3.6.9 is used which requires file formats like comma delimited CSV, ARFF and the like. The researcher preferred to use the CSV file format because the Oracle 11g database provided the data in such format. Using CSV file format enabled the researcher to use directly what is queried from Oracle database after applying the above data reduction techniques.

Attribute selection

In this regard, attribute selection is made based on WEKA attribute selection technique, other researchers' recommendations in the domain area and domain experts' information. This selection method is applied for all algorithms used in this study.

Facts from sampled data

By applying sampling, a total of 11,592 records are selected for this study. The different algorithms from classification used the sampled data to build different models for detecting fraudulent calls using SIM box to terminate international calls. In this section, some facts or figures from the sampled dataset is discussed for further analysis in this study. From the CDR description, call duration or 'duration' attribute field contains the duration of the call in seconds. For this attribute the minimum, average and maximum values are 1, 251.23 and 12870 seconds respectively. Similarly, charge CDR description attribute field holds the amount paid by the billing mobile number in cents. Under this attribute, the minimum, average and maximum values/cents paid by the user are 0, 880.04, and 48,328 cents respectively. This attribute has exactly similar value with FEE1 attribute.

The OCS derived attribute is derived from OCS CDR Database table from IN CCB database. The value in this field is what the customer recharged in terms of birr. The other derived attribute field is calls per day that hold average number of calls per day made by each mobile number. The minimum, average and maximum numbers of calls per day are 8, 47 and 150 respectively. The last derived attribute for discussion under this section is call ratio. It holds the total number of unique numbers called divided by the total number of calls made. It is meant to show the call dispersion rate of the subscriber. The minimum call ratio is 0 and average and maximum call ratio is 1.

Chapter Five

Experimentation and Modeling

5.1 Modeling

In the modeling phase of this research classification is used by applying different algorithms. The algorithms include tree based decision tree, rule based PART and function based multilayer perceptron. Multilayer perceptron is from neural network. The rules generated from classification using decision tree algorithm, specifically J48 and PART, are among the required output for predicting the fraudulent calls made via SIM boxes. In addition, the output from the neural network, multi-layer perceptron-WEKAs implementation of artificial neural network, is also used in this research. But the models obtained from multi-layer perceptron, PART and J48 decision tree algorithm are used for predicting fraudulent scenarios. Interesting rules are expected from J48 and PART due to the nature of the algorithms results.

The experimentation is made using WEKA data mining tool version 3.6.9. Different experiments are made using 15 attributes (with all determinant attributes), using only non-determinant attributes and using WEKA selected 4 attributes. The experimentation is made also by gradually reducing the number of determinant attributes in the domain area. Such experiments are conducted using J48, PART and multilayer perceptron algorithms. Different parameters are also adjusted to get an optimal result using each algorithm. Finally, the model with the best accuracy is selected by comparing the resulted models from the above three algorithms. A print screen image that shows the selected attributes in WEKA data mining tool interface is attached in Appendix IV.

5.1.1 Classification Modeling

The classification models are of three types in this study. The first model is using J48 algorithm from tree based, the other is using PART algorithm from rule based and the last classification modeling is using multi-layer perceptron algorithm from neural network. The algorithms for both decision tree and neural networks are tested using different parameters and the sampled dataset. Experimentations are conducted using the three algorithms to come up with the best predictive model for fraud detection. Finally, comparison among the best selected models is made to see and propose the best one for fraud prediction purpose.

5.1.1.1 Decision Tree modeling

As it is thoroughly discussed in section 2.2.3.3, decision tree is the most commonly used for prediction and classification purposes. On top of this, it does not require prior information about the data to be classified or predicted and the rules are also used for prediction purpose. It is also known for providing an easily interpretable solution (Bresfelean, 2007). In this section, experiments are conducted using J48 and PART with selected 15 attributes including the determinant variables in the domain area.

Experimentation 1 Using J48 Algorithm with All Attributes

This experiment for decision tree is conducted using different test options namely percentage split, cross-validation and use training set. For each test option parameters are changed to see the effect. The best performing models from each test mode is presented in table 5.1 below. But, the detail experimentation results using J48 algorithm with different number of attributes ranging from 15 to 9 by varying test options and parameters are summarized in Appendix I.

Experiments	Algorithm /function	Number of attributes	Number of leaves	Size of trees	Test modes	Time taken to build the model (sec)	Accuracy (%)
1	J48 (with all selected attributes)	15	11	21	Cross-validation 10-fold	3.3	99.9224
		15	11	21	Percentage split 66%	3.26	99.9746
		15	11	21	Use training set	3.26	99.9827

Table 5.1 Result of Decision Tree J48 Models Using Different Test Modes

In the above experiment table 5.1, the resulted models summaries using 15 attributes are presented. The change in the test mode has effect on the accuracy of the model, time taken to build the model and also on the size of the decision tree. The models resulted by changing the classifier test options, using training test, cross-validation and percentage split, are used to compare the models with each other.

The best result from the experimentations using 15 attributes is using training test mode. Its overall accuracy is 99.9827% with number of leaves 11 and tree size 21. It is the smallest number of leaves and tree size using decision tree for this study. The time taken to build the model is 3.26 seconds. The summary indicates that out of 11,592 records 11,590 (99.9827%) are correctly classified and the remaining 2 records are incorrectly classified which is 0.207%. The detail accuracy by class shows that ROC area and recall are all one for both fraud and non-fraudulent instances. On the other hand, precision and F-measure are 0.999 for fraudulent instances and 1 for non-fraudulent cases. When we come to the confusion matrix result, from the total of 9626 non-fraudulent records all (100%) are correctly classified and no record is incorrectly classified. From the total of 1966 fraudulent records 1964

(99.8983%) records are correctly classified and 2 (0.1017%) records are wrongly classified as non-fraudulent.

The second best model result is obtained using percentage split 66% as training and 34% as test dataset. In this experiment the parameters are set to the default. The overall accuracy for this model is 99.9746% with 11 leaves and 21 tree size. From 11592 records 3941(34%) records are separated as test set by using the algorithm test mode. From these records 3940 (99.9746%) of the records are correctly classified. The remaining 1 record (0.0254%) is incorrectly classified. From the detail accuracy by class, we can understand that ROC area, F-measure, recall and precision are 1 for non-fraudulent records. On the other hand, for fraudulent instances F-measure and precision are 0.999 but Roc area and recall are 1. As the confusion matrix indicates 695 (99.8563%) is correctly classified from the total of 696 fraudulent records and only 1 (0.1437%) record is wrongly classified. Concerning non-fraudulent records 3245 (100%) are classified accurately and no record is inaccurately predicted.

The third item in the summary table is test mode result of using cross validation 10-fold. In this experiment 11583 (99.9223%) records are correctly classified from the total of 11592 records. Only 9 records (0.0777%) are incorrectly classified. This accuracy figure shows the overall accuracy for the model but the detail accuracy by class shows: the true positive and false positive rates are 0.999 and 0.001 respectively. For non-fraudulent instances precision, F-measure and ROC area are 1 but recall is 0.999. On the other hand, for fraudulent instances the true positive and false positive rates are the same as non-fraudulent ones. But precision, recall, F-measure and ROC area are 0.996, 0.999, 0.998 and 1 respectively. The model resulted with 11 leaves and 21 tree size. Moreover, from the confusion matrix, we can understand that from the total of 9628 non-fraudulent records 9621 (99.9273%) records are correctly classified and the remaining 7 (0.0727%) records are wrongly classified. On the other hand, 1962 (99.8982%) records are correctly classified

from the total of 1964 records and the rest 2 (0.1018%) records are incorrectly classified.

From the discussions above, it is clear that the experiment conducted by using use training set test mode resulted in 99.9827% overall accuracy. This accuracy is taken as the best result using decision tree, J48 algorithm. The resulting confusion matrix for the best model is presented in table 5.2.

Actual	Predicted		Total	Accuracy
	Fraud	Non-fraud		
Fraud	1964	2	1966	99.8983%
Non-fraud	0	9626	9626	100%
Total	1964	9628	11592	99.9827%

Table 5.2 Confusion Matrix Result for J48 Algorithm Using Training Set

Experimentation 2 Using J48 with WEKA Selected Attributes

Here, WEKA's feature for selecting attributes is used and only 4 attributes are selected from the total of 15 attributes. The following experiment is conducted using these four attributes and default parameters. It is meant for comparison of the results proximity with that of the selected attributes. In this section only the best model is discussed since the purpose of this experiment is to compare the proximity of the result with other models.

Experiments	Algorithm /function	Number of attributes	Number of leaves	Size of trees	Test modes	Time taken to build the model (sec)	Accuracy (%)
1	J48 using WEKA selected attributes	4	15	29	Cross validation 10 fold	1.24	99.9482
		4	15	29	Percentage split 66%	0.79	99.8731
		4	15	29	Using training set	0.78	99.9655

Table 5.3 Experimentation Result Using J48 with WEKA Selected Attributes

This experiment is made using J48 tree based classification algorithm. The best result of the experiment using WEKA selected attributes shows that by using training set test mode 99.9655% overall accuracy is attained. Only 4 (0.0345%) records are incorrectly classified from the total of 11592 records. The model has 15 numbers of leaves and 29 tree size. The time taken to build the model is 0.78 seconds. The detail accuracy by class shows that true positive rate, false positive rate, precision, recall, F-measure and ROC area are 1, 0.002, 1, 1 and 1 respectively. On the other hand, the confusion matrix result shows that from the total of 9628 non-fraudulent records 9627 (99.9896%) records are correctly classified. The remaining 1 record is incorrectly classified. Additionally, from the total of 1964 fraudulent records 1961 (99.8473%) records are correctly classified and the rest 3 (0.1527%) records are wrongly classified as non-fraudulent. The result from WEKA selected 4 attributes for the rest test options are summarized in table 5.3 above.

Experimentation 3: Trend Experiment Using J48 Algorithm

Trend experiment using J48 algorithm is made by decreasing the number of attributes from 15 to 9. The trend using 15 to 10 attributes shows a maximum difference of 0.1% and the minimum is 0.02% from percentage split and using training set test option. But, the trend from 15 to 9 attributes resulted in 0.72% and 2.51% from using training set and percentage split respectively. The attributes gradually removed from the selected attributes are SMS, GPRS, international call made, OCS, Sub_Age and number of call per day. The result from trend experiment shows that these attributes on voice CDR are relevant in identifying fraudulent calls. For further detail please refer Appendix VIII.

Experiment 4 Using Part Algorithm with All Attributes

In this experiment, PART algorithm from classification which is rule based is used. Using this algorithm different experiments are made by applying different

test modes and parameters. The resulted models are summarized in Appendix II but the top 3 models summary is presented in table 5.4 below. The models resulted using 15 attributes are discussed in the following paragraphs.

Experiments	Algorithm /function	Number of attributes	Number of Rules	Test modes	Time taken to build the model (sec)	Accuracy (%)
1	PART algorithm	15	7	Cross validation 10 fold	3.61	99.9396
		15	**7**	**Percentage split 66**	**2.95**	**100**
		15	7	Using training set	3.33	99.9827

Table 5.4 Top 3 Models Using Part Algorithm

The first top experimentation result is using percentage split 66 percent test mode. The model resulted in 100% overall accuracy and the time taken to build the model is 2.95 seconds. Seven rules are generated from the model, which is the smallest from the entire experimentation result using PART algorithm. The algorithm assigned 3941 (34%) instances for test purpose and the result shows that all fraudulent and non-fraudulent instances are correctly classified with no wrong assignment. In addition, the detailed accuracy by class is 1 for all true positive rate, precision, recall, F-measure and ROC area for both fraudulent and non-fraudulent instances. The false positive rate value and incorrectly classified instance value in the confusion matrix is 0.

The second top experimentation resulted model is using training set test mode. The model resulted with 99.9827% overall accuracy and only 2 instances are incorrectly classified from the total of 11592 records. The number of rules generated by the model is 7 and 3.33 seconds are required to build the model. On the other hand, the detail accuracy by class shows that true positive rate,

precision, recall, F-measure and ROC area for non-fraudulent records is one but false positive is zero. Regarding fraudulent instances true positive, recall and ROC area are one but false positive rate, precision and F-measure is 0, 0.999 and 0.999 respectively. Moreover, the confusion matrix indicates that from the total of 1966 fraudulent records 1964 (99.8983%) records are correctly classified and the remaining 2 (0.1017%) are wrongly classified as fraudulent. Concerning, the non-fraudulent records all are correctly classified.

The top third model is using 10-fold cross validation test mode with overall accuracy of 99.9396%. Using this model from the total of 11592 records 11585 records are correctly classified and the remaining 7 (0.0604%) records are incorrectly classified. The model resulted in 7 rules and time taken to build the model is 3.61seconds. The detail accuracy by class result shows that true positive rate, false positive rate, recall and ROC area are 0.999, 0.001, 0.999 and 0.999 respectively for both fraudulent and non-fraudulent class. The precision and F-measure for non-fraudulent is 1 but it is 0.997 and 0.998 for fraudulent class respectively. The confusion matrix also indicate that from the total of 9628 non-fraudulent instances 9623 (99.9481%) records are correctly classified but the remaining 5 (0.0519%) records are wrongly classified as fraudulent. Additionally, from the total of 1966 fraudulent instances 1964 (99.8983%) instances are correctly classified but the rest 2 (0.1017%) records are not correctly classified.

Therefore, the model result using 15 attributes and by applying PART algorithm using percentage split 66% test option managed to score 100 percent accuracy when tested by 34 percent of the records.

Experimentation 5 Using Part Algorithm for WEKA Selected Attributes

By applying the feature for selecting attributes that uses different evaluation and search techniques. The supervised attribute selection technique resulted in four attributes among the 15 attributes given for the tool. This experiment is

conducted to compare the models obtained with more attributes. In this section only the best model is discussed to compare it with other best model. Table 5.5 presents summary of resulted models using different test modes.

Experiments	Algorithm /function	Number of attributes	Number of Rules	Test modes	Time taken to build the model (sec)	Accuracy (%)
1	PART using WEKA selected attributes	4	12	Cross-validation 10-fold	0.89	99.931
		4	12	Percentage split 66%	1.32	99.8224
		4	12	Using training set	0.87	99.9655

Table 5.5 Part Algorithm Resulted Models Summary Using WEKA Selected Attributes

The best model using WEKA selected four attributes by applying default parameters is using training set test mode. The model resulted in 12 rules and 0.87 seconds are required to build the model. It resulted 99.9655% overall accuracy. From the total of 11592 records 11588 records are correctly classified and the remaining 4 (0.0345%) instances are wrongly classified. The detail accuracy by class result indicates that true positive rate, precision, recall, F-measure and ROC area are 1 for non-fraudulent class. But, false positive is 0.002 for non-fraudulent and 0 for fraudulent class. Additionally, true positive rate, precision, recall, F-measure and ROC area are 0.998, 0.999, 0.998, 0.999 and 1 respectively. On the other hand, the result from the confusion matrix show that from the total of 9628 non-fraudulent instances 9627 (99.9896%) records are correctly classified but only 1 instance is wrongly classified. From the total of 9664 records 9661 (99.8473%) records are correctly classified but the remaining 3 (0.1527%) instances are wrongly classified as non-fraudulent.

Experiment 6: Trend Experiment Using Part Algorithm

Trend experiment is conducted by gradually reducing from the selected 15 attributes. The default parameters such as confidence factor, minimum number of object, number of fold and seed value are set to 0.25, 2, 3 and 1 respectively. The test options used in this experiment are use training set, 10-fold cross validation and percentage split 66%. The maximum accuracy level difference is observed using 15 to 10 attributes is 0.1 and the minimum is 0.01 from 10-fold cross validation and using training set respectively. But, using attributes 15 to 9 maximum difference is 3.86% from percentage split and minimum difference is 2.3% from use training set test option. The result from trend experiment shows that these attributes on voice CDR are relevant in identifying fraudulent calls. For further detail please refer Appendix VIII.

Best Rules from Decision Tree Algorithms

Some of the rules derived from the models are presented below and some interpretation is given for some of the rules.

Rule 1

```
If MSC <= 251911299721 AND Calls_per_day > 67: FRD (353.0)
```

There are 353 fraudulent cases where MSC number is less than or equal to 251911299721 and call per day is more than 67

Rule 2

```
If CALLED_NBR > 251111231603 AND BILLING_NBR_CHD <= 1811 AND
Calls_per_day > 68: FRD (517.0)
```

Rule 3

```
If CELL_A <= 636014000745242 AND BILLING_NBR_CHD > 2184 AND
CALLED_NBR > 251910025125: FRD (107.0)
```

```
Rule 4
```

```
If BILLING_NBR_CHD > 2081 AND BILLING_NBR_CHD <= 2094 AND
CALLED_NBR > 251910005640: FRD (291.0)
```

There are 291 fraudulent calls between mobile numbers 2081 and 2094

Rule 5

```
If BILLING_NBR_CHD > 2184 AND BILLING_NBR_CHD <= 2274 AND
MSC > 251911299701 AND CALLED_NBR > 251910048180: FRD (127.0)
```

There are 127 fraudulent calls between mobile numbers 2184 and 2274

Rule 6

```
If BILLING_NBR_CHD > 1876 AND CELL_A <= 636012004936057 AND
Sub_Age > 3 AND BILLING_NBR_CHD > 1941 AND CELL_A >
636012004926058: FRD (22.0)
```

Rule 7

```
If CALLED_NBR > 251118201949 AND BILLING_NBR_CHD > 1477 AND
CELL_A > 636012001422326 AND Sub_Age > 1 AND
CALLED_NBR > 251910038484 AND BILLING_NBR_CHD <= 2081 AND
CELL_A > 636014000745242: NFR (243.0)
```

Rule 8

```
If MSC > 251911299702 AND BILLING_NBR_CHD > 1477 AND
CHARGE > 1246 AND Call_Time = PH AND Sub_Age > 2 AND
MSC > 251911299722 AND CELL_A > 636012001422326 AND
BILLING_NBR_CHD > 1874: FRD (5.0)
```

Rule 9

```
If Call_Time = OP AND CELL_A <= 636014000845102 AND
BILLING_NBR_CHD > 2067 AND CALLED_NBR > 251116298888 AND
Sub_Age <= 2 AND
MSC <= 251911299702 AND CELL_A <= 636014000845102: FRD (33.0/4.0)
```

In MSCs 251911299701 and 02 and billing number greater than 2067 and BTS numbers not more than 636014000845102 there are 33 fraudulent instances and 4 non-fraudulent instances.

Rule 10

```
If Call_Time = OP AND MSC > 251911299701 AND BILLING_NBR_CHD > 1375
AND BILLING_NBR_CHD > 1387 AND Sub_Age <= 2 AND MSC <=
251911299702: FRD (62.0/23.0)
```

In MSCs 251911299701 and 02 and billing number above 1375 and new subscribers (not more than 2 month of subscription time) there are 62 fraudulent and 23 non-fraudulent calls.

Rule 11

```
If Call_Time = PH AND CELL_A > 636014000446492 AND MSC <=
251911299702: FRD (35.0/9.0)
```

Rule 12

```
If CELL_A > 636014000745242 AND BILLING_NBR_CHD <= 2175 AND
CELL_A > 6.36014000947427E14 AND CHARGE > 72 AND
BILLING_NBR_CHD <= 2083: FRD (8.0)
```

Rule 13

```
If CELL_A > 636014000845106 AND BILLING_NBR_CHD <= 2175: FRD (6.0)
```

Rule 14

```
If CELL_A > 636014000745242 AND BILLING_NBR_CHD > 2184 AND
MSC > 251911299701 AND Call_Time = PH AND
BILLING_NBR_CHD <= 2190: FRD (38.0)
```

Mobile numbers between 2184 and 2190 and not in MSC 251911299701 and BTS number greater than 636014000745242 there are 38 fraudulent calls.

Rule 15

```
If call_inter <= 3 and call_ratio <= 0.98 and Calls_per_day <= 73
and Sub_Age <= 2: FRD (21.0)
```

Rule 16

```
If call_inter <= 3 and call_ratio <= 0.98 and Calls_per_day <= 73
and Sub_Age > 2 and Calls_per_day > 63 and BILLING_NBR_CHD <= 1362:
FRD (3.0/1.0)
```

Rule 17

```
If call_inter <= 3 and call_ratio <= 0.98 and Calls_per_day > 73:
FRD (343.0)
```

Rule 18

```
If call_inter <= 3 and call_ratio > 0.98: FRD (1584.0)
```

Rule 19

```
If call_ratio > 0.46 and BILLING_NBR_CHD <= 1839 and
BILLING_NBR_CHD <= 1811 and Calls_per_day > 68: FRD (517.0)
```

There are 517 fraudulent calls using mobile numbers between 1811 and 1839

Rule 20

```
If call_ratio > 0.46 and BILLING_NBR_CHD > 1811 and CELL_A <=
636012001422326 and Calls_per_day > 104: FRD (4.0)
```

There are 4 fraudulent calls with calls per day above 104 and mobile number above 1811

```
Using training set 11 attributes

Rule 21
If call_ratio <= 0.99 and BILLING_NBR_CHD <= 1479 and CELL_A <=
636012001412327 and Sub_Age <= 3 and BILLING_NBR_CHD > 1409 and
call_ratio > 0.98 and MSC > 251911299721 and Sub_Age > 2 and
DURATION > 97: FRD (28.0/1.0)
```

There are 28 fraudulent calls and 1 non-fraudulent call mobile number between 1409 and 1479 and BTS number not more than 636012001412327 and subscriber age between 2 and 3 months and call ratio between 0.98 and 0.99

```
Rule 22
If call_ratio <= 0.99 and BILLING_NBR_CHD > 1839 and
BILLING_NBR_CHD <= 1854: FRD (158.0)
```

There are 158 fraudulent calls between mobile numbers 1839 and 1854 and call ration not more than 0.99

5.1.1.2 Artificial Neural Network Modeling

Here artificial neural network model is used for classification purpose. The multilayer perceptron classifier algorithm uses back propagation to classify instances. For this specific research, it is intended to classify the dataset as fraudulent or non-fraudulent. One of the tasks in artificial neural network is to normalize the data, in a range of -1 to 1, to make it suitable for the algorithm. This task is performed using WEKA preprocessing facility to normalize the attributes' values in the mentioned range.

WEKA normalizing preprocessing facility has been applied on the dataset to normalize the values between the range -1 and 1. All the values are numeric except the target and time of call attributes. The target attributes' values FRD (fraud) and NFR (non-fraud) remained as they are when the dataset is normalized using WEKA. In the same fashion time of call attribute's values PH (peak hour), OP (off-peak hour) and OP_WE (off-peak hour weekend) are given as they are but the normalizer changed them in the given range.

Experimentation 7 Using MLP Algorithm with All Attributes

Multilayer perceptron (MLP) is WEKA's implementation of artificial neural network classification algorithm. In table 5.6, experimentation results using 15 selected attributes resulted models are summarized. The resulted models using MLP algorithm scored different accuracy level 15 attributes. In the following discussion only the top 3 scoring models are discussed from the summary in table 5.6 below.

Using MLP algorithm many experiments are conducted by using different test options with number of attributes ranging from 15 to 9. In addition, these experiments are made with and without determinant variables in the domain area. Additional experiments using WEKA selected attributes are also conducted to compare the accuracy level and rules derived against the selected 15 attributes. The detailed summarized results for MLP algorithm are found in Appendix III.

Experiments	Algorithm /function	Number of attributes	Test modes	Time taken to build the model (sec)	Accuracy (%)
1	MLP algorithm	15	Cross-validation 10-fold	88.99	99.7153
		15	Percentage split 66%	222.41	99.5433
		15	Use training set	81.13	99.8361

Table 5.6 Summary of Experiments Using MLP Algorithm

The experimentation result using 10-fold cross validation test mode for 15 attributes scored 99.7153% overall accuracy. But, the time taken to build the model using 15 attributes is 88.99 seconds. From the total of 11592 instances only 33 (0.2847%) of them are incorrectly classified. The detail accuracy by class evaluation result indicates that true positive rate, false positive rate,

precision, recall, F-measure and ROC area are 0.998, 0.007, 0.999, 0.998, 0.998 and 0.999 respectively for non-fraudulent class. On the other hand, the true positive, false positive, precision, recall, F-measure and ROC area results for fraudulent record instances are 0.993, 0.002, 0.99, 0.993, 0.992 and 0.999 respectively. The result of the confusion matrix shows that from the total of 11964 fraudulent records 1951 (99.3381%) of them correctly classified. But, the remaining 13 (0.6612%) records are incorrectly classified as non-fraudulent. On the other hand, from the total of 9628 non-fraudulent records 9608 (99.7923%) of them are correctly classified and the remaining 20 (0.2077%) are wrongly classified as fraudulent.

The second experiment for discussion is using percentage split 66 percent test mode using 15 attributes. The model resulted 99.5433% overall accuracy and the time taken to build the model is 222.41 seconds. From the total of 3941 test records 3923 of them are correctly classified and the remaining 18 (0.4567%) records are wrongly classified. The result from the detail accuracy by class shows that true positive rate, false positive rate, precision, recall, F-measure and ROC area for non-fraudulent class are 0.999, 0.02, 0.996, 0.999, 0.997 and 0.999 respectively. However, the true positive rate, false positive rate, precision, recall, F-measure and ROC area results for fraudulent class is 0.98, 0.001, 994, 0.98, 0.987 and 0.999 respectively. Additionally, the result of the confusion matrix shows that from the total of 3246 non-fraudulent instances 3242 (99.8768%) of them are correctly classified and the remaining 4 (0.1232%) records are wrongly classified as fraudulent. From the total of 695 records 681 (97.9856%) of the records are correctly classified but 14 (2.0144%) records are wrongly classified as non-fraudulent.

The third experiment result for discussion is using training set test mode with 16 attribute. This model resulted in 99.8361% overall accuracy and time taken to build the model is 81.13 seconds. The above overall accuracy shows that from the total of 11592 records 11573 of them are correctly classified and the

remaining 19 (0.1639%) records are wrongly classified. In addition, the detail accuracy by class result for non-fraudulent class shows that true positive rate, precision, recall, F-measure and ROC area is 0.999 except for false positive rate which is 0.004. Regarding, fraudulent class the result for true positive rate, false positive rate, precision, recall, F-measure and ROC area is 0.997, 0.001, 0.994, 0.996, 0.995 and 0.999 respectively. On the other hand, the result from the confusion matrix shows that from the total of 9628 non-fraudulent records 9616 (99.8754%) of them are correctly classified and the remaining 12 (0.1245%) records are wrongly classified as fraudulent. Additionally, from 1964 fraudulent records 1957 (99.6436%) of them are correctly classified and the remaining 7 (0.3564%) records are wrongly classified as non-fraudulent.

From the above discussion, the model from MLP algorithm using training set test mode is selected as best model from neural network. Its overall accuracy is 99.8361%.

Experimentation 8 Using MLP Algorithm with WEKA Selected Attributes

Using WEKA's attribute selection feature selected four attributes from the selected 16 attributes. Experiments using training set, 10-fold cross validation and percentage split 66% test options are conducted to see the effect of using WEKA selected attributes. In the following paragraph, discussion is made on the best performing model. Summary of the resulted models is presented in table 5.7.

Experiments	Algorithm /function	Number of attributes	Test modes	Time taken to build the model (sec)	Accuracy (%)

1	MLP (WEKA selected 4 attributes	4	Cross-validation 10-fold	13.16	98.8527
		4	Percentage split 66%	13.74	98.6805
		4	Use training set	13.76	98.9303

Table 5.7 MLP Experimentation Result Using WEKA Selected Attributes

In the experiment using WEKA selected four attributes, use training set test option resulted in 98.9303% overall accuracy. This is the top result using MLP algorithm and the time taken to build the model is 13.76 seconds. The model is able to classify 11468 instances correctly from the total of 11592 records and the remaining 124 (1.0697%) is wrongly classified. In addition, the detail accuracy by class result shows that true positive rate, false positive rate, precision, recall, F-measure and ROC area is 0.989, 0.009, 0.998, 0.989, 0.994 and 0.991 respectively for non-fraudulent class. Similarly, the fraudulent class true positive rate, false positive rate, precision, recall, F-measure and ROC area is also 0.991, 0.011, 0.948, 0.991, 0.969 and 0.991 respectively. On the other hand, the result from confusion matrix shows that from the total of 9628 non-fraudulent records 9521 (%) of them are correctly classified and the rest 107 (%) instances are wrongly classified as fraudulent. Likewise, from the total of 1964 fraudulent records 1947 (%) of them are correctly classified and the rest 17 (%) instances are wrongly classified as non-fraudulent.

The models resulted by changing the classifier test options are used to compare the models with each other. Here the parameters for the algorithm like number of hidden layers, learning rate, momentum, seed, training time, validation set size and validation threshold are altered to see if the algorithm gives a better model. Neural network models that resulted best are selected for discussion under this section but the detail experimentation results using each test options are attached in Appendix III.

Experiment 9 Trend Experiment Using MLP Algorithm

In the experiments conducted using different attributes ranging from 15 to 9 attributes by gradually reducing the number of determinant attributes. The

maximum difference using 15 to 10 attributes is 0.25% from percentage split 66% test option and the minimum is 0.07% from 10-fold cross validation test option. But, from 15 to 9 attributes the maximum difference in accuracy level is 11.69% and the minimum is 11.55%. Attribute reduction is made starting from SMS, GPRS, CallInter, OCS, Sub_Age and call per day respectively. This indicates that the number of calls per day can be taken as a major determinant attribute. The result from trend experiment shows that these attributes on voice CDR are relevant in identifying fraudulent calls and MLP algorithm is the most sensitive to these attributes. For further detail please refer Appendix VIII.

5.1.2 Discussion on Impact, Rules and Trend

Based on the rules generated using J48 and PART algorithm, each SIM card that is used in SIM box on average makes 70 calls per day. This will be 2,100 calls per month and the average duration per call is more than 4 minutes (240 seconds). Therefore, the minimum total number of minutes for only one SIM card is 8,400 minutes. This will be 1,596 USD per SIM card. In this study a total of 171 mobile numbers are found from the sampled 11,592 records. By only taking this minimal figure, ethio telecom loses a multiple of 1596 USD for the identified mobile numbers. It is more than 272 thousand USD per month. This figure confirms the governments concern about telecom fraud, in addition to threats on national security.

The rules or models derived from decision tree (J48 and PART) shows that it is possible at least to approximate the location of SIM boxes. For instance, Rule # 4 and 7 gives the range of BTS numbers (CELL_A) that tells the approximate location of the mobile number(s). In addition, the rules also indicate the MSC numbers with more fraudulent calls, as an example we can take Rule # 1 and 6. Moreover, the rules also indicate the range of mobile numbers that are fraudulent or used for SIM box. Rule 22 to 25 can be good example for this case.

Trend analysis using J48, PART and Multilayer perceptron algorithms with different number of attributes is made. The experiments are conducted using all attributes and by gradually reducing determinant attributes. From all the algorithms, it is observed that the accuracy level difference is less than 1 in all experiments from attribute 15 to 10. But, the difference between the maximum and minimum accuracy level for attribute 15 to 9 is more than 2% except for J48 using training set test option, which is 0.72%. Significance accuracy level difference is observed for multilayer perceptron algorithm for attributes 15 to 9, which is greater than 11.54%. In general, neural network algorithm is more sensitive and PART algorithm is the second sensitive algorithm for these determinant attributes.

5.1.3 Comparison of Classification Models

In this study J48 algorithm from tree based, PART algorithm from rule based and multilayer perceptron from function based are used. In WEKA data mining tool, the above algorithms are grouped under classifier but multilayer perceptron is from the group of artificial neural network, J48 is from decision tree and PART from rule base algorithms. The following paragraphs focus on comparison of the best models from each algorithm. The detail experimentation results for each algorithm are found in Appendix I, II and III. The top three models derived from each algorithm are discussed in the previous sections. It is now time to compare the models resulted from the three algorithms. Table 5.8 presents the summary of top scoring models obtained from J48, PART and MLP algorithms.

The model resulted from decision tree J48 algorithm by using full training set (use training set) test option has an overall accuracy of 99.9827%. The result of the confusion matrix indicates that prediction for non-fraudulent records is 99.9792% accurate. This is from the total of 9628 non-fraudulent records 9626 of them are correctly classified. The remaining 2 (0.0173%) records are classified wrongly. On the other hand from the total of 1964 fraudulent records

all of them are correctly classified. The time taken to build this model is 3.26 seconds with number of leaves 11 and size of tree 21. In addition, the detail accuracy by class result shows that true positive rate, precision, recall, F-measure and ROC area is 1 and false positive rate is 0, for non-fraudulent class. Similarly, true positive rate, false positive rate, precision, recall, F-measure and ROC area is 1, 0, 0.999, 1, 0.999 and 1 for fraudulent class respectively.

In the experiment using PART algorithm with 15 attributes and using percentage split 66% resulted in 100 percent overall accuracy. The model also delivered 7 rules. The detail accuracy by class shows that true positive rate, false positive rate, precision, recall, F-measure and ROC area are all 1 except false positive which is 0 for both fraudulent and non-fraudulent classes. Since, the accuracy is 100 percent both fraudulent and non-fraudulent instances are correctly classified. No record is incorrectly classified.

The model from artificial neural network is obtained with multilayer perceptron algorithm and using training set test option. The overall accuracy of this model is 99.8533% and time taken to build the model is 141.08 seconds. This shows that 11573 records are correctly classified from the total of 11592. The confusion matrix indicates that from the total of 9628 non-fraudulent records 9616 (99.8754%) of them are accurately classified and only 12 (0.1246%) instances are wrongly classified as fraudulent. Additionally, from the total of 1964 fraudulent records 1957 (99.6436%) of them are correctly classified but 7 (0.3564%) records are wrongly classified. On the other hand, the detail accuracy by class result for non-fraudulent class shows that true positive rate, false positive rate, precision, recall, F-measure and ROC area are 0.999 for all except false positive rate which is 0.003. Similarly, for fraudulent class the true positive rate, false positive rate, precision, recall, F-measure and ROC area are 0.997, 0.001, 0.994, 0.997, 0.996 and 0.999 respectively.

From the above discussion regarding top scoring classification models PART algorithm resulted in 100 percent accuracy with both 15 attributes. The algorithm used is percentage split 66% test option. This test option split 66 percent of the dataset for training and the remaining 34 percent for testing. After training using the 66 percent dataset, testing is conducted using 3941 (34%) instances of the data. Finally, the test result showed that 100 percent accuracy is achieved. Therefore, the model from PART algorithm is then selected as the first best model of the study and the resulted model is found in Appendix V. The second best model of the study is using J48 algorithm by applying use training set test option with 99.9827% overall accuracy. J48 algorithm resulted model and tree is attached in Appendix VI and VII. The third best model of the study is from artificial neural network using multilayer perceptron by applying use training set test option with 99.8533% overall accuracy.

Algorithm	Test Options & Parameters	Number of Records & Percentage	Predicted		Total
			Fraud	Non-fraud	
J48	use training set & default paramenters	Total records (T)	1964	9628	11592
		Correctly classified (C)	1964	9626	11590
		Incorrectly classified (I)	0	2	2
		Correctly classified (C/T) %	100	99.9792	99.9827
		Incorrectly classified (I/T) %	0	0.0208	0.0173
PART	Percentage split 66% & default paramenters	Total records (T)	695	3246	3941
		Correctly classified (C)	695	3246	3941
		Incorrectly classified (I)	0	0	0
		Correctly classified (C/T) %	100	100	100
		Incorrectly classified (I/T) %	0	0	0
MLP	use training set & default paramenters	Total records (T)	1964	9628	11592
		Correctly classified (C)	1957	9616	11573
		Incorrectly classified (I)	7	12	19
		Correctly classified (C/T) %	99.6436	99.8754	99.8361
		Incorrectly classified (I/T) %	0.3564	0.1246	0.1639

Table 5.8 Summary Top Scored Models from J48, PART and MLP Algorithms

5.2 Evaluation

At this point, evaluation of the model is made if it meets the objectives of the research. Among the business objectives of the company, offering best quality of service and building financially sound company are among the four objectives. To meet these objectives predicting fraudulent calls coming through SIM boxes will help to minimize revenue loss and maximize quality of service. Loss minimization or avoiding revenue loop hole is one way of profit maximization in any business environment. The finding of this study will help the company to give due emphasis for SIM box fraud.

By applying data mining on the dataset collected from ethio telecom different models are obtained. Among these models comparison have been made to select the best one that resulted with highest accuracy level. A comparison among the top best models from the three algorithms have been made in order to propose the one that fits for predicting fraudulent calls using SIM boxes for terminating international calls. It is now time to evaluate the resulted model on predicting fraudulent calls based on the dataset at hand.

The resulted classification models both for decision tree and neural network summarized discussion are presented as follows. From decision tree, J48 algorithm, using training set test option and 11592 records, 99.9827% accuracy is achieved. From the PART algorithm using percentage split 66 percent test option resulted in 100% accuracy level for the same number of records. The model from multilayer perceptron algorithm also showed overall accuracy of 99.8533% by applying use training set test option. Both decision tree and artificial neural network models resulted in best accuracy level but the one from PART algorithm is selected since its accuracy level is higher than the others.

The identified fraudulent mobile numbers using the data mining classification models are cross checked using Z-smart CCB application software. This is done

by the researcher's access privilege on the system. The result of checking on the system shows that more than 90 percent of the identified mobile numbers are found to be fraudulent. For the purpose of illustration and demonstration print screen image is taken from Z-smart software when the mobile (billing) numbers are checked in Z-smart system. This is how the researcher tried to evaluate the results obtained from the selected algorithms models. In addition, Z-smart CCB software is used in departments or divisions like security, IT operation, customer service contact center, sales and collection, activation and registration, retention and loyalty, and many others. The identified fraudulent mobile numbers are also blocked or terminated using this application software by the order of security division. Therefore, the researcher used this tool to validate the results from the derived models. As mentioned earlier, the researcher has been working in telecom sector, specifically Ethio telecom, since 2002. The different departments that the researcher came across with different responsibilities are sales, technical audit or IT/management audit, customer service contact center, network trouble ticket expert, and others. The application interface on Z-smart used for cross checking the result from the research is found in Appendix IX.

In addition to the above evaluation method, domain experts are invited to comment on the derived models and rules. But, up to this minute the manager and experts from fraud management section didn't show willingness to participate or evaluate the resulted models. The people in the domain area, specifically the concerned manager, is suspicious and reserved fearing not to disclose the policies or rules and checking mechanisms in the section. But, the researcher as domain expert in telecom sector has used the access privileges to evaluate the results from this study and assure that the developed models are worth applying.

5.3 Deployment of the Result

Deployment of the result is the final stage of CRISP-DM process model. The output of this research, models, rules and patterns can now be deployed by creating an interface with the existing system to detect fraudulent calls. The result of this research enables telecom companies, specifically ethio telecom, to use these data mining models in order to detect SIM box or gateway frauds. In addition to providing a clue about the problem in a scientific manner, it provides the knowledge, pattern and model to identify fraudulent calls. The SIM gateways that are used to terminate international calls can be identified easily using the rules derived from PART and J48 algorithms.

Moreover, the rules obtained from PART and J48 algorithm can be used as SIM box fraud detection policy for fraud management section of ethio telecom. It can also be used to update SIM box fraud detection policies, if there is any.

On top of its implementation on existing fraud management system (FMS), the finding of this research can also be an input for the projects in quality circle. Quality circle projects are projects where major problems in the company that need lasting solution are addressed. The projects are organized by incorporating different stake holders from different departments. This SIM box fraud problem is one of the project issues in quality circle.

Chapter Six

Conclusions and Recommendations

6.1 Conclusions

In this study an effort has been made to build a predictive modeling for fraud detection using classification method. Decision tree and neural network classification algorithms such as J48, PART and multilayer perceptron are used. This study mainly used J48 and PART from decision tree and multilayer perceptron from neural network. PART algorithm from decision tree resulted with best accuracy than multilayer perceptron and J48. The result from J48 and multilayer is still promising but to propose one model as the best for this study, the model from PART algorithm is selected. This algorithm using percentage split 66% test option resulted in 100 percent accuracy. It is mainly due to the use of additional attributes such as SMS, GPRS, calls per day, call ratio in addition to voice CDR data. In general, it is proved that decision tree has the best accuracy than that of neural network in fraud prediction.

The rules generated from decision tree tries to group mobile numbers that are used for SIM box purpose. In addition, it tries to locate the area by ranging the group of BTS numbers. When a mobile number uses few number of BTS numbers, it can be suspected by looking additional parameters like the number of calls made daily, incoming calls, GPRS, SMS, call ratio and international traffic. The rules from the models provide information by indicating the MSC numbers, range of mobile numbers and BTS numbers. Each regional state or major cities of the country has at least one MSC. The model tried to indicate how many frauds are found under a certain range of MSC numbers. One can understand using the model that these frauds are not limited in one area like Addis Ababa. These frauds are found in MSCs less than or equal to 251911299703 which is Addis Ababa and greater than 251911299715 which is

outside Addis Ababa. For security reason, it may be enough to mention some of MSC numbers (251911299701 to 251911299703) in Addis Ababa and leave the remaining for the company. The ethio telecom has detailed data on which BTS number is located where and under which MSC. Therefore, it is easy to trace the fraudsters and take legal and remedial action. This implies that the SIM box fraud is spread all over the country. It is also possible to say that individuals in regional cities or major cities of the country are involved in this fraud activity.

Based on the rules generated using J48 or PART, each SIM card that is used in SIM boxes at least makes 70 calls per day. This will be 2,100 calls per month and the average duration per call is more than 4 minutes (240 seconds). Therefore, the minimum total number of minutes for only one SIM card is 8,400 minutes. This will be 1,596 USD per SIM card. In this study a total of 171 mobile numbers are found from the sampled 11,592 records. By only taking this minimal figure, ethio telecom loses a multiple of 1596 USD for the identified mobile numbers. It is more than 272 thousand USD per month. This also confirms the governments concern on telecom fraud in addition to security threats.

The rules generated from the models can help ethio telecom or other telecom companies to revise or cross check the SIM box fraud detection policies they have. At least this can help as a starting point to further work on SIM box frauds currently spreading in Africa, in general (Adu-Boafo, 2013)

6.2 Recommendations

The GSM network is capable of reading and registering mobile IMEI numbers. Such kind of data is obtained from mobile switches. If ethio telecom enables the network devices to register IMEI numbers for pre-paid mobiles, it also helps to identify mobile numbers under a specific IMEI and take some kind of action on the devices. For fraudsters losing the SIM cards is not a big loss as such but

the devise. Once the devise is disabled from the network, it is no more useful in that country. If a range of SIM cards are blocked today, he/she can get many more since there is excess supply with cheap price. The researcher was intending initially to identify the IMEI numbers that are used to terminate international calls but unfortunately this number is not in the CDR. This value could be found in the CDR from switch due to security reason. The company can consider this attribute to facilitate the search for these fraudulent mobile numbers and the device.

The other recommendation as (Asfaw, 2006) and (Lokanathan & Samarajiva, 2012) indicated further reduction on TAR or termination rate could weaken fraudsters and minimize the revenue they earn by practicing such fraud. If we see the total accounting rate per minute trend in 2000, 2002, 2003, 2005, and 2012, it was 1, 0.8, 0.46, 0.40 and 0.19 respectively. If the rate reduced reasonably, they will be discouraged and go for other options. This action should be backed by strict follow up of such calls, appropriate law regarding such frauds and taking appropriate legal action.

This research is limited to prepaid mobiles only. One can conduct similar researches on postpaid mobiles. Since, similar frauds can be done using postpaid mobiles.

It is also possible to conduct similar researches on fraud by using signaling protocols or data used. But, there is a need to check if the data is accessible for research before submitting the title.

For this research WEKA is used for building predictive modeling for fraud detection in telecommunications. Similar researches can be made that has more options for clustering and neural network. Clustering algorithms like Genetic K-means Algorithm and Fast Genetic K-means Algorithm can be used to do similar research. These clustering algorithms are powerful but they are not found in WEKA. Further detail about these algorithms could be found in

(Elavarasi, Akilandeswari, & Sathiyabhama, 2011). In addition, the neural network in WEKA has only limited features. Related literatures suggests that neural network is best in identifying fraudulent calls in telecom sector but the result using WEKA is not the case. Therefore, the researcher recommends cross checking this result using other tools to validate it.

References

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

Acuna, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy *Classification, Clustering, and Data Mining Applications* (pp. 639-647): Springer.

Adu-Boafo, N. (2013). The big issue: Perspective on SIM Box Fraud in Ghana. *Africa Telecom & IT, 4*, 10-17.

Akhter, M. I., & Ahamad, M. G. (2012). Detecting Telecommunication Fraud using Neural Networks through Data Mining. [Journal]. *International journal of Science & Engineering Research, Volume 3*(Issue 3).

Ali, S., & Smith, K. A. (2006). On learning algorithm selection for classification. *Applied Soft Computing, 6*(2), 119-138.

Anderson, D., & McNeill, G. (1992). Artificial neural networks technology. *Kaman Sciences Corporation, 258*, 6.

Asfaw, N. (2006). Challenges Facing International Telecom Business and the Way Forward, Ethiopian Telecommunication Corporation's Perspectives. *Masters Thesis (Telecom MBA), College of Telecommunication and Information Technolgy, Management Department*.

Augustin, S., Gaißer, C., Knauer, J., Massoth, M., Piejko, K., Rihm, D., et al. (2012). *Telephony Fraud Detection in Next Generation Networks.* Paper presented at the AICT 2012, The Eighth Advanced International Conference on Telecommunications.

Bella, M. B., Olivier, M., & Eloff, J. (2005). *A fraud detection model for Next-Generation Networks.* Paper presented at the Proceedings of the 8th Southern African Telecommunications Networks and Applications Conference (SATNAC 2005), Central Drakensberg, KwaZulu-Natal, South Africa.

Berhanu, H. (2006). Fraud Detection in Telecommunication Networks Using Self-Organizing Map: The Case of Ethiopian Telecommunication Corporation (ETC). *Masters Thesis, College of Telecommunication and Information Technology, Department of Information Technology*.

Berry, M. J., & Linoff, G. S. (2004). Data Mining Techniques For Marketing, Sales, and Customer Relationship Management Second EdiM tion: Wiley Publishing, Inc.

Berry, M. W., & Browne, M. (2006). *Lecture notes in data mining*. Singapore: World Scientific.

Bounsaythip, C., & Rinta-Runsala, E. (2001). Overview of data mining for customer behavior modeling. *VTT Information Technology, 18*, 1-53.

Bresfelean, V. P. (2007). *Analysis and predictions on students' behavior using decision trees in Weka environment.* Paper presented at the Information Technology Interfaces, 2007. ITI 2007. 29th International Conference on.

Burge, P., Shawe-Taylor, J., Cooke, C., Moreau, Y., Preneel, B., & Stoermann, C. (1997). *Fraud detection and management in mobile telecommunications*

networks. Paper presented at the Security and Detection, 1997. ECOS 97., European Conference on.

Cerny, P. A., & Proximity, M. A. (2001). Data mining and neural networks from a commercial perspective. *Auckland, New Zealand Student of the Department of Mathematical Sciences, University of Technology, Sydney, Australia*.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., et al. (2000). CRISP-DM 1.0. *CRISP-DM Consortium*.

Choudhary, M. A., & Aftab, H. (2011). *Optimizing financial parameters to disincentivise international grey traffic and rationalization of measures to curb illegal international telephony in Pakistan.* Paper presented at the Technology Management Conference (ITMC), 2011 IEEE International.

CxB-Limited. (2013). CxB Solutions, SIM Box Detector. from http://cxbsolutions.com/html/sim_box_detection.html (Date accessed April 12, 2013)

De Ville, B. (2006). *Decision trees for business intelligence and data mining: using SAS enterprise miner*: Sas Inst.

Elavarasi, S. A., Akilandeswari, J., & Sathiyabhama, B. (2011). A survey on partition clustering algorithms. *International Journal of Enterprise Computing and Business Systems (IJECBS), 1*(1).

Estévez, P. A., Held, C. M., & Perez, C. A. (2006). Subscription fraud prevention in telecommunications using fuzzy rules and neural networks. *Expert Systems with Applications, 31*(2), 337-344.

Ethio-telecom. (2012a). Monthly Report of Internation Trafic Monitoring. *Unpublished, Ethio-telecom Internal Report*.

Ethio-telecom. (2012b). Performance related press confrence. from http://www.ethiotelecom.et/news/news.php?id=74 (Date accessed December 21, 2012)

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

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39*(11), 27-34.

Fekadu, M. (2004). Application of Data Mining Techinques to Customer Relationship Management (CRM): The case of Ethiopian Telecommunications Corporation. *AAU, Faculty of Informatics, Department of Information Science*.

Ferreira, P., Alves, R., Belo, O., & Cortesão, L. (2006). Establishing fraud detection patterns based on signatures *Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining* (pp. 526-538): Springer.

Frank, E., & Witten, I. H. (1998). Generating accurate rule sets without global optimization.

Gardner, M., & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)--a review of applications in the atmospheric sciences. *Atmospheric environment, 32*(14-15), 2627-2636.

Gebremeskal, G. (2006). Data Mining Application in Supporting Fraud Detection on Ethio-Mobile Services. *Masters Thesis, AAU, Faculty of Informatics, Department of Information Science*, 67.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. London The MIT press.

Harding, J., Shahbaz, M., & Kusiak, A. (2006). Data mining in manufacturing: a review. *Journal of Manufacturing Science and Engineering, 128*, 969.

Hilas, C. S., & Mastorocostas, P. A. (2008). An application of supervised and unsupervised learning approaches to telecommunications fraud detection. *Knowledge-Based Systems, 21*(7), 721-726.

Hornick, M. F., Marcadé, E., & Venkayala, S. (2007). *Java data mining: strategy, standard, and practice: a practical guide for architecture, design, and implementation*: Morgan Kaufmann.

Jember, G. (2005). Data Mining Application in Supporting Fraud Detection on Mobile Communication: The Case of Ethio-Mobile. *Masters Thesis, AAU, Informatics Faculty, Department of Information Science*, 98.

Jiawei, H., & Kamber, M. (2001). Data mining: concepts and techniques. *San Francisco, CA, itd: Morgan Kaufmann, 5*.

Jonsson, E., Lundin, E., & Kvarnström, H. (2000). *Combining fraud and intrusion detection-meeting new requirements.* Paper presented at the NORDIC WORKSHOP ON SECURE IT SYSTEMS-NORDSEC.

Kivi, A. (2009). Measuring mobile service usage: methods and measurement points. *International Journal of Mobile Communications, 7*(4), 415-435.

Kou, Y., Lu, C.-T., Sirwongwattana, S., & Huang, Y.-P. (2004). *Survey of fraud detection techniques.* Paper presented at the Networking, sensing and control, 2004 IEEE international conference on.

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

Lokanathan, S., & Samarajiva, R. (2012). Carrots and sticks: Principles of effective regulation to curb illegal bypass in international voice traffic.

Melaku, G. (2009). Application of Data Mining Techinques to Customer Relationship Management (CRM): The case of Ethiopian Telecommunications Corporation's (ETC) Code Division Multiple Access (CDMA) Telephone Service. *AAU, Faculty of Informatics, Department of Information Science*.

Negarit, G. (2012). *Telecom Fraud Offence Proclamation*

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.

SAS-Institute. (2003). *Data Mining Using SAS Enterprise Miner: A Case Study Approach*. Cary, NC, USA: SAS Institute Inc.

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

Sumathi, S., & Sivanandam, S. (2006). *Introduction to data mining and its applications* (Vol. 29): Springer.

Tariku, A. (2011). Mining Insurance Data For Fraud Detection: The Case of Africa Insurance Share Company. *AAU, Faculty of Informatics, Department of Information Science*.

Tesema, T. B., Abraham, A., & Grosan, C. (2005). Rule mining and classification of road traffic accidents using adaptive regression trees. *International Journal of Simulation, 6*(10), 80-94.

Tesfaye, H. (2002). Predictive Modeling Using Data Mining Techniques in Support of Insurance Risk Assessment. *AAU, School of Graduate Studies, School of Information Studies for Africa*.

Tibebe, B. (2005). Application of Data Mining Technology to Support Road Traffic Accident Severity Analysis at Addis Ababa Traffic Office. *AAU, Faculty of Informatics, Department of Information Science*.

Two-Crows. (1999). *Introduction to Data Mining and Knowledge Discovery* (3rd edition ed.): Two Crows Corporation.

Umayaparvathi, V., & Iyakutti, K. (2011). A Fraud Detection Approach in Telecommunication using Cluster GA. *International Journal of Computer Trends and Technology (IJCTT)*(May to June).

Weiss, G. M. (2005). Data mining in telecommunications *Data Mining and Knowledge Discovery Handbook* (pp. 1189-1201): Springer.

Willassen, S. (2003). Forensics and the GSM mobile telephone system. *International Journal of Digital Evidence, 2*(1), 1-17.

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining.* Paper presented at the Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.

Witten, I. H., & Frank, E. (2000). Nuts and bolts: Machine learning algorithms in java. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, 265-320.

Ye, N. (2003). *The handbook of data mining*: Lawrence Erlbaum Associates, Publishers.

Yigzaw, M., Hill, S., Banser, A., & Lessa, L. (2010). *Using Data Mining to Combat Infrastructure Inefficiencies: The Case of Predicting Non-payment for Ethiopian Telecom.* Paper presented at the 2010 AAAI Spring Symposium Series.

Appendix I: Summary of J48 algorithm experimentation

(using different attributes, test options and parameters).

Experiments	Algorithm /function	Number of attributes	Number of leaves	Size of trees	Test modes	Time taken to build the model (sec)	Accuracy (%)
1	J48 (with all attributes and no repeatition)	15	11	21	Cross-validation 10-fold	3.3	99.9224
		15	11	21	Percentage split 66%	3.26	99.9746
		15	11	21	Use training set	3.26	99.9827
2	J48 (WO SMS)	14	19	37	Percentage split 66	2.24	99.797
		14	19	37	Using training set	2.85	99.9051
		14	19	37	Cross validation 10 fold	2.57	99.7757
3	J48 (WO SMS, GPRS	13	20	39	Cross-validation 10-fold	2.95	99.7757
		13	20	39	Percentage split 66%	2.61	99.797
		13	20	39	Use training set	2.73	99.9137
4	J48 (with OCS, Sub_Age, CallRatio & CallPerDay)	12	13	25	Cross-validation 10-fold	2.83	99.8365
		12	13	25	Percentage split 66%	2.56	99.8731
		12	13	25	Use training set	2.33	99.9655
5	J48 (with OCS, Sub_Age & CallRatio)	11	122	238	Cross-validation 10-fold	2.74	98.1453
		11	122	238	Percentage split 66%	2.45	98.1984
		11	122	238	Use training set	2.37	99.4651
6	J48 (WO SMS,GPRS, Inter call, Repeated removed) or wz calls per day	10	27	52	Cross-validation 10-fold	1.92	99.5773
		10	27	52	Using training set	1.74	99.8188
		10	27	52	Percentage split 66%	1.82	99.6955
7	J48 (with OCS and Sub_Age)	10	138	272	Cross-validation 10-fold	2.99	97.8606
		10	138	272	Percentage split 66%	3.04	97.0566
		10	138	272	Use training set	2.77	99.1373
8	J48 (WO det att)	9	89	175	Percentage split 66%	2.57	98.5283
		9	89	175	Using training set	2.5	99.2581
		9	89	175	Cross validation 10-fold	2.48	98.568
9	Seed 2 numFold 3 minNumObj 2 confidence 0.25	9	89	175	Cross validation 10-fold	2.6	98.568
		9	89	175	Percentage split 66%	2.67	98.5283
		9	89	175	Using training set	2.73	99.2581
10	Seed 2 numFold 3 minNumObj 2 confidence 0.35	9	102	200	Cross validation 10 fold	2.43	98.5594
		9	102	200	Percentage split 66%	2.58	98.5029
		9	102	200	Using training set	2.23	99.3703

Experiments	Algorithm /function	Number of attributes	Number of leaves	Size of trees	Test modes	Time taken to build the model (sec)	Accuracy (%)
11	Seed 2 numFold 4 minNumObj 2 confidence 0.35	9	102	200	Cross validation 10 fold	2.99	98.5594
		9	102	200	Using training set	2.9	99.3703
		9	102	200	Percentage split 70%	2.32	98.7062
		9	102	200	Percentage split 80%	2.56	98.6195
		9	102	200	Cross validation 12 fold	3	98.706
		9	102	200	Cross validation 8-fold	2.95	98.568
		9	102	200	Percentage split 50%	2.54	98.1021
		9	102	200	Percentage split 60%	3.1	98.5767
12	Seed 2 numFold 4 minNumObj 3 confidence 0.35	9	94	185	Cross validation 10 fold	2.37	98.6025
		9	94	185	Percentage split 66%	2.91	98.4529
		9	94	185	Using training set	2.49	99.2926
13	J48 using WEKA selected attributes	4	15	19	Cross validation 10 fold	1.24	99.9482
		4	15	19	Percentage split 66%	0.79	99.8731
		4	15	19	Using training set	0.78	99.9655
NB: J48 algorithm experiment using different parameters and number of attributes							

Appendix II: Summary of PART algorithm experimentation results

(using different attributes, test options and parameters).

Experiments	Algorithm /function	Number of attributes	Number of Rules	Test modes	Time taken to build the model (sec)	Accuracy (%)
1	PART Using all attributes	15	7	Cross validation 10 fold	3.61	99.9396
		15	**7**	**Percentage split 66**	**2.95**	**100**
		15	7	Using training set	3.33	99.9827
2	PART (WO SMS) 2 repeated attributes	14	15	Cross validation 10 fold	3.52	99.7412
		14	15	Percentage split 66	3.07	99.594
		14	15	Using training set	3.28	99.8965
3	PART (WO SMS, GPRS) 2 repeated attributes	13	19	Cross-validation 10-fold	3.16	99.7585
		13	19	Percentage split 66%	3.34	99.594
		13	19	Use training set	2.52	99.8965
4	PART (WO SMS, GPRS, Inter call) 2 repeated attributes	12	16	Cross-validation 10-fold	2.44	99.6722
		12	16	Percentage split 66%	2.45	99.7746
		12	16	Use training set	2.42	99.8361
5	PART (WO SMS, GPRS, Inter call) (WZ ocs Sub_Age callRatio)	12	9	Cross-validation 10-fold	2.88	99.8792
		12	9	Percentage split 66%	2.63	99.9239
		12	9	Use training set	2.84	99.9741
6	PART (no determinant but 2 repeated attributes)	11	51	Cross-validation 10-fold	5.91	97.5414
		11	51	Percentage split 66%	5.74	97.9497
		11	51	Use training set	5.5	98.197
7	PART (WZ OCS SUB_AGE & CallRatio)	11	62	Cross-validation 10-fold	5.94	97.921
		11	62	Percentage split 66%	7.09	96.5237
		11	62	Use training set	6.62	98.7578
8	PART (WO FEE1)	10	51	Cross-validation 10-fold	5.09	97.8778
		10	51	Percentage split 66%	4.92	97.9447
		10	51	Use training set	5.29	98.197
9	PART (WZ OCS & SUB_AGE)	10	74	Cross-validation 10-fold	6.75	96.6701
		10	74	Percentage split 66%	7.32	96.1431
		10	74	Use training set	4.79	97.4206
10	PART (WZ OCS)	9	61	Cross-validation 10-fold	6.88	95.1259
		9	61	Percentage split 66%	7.37	94.6968
		9	61	Use training set	5.58	97.5155
11	PART (With no	9	44	Cross-validation 10-fold	5.8	97.8778

Experiments	Algorithm /function	Number of attributes	Number of Rules	Test modes	Time taken to build the model (sec)	Accuracy (%)
	repeated)	9	44	Percentage split 66	6.16	97.9193
		9	44	Using training set	6.09	97.6967
12	PART (seed 2 numFold 3 minNumObj 2 Confidence Factor 0.25)	9	44	Cross-validation 10-fold	6.63	97.8778
		9	44	Percentage split 66%	6.55	97.9193
		9	44	Using training set	6.09	97.6967
13	PART (seed 2 numFold 3 minNumObj 2 Confidence Factor 0.35)	9	44	Cross-validation 10-fold	5.97	97.8778
		9	44	Percentage split 66%	6.6	97.9193
		9	44	Using training set	5.84	97.6967
14	PART (seed 2 numFold 3 minNumObj 3 Confidence Factor 0.35)	9	54	Cross-validation 10-fold	5.29	97.8269
		9	54	Percentage split 66%	4.69	97.4372
		9	54	Using training set	5.3	98.7405
15	PART (seed 2 numFold 4 minNumObj 3 Confidebce Factor 0.35)	9	54	Cross-validation 10-fold	6.5	98.7405
		9	54	Percentage split 66%	6.89	97.4372
		9	54	Using training set	5.3	98.7405
		9	54	Cross validation 8-fold	2.83	97.5155
		9	54	Cross validation 12-fold	5.44	97.3602
		9	54	Percentage split 50%	6.95	97.3775
		9	54	Percentage split 60%	4.92	97.5199
		9	54	Percentage split 70%	6.95	97.326
		9	54	Percentage split 80%	6.91	97.239
16	PART using WEKA selected attributes	4	12	Cross-validation 10-fold	0.89	99.931
		4	12	Percentage split 66%	1.32	99.8224
		4	12	Using training set	0.87	99.9655

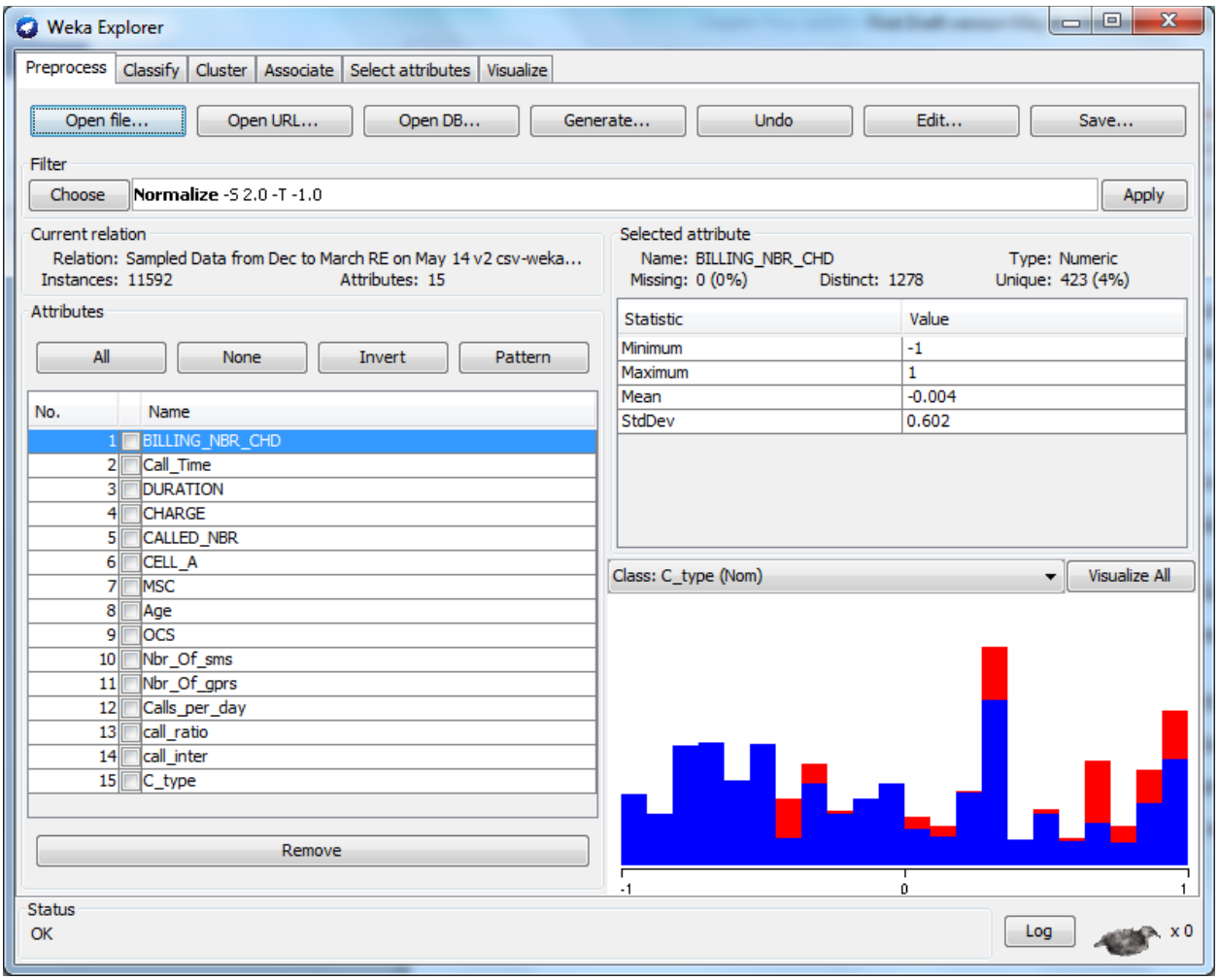
Appendix III: Summary of multilayer perceptron (MLP) algorithm experimentation results

(using different attributes, test options and parameters).

Experiments	Algorithm /function	Number of attributes	Test modes	Time taken to build the model (sec)	Accuracy (%)
1	MLP	15	Cross-validation 10-fold	88.99	99.7153
		15	Percentage split 66%	222.41	99.5433
		15	Use training set	81.13	99.8361
2	MLP (without SMS)	14	Cross-validation 10-fold	64.96	99.7412
		14	Percentage split 66%	62.74	99.6194
		14	Use training set	65.06	99.8275
3	MLP (without SMS & Callinter)	13	Cross-validation 10-fold	59.78	99.7153
		13	Percentage split 66%	59.37	99.6448
		13	Use training set	60.09	99.8533
4	MLP (without Call_ratio & Sub_Age)	13	Cross-validation 10-fold	63.06	99.4997
		13	Percentage split 66%	63.18	99.6701
		13	Use training set	63.12	99.7326
5	MLP (without SMS, GPRS & Callinter)	12	Cross-validation 10-fold	51.02	99.7067
		12	Percentage split 66%	51.03	99.7209
		12	Use training set	53.03	99.8188
6	MLP (without OCS, SMS, GPRS & Callinter)	11	Cross-validation 10-fold	48.02	99.7153
		11	Percentage split 66%	49.48	99.6701
		11	Use training set	47.93	99.7498
7	MLP (without SMS, Sub_Age, call_ratio & call_inter)	11	Cross-validation 10-fold	53.92	98.0331
		11	Percentage split 66%	54.92	97.0566
		11	Use training set	55.36	98.0245
8	MLP (without Sub_Age, OCS, SMS, GPRS & call_inter)	10	Cross-validation 10-fold	42.58	99.7153
		10	Percentage split 66%	41.87	99.797
		10	Use training set	44.62	99.8188
9	MLP (without SMS, Sub_Age, call_ratio, GPRS & call_inter)	10	Cross-validation 10-fold	44.48	98.0504
		10	Percentage split 66%	43.12	97.8187
		10	Use training set	43.39	98.2057

Experiments	Algorithm /function	Number of attributes	Test modes	Time taken to build the model (sec)	Accuracy (%)
10	MLP (without Sub_Age, OCS, SMS, GPRS, call_per_day & call_inter)	9	Cross-validation 10-fold	39.14	88.1039
		9	Percentage split 66%	38.83	88.0741
		9	Use training set	38.32	88.147
11	MLP (WEKA selected 4 attributes	4	Cross-validation 10-fold	13.16	98.8527
		4	Percentage split 66%	13.74	98.6805
		4	Use training set	13.76	98.9303

Appendix IV: Print shot of WEKA data mining tool for MLP algorithm using 15 attributes.



Appendix V: Output of PART algorithm resulted model

(using percentage split 66% with 15 attributes, which is the best model selected for this study).

```
=== Run information ===

Scheme:weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1
Relation:     Sampled Data from Dec to March RE on May 14 v2 csv-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R8
Instances:    11592
Attributes:   15
              BILLING_NBR_CHD
              Call_Time
              DURATION
              CHARGE
              CALLED_NBR
              CELL_A
              MSC
              Sub_Age
              OCS
              Nbr_Of_sms
              Nbr_Of_gprs
              Calls_per_day
              call_ratio
              call_inter
              C_type
Test mode:split 66.0% train, remainder test

=== Classifier model (full training set) ===

PART decision list
------------------

Calls_per_day <= 60 AND
Calls_per_day <= 59: NFR (8159.0)

call_ratio > 0.46 AND
call_inter <= 3 AND
call_ratio > 0.98: FRD (1594.0/1.0)

call_ratio <= 0.46: NFR (1378.0)

MSC <= 251911299721 AND
Calls_per_day > 67: FRD (353.0)
```

Calls_per_day > 66: NFR (85.0)

call_ratio > 0.97: FRD (19.0/1.0)

: NFR (4.0)

Number of Rules : 7

Time taken to build model: 2.95 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances	3941	100 %
Incorrectly Classified Instances	0	0 %
Kappa statistic	1	
Mean absolute error	0.0002	
Root mean squared error	0.0026	
Relative absolute error	0.0812 %	
Root relative squared error	0.6743 %	
Total Number of Instances	3941	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	1	NFR
	1	0	1	1	1	1	FRD
Weighted Avg.	1	0	1	1	1	1	

=== Confusion Matrix ===

a	b	<-- classified as
3246	0	a = NFR
0	695	b = FRD

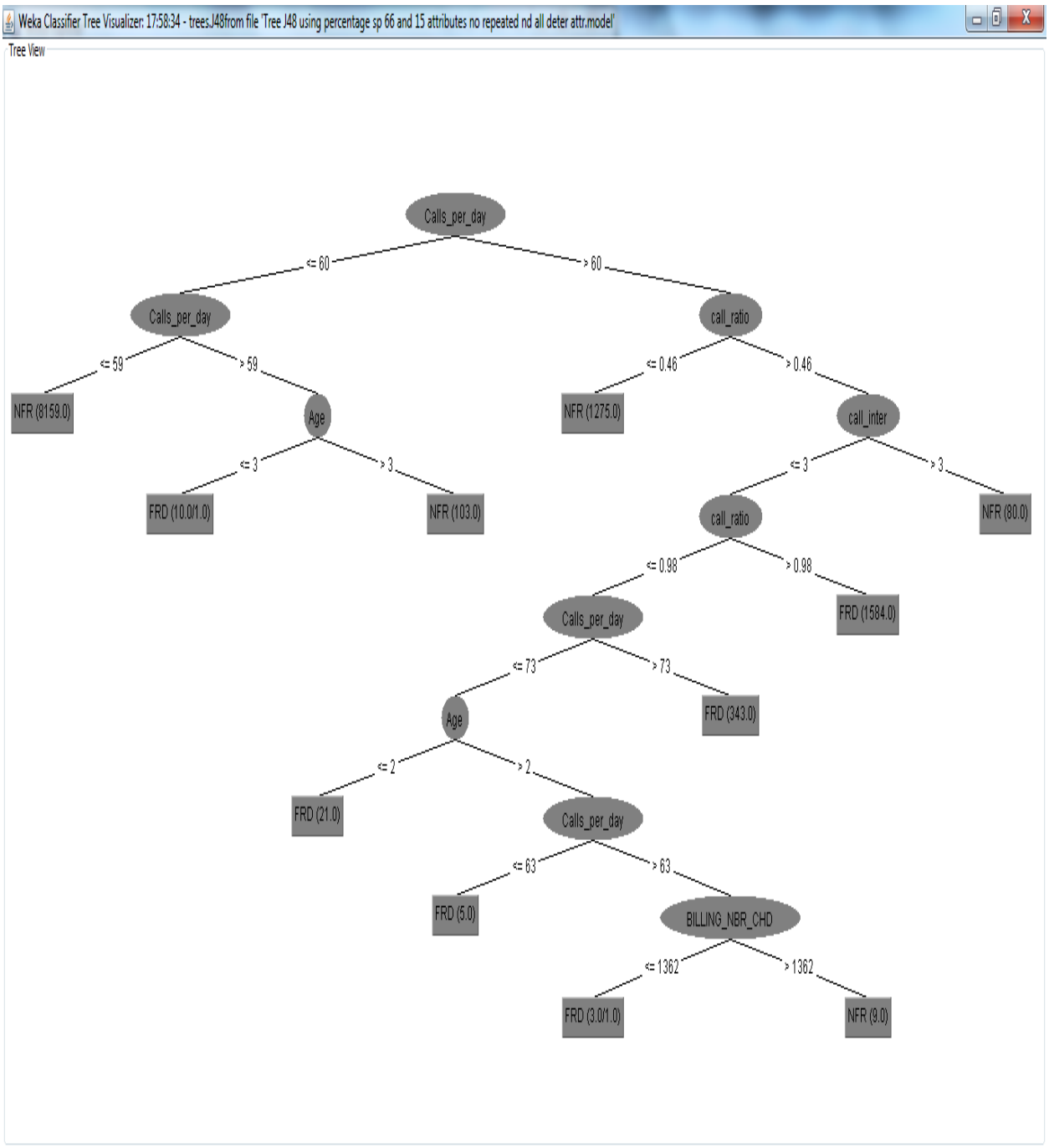
Appendix VI: Output of J48 algorithm resulted model

(using training set test option).

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Sampled Data from Dec to March RE on May 14 v2 csv-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R8
Instances:    11592
Attributes:   15
              BILLING_NBR_CHD
              Call_Time
              DURATION
              CHARGE
              CALLED_NBR
              CELL_A
              MSC
              Sub_Age
              OCS
              Nbr_Of_sms
              Nbr_Of_gprs
              Calls_per_day
              call_ratio
              call_inter
              C_type
Test mode:evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Calls_per_day <= 60
|   Calls_per_day <= 59: NFR (8159.0)
|   Calls_per_day > 59
|   |   Sub_Age <= 3: FRD (10.0/1.0)
|   |   Sub_Age > 3: NFR (103.0)
Calls_per_day > 60
|   call_ratio <= 0.46: NFR (1275.0)
|   call_ratio > 0.46
|   |   call_inter <= 3
|   |   |   call_ratio <= 0.98
|   |   |   |   Calls_per_day <= 73
|   |   |   |   |   Sub_Age <= 2: FRD (21.0)
|   |   |   |   |   Sub_Age > 2
|   |   |   |   |   |   Calls_per_day <= 63: FRD (5.0)
|   |   |   |   |   |   Calls_per_day > 63
|   |   |   |   |   |   |   BILLING_NBR_CHD <= 1362: FRD (3.0/1.0)
|   |   |   |   |   |   |   BILLING_NBR_CHD > 1362: NFR (9.0)
|   |   |   |   Calls_per_day > 73: FRD (343.0)
|   |   |   call_ratio > 0.98: FRD (1584.0)
|   |   call_inter > 3: NFR (80.0)

Number of Leaves  :  11
```

```
Size of the tree :   21


Time taken to build model: 3.26 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances       11590               99.9827 %
Incorrectly Classified Instances         2                0.0173 %
Kappa statistic                          0.9994
Mean absolute error                      0.0003
Root mean squared error                  0.0116
Relative absolute error                  0.096  %
Root relative squared error              3.0991 %
Total Number of Instances            11592

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure
ROC Area  Class
                 1         0          1         1         1
1        NFR
                 1         0          0.999     1         0.999
1        FRD
Weighted Avg.    1         0          1         1         1
1

=== Confusion Matrix ===

    a    b   <-- classified as
 9626    2 |    a = NFR
    0 1964 |    b = FRD
```

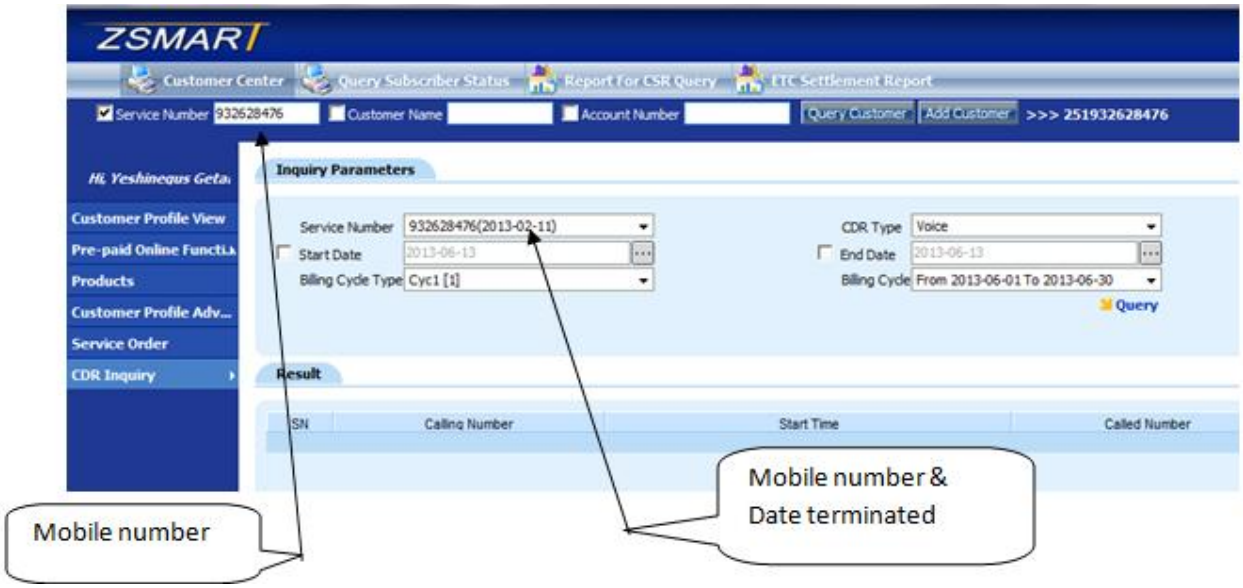
Appendix VII: View of J48 algorithm resulted model tree (second best model)



Appendix VIII: Summary of trend experimentation result

Algorithm	Test Options & Parameters	Number of Attributes	Number of rule /tree size	Accuracy	Number of Attributes	Number of rule /tree size	Accuracy	Number of Attributes	Number of rule /tree size	Accuracy	Number of Attributes	Number of rule /tree size	Accuracy	Number of Attributes	Number of rule /tree size	Accuracy	Number of Attributes	Number of rule /tree size	Accuracy	Number of Attributes	Number of rule /tree size	Accuracy
J48	10 fold C.V	15	21	99.92	14	21	99.92	13	21	99.92	12	25	99.90	11	25	99.88	10	25	99.87	9	243	97.92
	Perce. split 66%	15	21	99.97	14	21	99.97	13	21	99.97	12	25	99.87	11	25	99.87	10	25	99.87	9	243	97.46
	training set	15	21	99.98	14	21	99.98	13	21	99.98	12	25	99.97	11	25	99.97	10	25	99.97	9	243	99.27
PART	10 fold C.V	15	7	99.94	14	7	99.98	13	7	99.98	12	9	99.88	11	9	99.88	10	9	99.90	9	60	96.83
	Perce. split 66%	15	7	100.0	14	7	100.0	13	7	100.0	12	9	99.92	11	9	99.92	10	9	99.92	9	60	96.14
	training set	15	7	99.94	14	7	99.98	13	7	99.98	12	9	9.97	11	9	99.97	10	9	99.97	9	60	97.68
MLP	10 fold C.V	15		99.72	14		99.74	13		99.67	12		99.71	11		99.72	10		99.72	9		88.10
	Perce. split 66%	15		99.54	14		99.62	13		99.64	12		99.72	11		99.67	10		99.80	9		88.07
	training set	15		99.84	14		99.83	13		99.82	12		99.82	11		99.75	10		99.82	9		88.15

Appendix IX: Z-Smart Interface Used to Check the Identified Mobile Numbers



Appendix X: Attributes of CDR with Selection or Rejection Reason and Sample Data

No	Field name	Sample data	Reason for selection or rejection
1.	BILLING_NBR	931736703	Selected based on domain experts and literature recommendation
2.	START_TIME	1/1/2013 20:58	Selected based on domain experts and literature recommendation
3.	DURATION	747	Selected based on domain experts and literature recommendation
4.	CHARGE	546	Selected based on domain experts and literature recommendation
5.	CALLING_NBR	251931736703	Because it is similar with billing number
6.	CALLED_NBR	251913038519	Selected based on domain experts and literature recommendation
7.	BILLING_IMSI		Because it is blank or no data.
8.	MSRN		Because it is blank or no data.
9.	LAC_A	636014000845027	Not selected because it is similar with CELL_A
10.	CELL_A	636014000845027	Selected based on domain experts and literature recommendation
11.	LAC_B		Because it is blank or no data.
12.	CELL_B		Because it is blank or no data.
13.	TRUNK_IN		Because most of the value is blank or no data (when the call is international).
14.	TRUNK_OUT		Because most of the value is blank or no data (when the call is international).
15.	BILLING_CYCLE_ID	1498	Constant value (represent the month of January)
16.	FILE_ID	1269724093	File name from CDR data department
17.	RECORD_SEQ	3050	Because it is a sequence number and not recommended by literatures or domain experts.
18.	EVENT_INST_ID	3145509095	Because it is a unique ID given for a single call.
19.	RE_ID	2461	Constant value
20.	MSC	251911299702	Selected based on domain experts and helps to identify the region or city like Jima or Addis Ababa.
21.	FEE1	546	Because it is similar with CHARGE
22.	FEE2	0	Constant value /reserved for benefit packages
23.	FEE3	0	Constant value /reserved for benefit packages

No	Field name	Sample data	Reason for selection or rejection
24.	FEE4	0	Constant value /reserved for benefit packages
25.	CDR_TYPE		Because it is blank or no data.
26.	CALLING_AREA		Because it is blank or no data.
27.	CALLED_AREA	0	Constant value
28.	PLAN_NBR	3145509095	Because it is randomly generated.
29.	OFFER_NBR	11	Because it is randomly generated.
30.	THIRD_NBR		Because it is blank or no data.
31.	PART_ID	1	Constant value
32.	OCS_RE_ID	1008	Constant value
33.	BASIC_CHARGE	0	Constant value
34.	BENEFIT_CHARGE	0	Constant value