Addis Ababa University

College of Natural Science

School of Information Science

Predictive Modeling for International Roaming Fraud Detection in 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

Tarikua Worku

March, 2018

Addis Ababa, Ethiopia

Addis Ababa University

College of Natural Science

School of Information Science

Predictive Modeling for International Roaming Fraud Detection in ethio telecom

By

Tarikua Worku

Name and signature of Member of the Examining Board

Dereje Teferi (PhD)	_____
Advisor	Signature
Million Meshsha (PhD)	_____
Internal Examiner	Signature
Solomon Teferra (PhD)	_____
Internal Examiner	Signature

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any otherq degree or professional qualification.

Tarikua Worku _____
Name Signature

Place: Addis Ababa

Date of submission: March 2018

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

Dereje Teferi (PhD) _____
Advisor's Name Signature

Acknowledgement

Above all, my gratitude goes to the almighty GOD. To give me the strength and with me all times and will always be.

I am also thankful to Dr. Dereje Teferi (PhD) who helped me to reach this level. I am also indebted to all instructors who support me while my education period.

I would like to thank IT Fraud Operation, Revenue assurance, Core network, Signaling network, and Roaming and Interconnect stuffs of ethio telecom. For their support in sharing valuable experiences, and for their welcoming approach.

I also thank my beloved family, my Mother Wude Mengesha who encourage me while I am working and My Sisters. Last but not the least my son Athnatios I am grateful for your patience when I was busy and not able to give the time and attention you deserve.

Abstract

Telecommunication fraud is remaining a challenging task since the beginning of commercial telecom service. There are various reasons that makes telecom fraud detection and prevention challenging. Integration of new technologies with existing technologies without evaluating the security hole is main reason. International roaming service is one of the immerged service in mobile technology. Roaming service allow subscribers to continue to use their home operator phone number, and other services while they are in another country. This geographical difference between service providers and subscribers make the roaming service open for different types of fraud attacks. So prevention and detection of international roaming fraud is crucial for service providers.

In this study an effort has been made to build a predictive modeling for fraud detection using classification method. From decision tree and rule based classification algorithms: random forest, J48, and ZeroR are used. Random Forest meet the highest accuracy 99.8154% with around 0.0109% false positive rate. So Random Forest algorithm is proposed as the best algorithm to detect international roaming fraud than the other two algorithms (J48 and ZeroR).

Keywords: Telecom fraud, International Roaming, Data mining, random forest

Table of Contents

List of Tables

List of Figures

List of Appendixes

List of Acronyms

CAMEL:	Customized Applications for Mobile Network Enhanced Logic
CBS:	Convergent Billing System
CDR:	Call Detail Records
CFCA:	Communications Fraud Control Association
CLI:	Calling Line Identification
CRISP-DM:	Cross-Industry Standard Process -Data Mining
CRM:	Customer Relation Management
ESN:	Electronic Serial Number
FASG:	Fraud and Security Group
FDK:	Fraud Detection Tool
GPRS:	General Packet Radio Service
GSM:	Global Stations for Mobile communications
GSMA:	GSM Association
HPMN:	Home Public Mobile Network
HPMNO:	HPMN Operator
HUR:	High Usage Report
IMEI:	International Mobile Equipment identity
IMR:	International Mobile Roaming
IMSI:	International Mobile Subscriber Identity
IRSF:	International Revenue Share Fraud

ITU:	International Telecommunication Union
KDD:	Knowledge Discovery Database
LTE:	Long time Evolution
MMS:	Multimedia Service
MSISDN:	Mobile Station integrated Service Digital Network
NGN:	Next Generation Network
NRTRDE:	Near Real Time Roaming Data Exchange
PBX:	Privet Branch Exchange
PRM:	Partner Relationship Management
SIM:	Subscriber Identity module
SMS:	Short Message Service
SSM:	Special Sensor Microwave
VLR:	Visitor Location Register
VPMN:	Visited Public Mobile Network
WCDMA:	Wideband CDMA
WEKA:	Waikato Environment for Knowledge Analysis

This page is intentionally left blank

Chapter One

1. Introduction

1.1.Background

In less than thirty years, mobile communications have outshined the traditional fixed line telephony and become an integral part of everyday life. People are now more and more reliant on their mobile phone and expect to be connected anywhere at any time for work, social life, and contact emergency services [1].

Roaming is one of rapidly growing mobile communication services since1990s, and the first inter-operator roaming agreement was signed in 1992 (between Telecom Finland and Vodafone UK) [2]. Roaming is a feature of cellular networks allowing a customer of one operator to use the network of another operator, based on a wholesale inter-operator agreement. The most well-known form of roaming is international roaming, which allows users to use their mobile devices when abroad [2]. National roaming is roaming on networks of operators within the same country.

International Mobile Roaming (IMR) effectively extends the coverage of a roaming customer's home operator's retail voice and SMS services, allowing the customer to continue to use their home operator phone number, and data services while in another country. This seamless extension of coverage is enabled by a wholesale roaming agreement between a roaming customer's home operator and the visited network in the visited country, which addresses the technical and commercial components required to enable the service [3].

There are three major actors that intervene in roaming scenarios: the subscriber, who makes use of the telecommunications services provided; the Home Public

Mobile Network (HPMN), which handles the user's subscription and services; and the Visited Public Mobile Network (VPMN), in whose geographical coverage area users gain access to the services contracted with the HPMN. To enable the operator's subscribers to engage roaming facilities in a given VPMN, a roaming agreement must previously be negotiated between the two telecommunications operators. The procedures for drafting this business aspect are usually standardized, as in the case of Global System for Mobile Communications (GSM) service through the GSM Association (GSMA) [4] procedures.

The greatest cause of income losses in the telecommunication industry is fraud [5]. Telecommunication fraud can be defined as a deliberate attempt to misuse the products or services Processes of an operator without having to make any or at least a part of the payments. In the area of mobile Telecommunications, one of the business aspects that most contributes to income loss due to fraud is roaming [6].

Roaming fraud consists of access by the subscriber to the resources of the HPMN via the VPNM in such a way that the operator of the HPMN is unable to charge the subscriber for the services provided and is obliged to pay the operator of the VPNM for the facilities provided in the roaming scenario.

1.2.Ethio Telecom Scenario

Ethio telecom is government owned and the sole telecom operator in Ethiopia. The introduction of telecommunications in Ethiopia dated back to 1894 [7]. The service is growing from fixed telephone to fixed /wireless, Internet (dialup and broadband), mobile (pre-paid and post-paid), Code Division Multiple Access (CDMA) voice, internet and data, Wide (WCDMA) high speed internet and voice, 3G and recently 4G Long Time Evolution (LTE) and other value-added services (VAS) like SMS, MMS, are among the major telecom services provided by ethio telecom [8].

Ethio telecom introduced international mobile roaming services to its postpaid subscribers in 2003, and provides a roaming service in partnership with global partners and currently has a roaming agreement with over 461 operators all over the world [8]. The growth of Ethiopian economy leads the country to be more attractive for investment, tourism and to held international conferences. Because of the above situations the number of subscribers who request for International roaming services are growing. In October 2015 ethio telecom mobile subscribers registered and got service to foreign networks using their mobile while abroad are one hundred twenty five thousand and thirty-one (125,031) [8].

Roaming is the most expensive telecom service and as described above telecom service providers lose high revenue because of roaming fraud and ethio telecom is also one of the victim.

1.3. Statement of the Problem

Because of the evolution of telecom services that are supplied by telecommunication operators, international fraud organized networks have been developing complex fraud techniques that make it possible to generate substantial losses in a company's earnings. These losses may later have effects for the rates that these companies charge to their subscribers, which leads to a rise in prices [3].

Telecommunications fraud is considered as a bigger business than international drug trafficking, [3]. The motivation for perpetrators of such criminal activities are characterized by demographics, penetrative technology, culture, staff dissatisfaction, operational inefficiencies, lack of proper business models, greed, money laundering, geopolitical and socio-economic factors.

2013 Global Fraud Loss Survey of CFCA analyzed mobile communications industry estimated fraud loss is around USD 2.2 trillion. This estimation shows that the industry is losing USD 46.3 billion per Year from fraud, increasing at a rate

of 15% from 2011. This amounts to 2.09% of revenue and 6.11 billion dollars is lose annually because of roaming fraud [9]. It is therefore necessary, not only for operators but also for governments and users, to establish and facilitate technical, political, economic, and social measures that hamper roaming fraud [1].

Detection and prevention of frauds is one of the main objectives of the telecommunication industry. However, the volume of data being generated nowadays is increasing at phenomenal rate. So, extracting useful knowledge from such data collections is an important and challenging issue. In order to build such a non-trivial model, many researches were carried out on the feasibility of using the Data Mining (DM) techniques which comes from the need of analyzing high volumes of data collected by the telecommunication companies (customer data, unbilled calls, etc.) and related to different kinds of transactions between the company and its customers [10].

Ethio telecom is one of the victim telecom service provider and loss its revenue because of different type of telecom frauds, in 2015 "Ethio Telecom is reported to have lost around USD33 million to fraudsters" [4]. By considering all the above problems the research will analyze the questions below and give predictive analysis on international roaming fraud attacks.

- What possible fraud prevention and detection mechanisms can be setup in international roaming fraud?
- What are the methods used in ethio telecom to detect and prevent roaming fraud?
- Which Data Mining algorithm can be more suitable for the purpose of predicting International Roaming Fraud?

1.4. Objective of the study

1.4.1. General Objective

The general objective of the study is to analyze the international roaming traffic and come up with a predictive model for preventing and detecting international mobile roaming fraud.

1.4.2. Specific Objectives

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

- Understand the domain area through reviewing international roaming agreements, roaming policies and review different literatures on telecom fraud specifically in international roaming fraud area.
- Select appropriate data to predict fraudulent international roaming usage/calls.
- Build predictive models to evaluate and interpret the results.
- Evaluate and compare the performance of different DM algorithms and recommend the overall best results of DM models.

1.5. Scope and Limitation of the Study

1.5.1. Scope of the Study

This study is intend to introduce a better way to detect international roaming fraud (IRF) in the case of ethio telecom by using data mining techniques. The analysis is done using four months (June to September 2015) ethio telecom roam-in subscribers call detail records (CDR). Totally 500,000 records with 12 attributes are used to predict and model a better detection algorithm.

1.5.2. Limitation of the Study

The scope of this thesis is limited to understand International Roaming Fraud detection and prevention techniques, and standards which are used in ethio telecom to predict a better data mining model. The output of this work is constructed based on the data collected from ethio telecom (June-September 2015) roam-in CDR by applying on three different datamining algorithms (J48, ZeroR and Random Forest) on WEKA tool. The study is only limited on the above tool, algorithms and data.

1.6.Significance of the Research

Telecom fraud remain a serious global issue for communication network and services. As telecom technologies and services are changes, telecom fraud techniques also become more complex to identify them. So understand fraud causes and identify detection and prevention methods is significant.

This research enables ethio telecom to predict and on time detection of fraudulent international roaming calls. It fills the gaps of previously conducted researches in telecommunication fraud detection methods. It can be used as a reference for international roaming fraud scenario. It also can be used as a guide for future works.

1.7.Methodology of the Study

1.7.1. General Approach

Methodology is a way that deals with data collection, analysis and interpretation that shows how researcher achieves the objectives and answers the research questions. Hence, in order to achieve the general and specific objectives of the study CRISP DM model is in this research. It is because CRISP-DM (Cross-Industry

Standard Process for Data Mining) model is a fully documented, freely available, robust, and nonproprietary data mining model. The data mining tool selected for this research is Waikato Environment for Knowledge Analysis (WEKA). It is developed at the University of Waikato in New Zealand, written in java (object oriented programming language) and tested under different operating systems [11]

MS Excel, MS Access and MySQL are used for data analysis and experimentation. For data preparation, data understanding working with ethio telecom domain experts. Such as roaming service experts, Fraud Operation experts, IT operation experts, also domain related ethio telecom documents are reviewed.

1.7.2. Business Understanding

As per the impact and hugeness of telecom fraud. There are a lot of books, magazines, proprietary white papers, journal articles, conference papers are available. For this particular research international roaming fraud detection and prevention tools and techniques related literatures, roaming agreement and policy documents are reviewed including local research papers (ethio telecom).

In addition to this the researcher has telecom related knowledge to understand International Roaming related issues and the Call Detail Records (CDR) data attributes and it meanings.

1.7.3. 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 then Information system division. As per CIO direction the departments under IS division are cooperate for consulting and give the data as per the researcher requirements.

The data generated in the telecommunication industry is huge. A vast amount of data collect from different devices in every second [12].

1.7.4. Data Understanding and Preparation

Without business understanding data has no mining, in order to understand the data business context awareness, business strategy, business architecture, high level data architecture are important concepts. As it is explained the researcher has a domain knowledge on telecom environment so it was not difficult to understand ethio telecom domain experts' explanation of data and business rules. Data preprocessing phase helped to clean the data by avoiding empty columns, static values on the column, dealing with missing values, data reduction, data integration and transformation and other activities.

1.7.5. Modeling and Experimental Techniques

CRISP-DM process model is selected for data mining process of this research. CRISP is industry standard process which is applicable on telecom environment. Classification and neural network methods are also used for prediction. WEKA tool is for data analysis.

1.7.6. Evaluation Technique

Evaluation must be appropriate and crucial in order to assess the output of the study. The result of this research is evaluated in different ways. The firs technique is using Weka to test the data. The other evaluation is by comparing the model output with existing ethio telecom Fraud Management System (FMS) output. In addition to this discussion is conducted with ethio telecom IT fraud operation (ITFO), Revenue Assurance, and Roaming and Interconnect Agreement teams' based on the model output.

1.7.7. Deployment Technique

The Model is not require any special environment. It can be work in under any Operating system and can be integrated with the existing FMS.

1.8.Organization of the Thesis

This paper is organized as follow: the first chapter contain the introductory part which introduce the whole idea of the thesis. The subsequent chapter contain concepts of telecom fraud, telecom fraud classification methods, and related works are discussed in detail. In chapter three covered the discussion on data mining concept, tools and data mining 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 conclusion and recommendations part of this study.

Chapter Two

2. Review of Literature and Related Works

2.1. Telecommunication Fraud

Since the beginning of commercial telecommunications begins, the fraudsters have been causing financial damage to the companies who offered these services [1, 13]. In modern days, new technologies and services are emerged and have a great impact on people's life. This has resulted in the increase of frauds in today's technological environment [13]. Telecommunications fraud became fundamental issue for telecom operators. Telecommunications Worldwide Industry Experts surveyed and estimate annual global fraud losses to be in the range of $60-$70 billion (USD) [14]. These fraud losses represent approximately 4%-10% of telecom revenues. One of the largest contributing factors to losses due to fraud is the one coming from roaming scenarios [6].

Telecommunication fraud has been a major hindrance to the rapid growth of this industry as it has caused both the telecommunication operators and its subscriber's loss of revenue [15]. Fraud negatively impacts on the telephone company in four ways which are finance, market, customer relation and shareholders perception [16]. Fraud identification and detection is an increasingly important, expensive and difficult task in today's technological environment and the most difficult aspect of fighting fraud is identifying it.

The following topics of the chapter is classified in four main topics. The first is discuss about what telecommunication fraud is, and how fraud is categorized. The second section is discuss about international roaming and related frauds. The

third is about fraud effect and detection tools. The last topics of the chapter is about reviews data mining based fraud detection related literature.

2.1.1. Telecom Fraud Definition

Fraud is the act of deceiving others for personal gain. The word comes from the Latin fraudem, meaning deceit or injury, and over the years has come to represent a wide array of injustices, including forged artwork, confidence schemes, academic plagiarism, and email advance-fee [13, 17].

Telecommunication fraud can be defined as a deliberate attempt to misuse the products or services processes of an operator without having to make any or at least a part of the payments [4]. On a simple level, fraud can be described as any activity by which service is obtained without intention of paying. Organizations sometimes calculate how much money they lose through fraud by defining it as the money that is lost on accounts where no payment in received [13]. However, for detection purposes, such a definition is of little use, as using this definition fraud can only be detected once it has occurred. In fact, specifying what fraud is may be impossible, as the differences between fraudulent and non-fraudulent behavior may be indefinably small. However, what can be used for specifying fraud are examples of fraudulent behavior. When asking the question again, 'what is fraud?' it is these examples that provide the answer [13]

Telecommunication fraud is also defined as the unauthorized use, tampering or manipulation of a mobile phone or service [18]. Fraud can be referred to as the particular actions of employment of any services without being charged intention as in [19]. The main definition of telecommunication fraud correspond to the abusive usage of an operator infrastructure. This means using the resources of a carrier (Telecommunications Company) without intention of paying [20]. Currently no standard definition for 'telecommunication fraud' [4]. But it essentially involves using deception to make a personal gain dishonestly for oneself and /or

create a loss for another. Ethiopian criminal law, Telecom fraud offence proclamation No. 761/2012 define fraud as "fraud is the crime" [21].

Telecommunication fraud can be simply described as any activity by which telecommunications service is obtained without intention of paying [6]. Telecommunication fraud has certain characteristics that make it particularly attractive to fraudsters and it is a worldwide problem with substantial annual revenue losses of telecom companies [3].

It is known that Telecommunication industry has expanded dramatically in the last few years with the development of affordable mobile phone technology [18], with the increasing number of mobile phone subscribers, global mobile phone fraud is also set to rise. And became major financial concern for operators despite increased efforts to eradicate it [22]. The theft of service and misuse of voice as well as data networks of telecom providers is considered as fraud.

Telecom fraud is fascinating case study. People have been cheating phone companies for decades, and recently the phone companies have been dynamically returning the compliment. At the beginning of the twenty first century, the convergence of computing and communication technologies has altered considerably the way in which industrialized communities function [13]. Cybercriminals are increasingly adopting hyper-scale techniques to help them perpetrate fraud faster and more efficiently than ever before. Criminals are highly adaptive and continually evolving, able to work around old technology and old approaches [10]. At the late twentieth century, fraud matured in the area of transactional businesses, most notably in the telecommunications and credit card industries [10, 1].

When the first analogue mobile communication networks were launched, weaknesses in the security, particularly the lack of encryption of both the voice channel and the authentication data made the networks susceptible to eavesdropping

and cloning [1]. As the technology evolved from analogue to digital, so the nature of fraud changed as it became more difficult to eavesdrop and clone, and this led to a shift away from technical fraud towards more procedural and contractual types of fraud [13].

The development of new technologies has also provided criminals more sophisticated way to commit fraud and has required more advanced techniques to detect and prevent such events. Detection and prevention of frauds is one of the main objectives of telecommunication industry [23]. However the possibility of technical fraud cannot be ruled out forever in GSM as one door is closed on a fraudster, so the fraudster will attempt to open another [13]. The fraudster will always seek a way to beat the system and any fraud detection mechanism has to be cost effective [23]. In recent years, fraud is increasing rapidly with the development of modern technology and global communication [5]. Also the danger of localization is small, because all actions are performed from a distance which in conjunction with the mass topology and the size of network makes the process of localization time consuming and expensive. Additionally no particularly sophisticated equipment is needed. Simple knowledge of an access code, which can be acquired even with methods of social engineering, makes the implementation of fraud feasible.

One of the problem with telecommunication fraud is the huge loss of revenue and it can affect the credibility and performance of telecommunication companies [6]. The most difficult problem that faces the industry is the fact that fraud is dynamic. This means whenever fraudster's feel that they will be detected them find other ways to circumvent security measures. Telecommunication fraud also involves the theft of services and deliberate abuse of voice and data networks [10].

CFCA experts periodically estimates the level of worldwide telecommunications fraud since end of 20th century. In 1999 international fraud loss survey estimated around $12 billion, in 2003 between $35 and $40 billion, in 2006 between $55 and $60 billion, in 2009 between $70 and $78 billion [9, 24], and in 2015 estimated around $38.1 billion (USD) [6], down 18% from 2013 [14]. Annual global telecom revenues losses for fraud are approximately 1.69% in 2013 and it is decreased by 0.40% in 2015 [6], but 89% of operators surveys showed fraud losses had increased or stayed the same within their own companies [6]. CFCA 2015 survey shows the top five methods for committing fraud are PBX Hacking, IP PBX Hacking, and Subscription Fraud. Dealer Fraud, also the top five types of frauds are International Revenue Share Fraud (IRSF), Interconnect Bypass, Premium Rate Service (PRS), Arbitrage, and theft [6].

Many industry forums and fraud management companies are working with telecom operators, device manufactures and regulators. Fraud classification methods of different organizations is not similar. Some of the classification methods are in terms of penetration of the service, security weakness of the network and protocols, financial gain, and the technology used. Fraud classification will be described in the next sections.

2.1.2. Classification of Telecom Fraud

Telecommunication fraud is classified differently, earlier classifications complies in two categories which are subscription fraud and superimposed fraud. Subscription fraud denotes the behavior of using false identity to subscribe a service and evade payment. Superimposed fraud is the use of a service without having the necessity. Authority and is usually detected by the appearance of "phantom" call on a bill. Superimposed fraud includes mobile phone cloning, ghosting, insider fraud and tumbling [25, 16].

Recently telecommunication frauds are classified in wide-ranging categories. Based on behavior, data source, technology, service type, users or usage and many more. Fraud in telecommunications have been classified by the technical manner in which they are committed and non-technical purpose just for financial gains. A further classification can be done by considering whether the network abuse is the result of administrative fraud, procurement fraud, or application fraud [13]. In some documents telecom fraud classify simply in four groups contractual fraud, hacking fraud, technical fraud and procedural fraud [13, 4].

There are a lot of trade associations and fraud management companies are working on telecommunication fraud, by collaborating with telecom equipment manufacturers, regulators and telecom service providers. In this paper some of international associations which are GSM Association (GSMA), Communication Fraud Control Association (CFCA), TM Forum, SYNVERS, and ARGYLE fraud classification methods are discussed.

GSM Association represent the interests of mobile operators since 1987 worldwide. Spanning more than 220 countries, the GSMA units nearly 800 of the world's mobile operators with more than 230 companies [26]. GSMA FASG (Fraud and Security Group) was established in December 2014 and the mission of the group is to drive the industry's management of fraud and security matters related to GSM technology, networks and services [27]. GSMA first version fraud manual produced in 1996. In order to list and categories the various types of frauds. The 2014 version of GSMA fraud manual classify telecom frauds under the broad categories of; Technical fraud, Subscription fraud, distribution fraud, business fraud and prepaid fraud [27].

Under each category a lot of related fraud types are listed. For instance under technical fraud fourteen fraud types are categorized and some of them are SIM

Card Cloning, Spamming (SMS & IP services), and PBX Hacking. Under Subscription fraud there are nine fraud types are disclosed and some of them are Proxy Fraud and Call selling, Under Distribution fraud Five different fraud types are listed and some of them are Dealer Fraud, False Agent/ Remote Activation. Under Business fraud fifteen different fraud types are classified and some of them are PRS, Roaming, Internal Fraud, and Wangiri. Under the last category Prepaid Fraud Five fraud types are grouped and some of them are Scratch Card Abuse, Manual Recharging and Prepaid Services Abuse Fraud [27].

Communication Fraud Control Association (CFCA) is a not-for-profit global educational association that working to combat communication fraud. The mission of the CFCA is to be the premier international association for revenue assurance, loss prevention and fraud control through education and information [6]. CFCA classify telecommunication fraud in two main categories which are Fraud Method and Fraud Types.

Fraud method is how they access the network or service to enable revenue gain from the attack and fraud type is how they use the service or network to generate revenue from the attack [6]. On CFCA 2015 survey, the top five fraud methods are PBX Hacking, IP PBX Hacking, Subscription Fraud (Application), Dealer Fraud and Subscription fraud (Identity) also top five types of frauds are International Revenue Share Fraud (IRSF), Interconnect Bypass, Premium Rate Service, Arbitrage, Theft/ stolen Goods [6, 24].

Tele Management (TM) forum Telecommunication fraud classification guide is developed as a periodically growing resource to arm operators with fraud type information and offer them a best practice for a common fraud cases classification model. And the guide is structured to support Fraud Operations activities associated with the TM Forum Business Frameworks model [28]. Classified tel-

ecommunication frauds as Enabler Technique and Fraud Type [29]. Fraud enabler is the method or technique of getting access to the goods or service and perpetrating the Fraud, A Fraud Enabler can be an illegal action by itself. It is possible to have one or a combination of a set of Fraud Enablers for a specific fraud type [29]. Fraud Enables can be classified in four sub groups Attack Type, Fraudster Type, Location and Environment. Under each sub group there are various types of fraud like Subscription Fraud, PBX hacking, Arbitrage, SMS-C abuse and other. Fraud Types also classified in six sub groups Location, Environment, Objective, technology, Service and Supplementary Service and under each sub group there are other categories [29].

SYNVERS established in 1987, and become global leader in mobile inter-operability, mobile communications and mobile expertise. [30] SYNVERS classified fraud in five broad range of threats there are Domestic, Roaming, Prepaid, Data and Subscription and under each broad category there are vast number of fraud types are categorized [31]. Under Domestic threats there are around eighteen fraud types are listed and some of them are LTE fraud, SIM cloning, SIMBox fraud, IRSF/PRS. Under Roaming threats around twelve Fraud types are categorized and some of them are SIM Cloning, SMS fraud, Call selling, GPRS High usage and Subscription. Under Prepaid threat around eight fraud types are sub categorized some of them are Manual recharge fraud, high balance fraud, scratch card abuse and internal fraud. Also under Data threat five types Data usage frauds are listed. In the last category Subscription Threat around seven fraud types are listed and some of them are Identity theft fraud, dealer fraud, and Handset subsidy loss.

ARGYLE data is working on real-time fraud, security and revenue threat analytics applications that help mobile providers protect themselves from fraud threats, profit threats, SLA threats, and forensic threats [32]. The ARGYLE data Fraud classification guide book classify fraud differently [4].

There are many more respective organizations working in telecommunication fraud like KPMG, SUbex, i3 Forum and many more. But it is impossible to write about all in this paper.

Some of Fraud classification techniques are quite sophisticated and combine more than one known method [13]. Telecommunication fraud is not static, new techniques evolve as the telecom companies put up defenses existing ones. The fraudsters are smart opponents, continually looking for exploitable weaknesses in the telecom infrastructure.

In this limited resource it is difficult to describe about all fraud types. Instead this paper focused on International roaming related fraud types the description also focused in the fraud type in roaming scenario.

2.1.3. International Roaming Fraud

Roaming fraud is one of the highest revenue earners in the telecommunications industry, which means that it is also the most vulnerable to fraudulent attacks, Telecommunications analysts estimate that international mobile roaming rates generate approximately 5-10 per cent of operator's revenues globally, and constitute an even bigger slice of their profits [33].

Roaming fraud can starts as an internal or subscription fraud in the home network, when obtained SIM cards are sent to a foreign network the fraud or abuse starts. Roaming fraud is a gate for other fraudulent activities such as IRSF, Call Selling, and abusive PRS [29]. Roaming fraud can be very well-organized, involving groups of fraudsters working across international boundaries and at different points in the traffic chain. Suppose you made a call to America and it was directed to a specific location in a Pacific Island, but the number on that Pacific Island doesn't exist as it was sold off to a totally different location and turned into a premium rate service. Now, ask yourself: How am I going to investigate

that? What police force can you call to check on that, especially if it's viewed as an isolated problem? [34].

Roaming fraud looks a lot like other kinds of mobile fraud it's exploited by illegally obtained SIM cards. It's a launch pad for premium rate and international revenue share fraud and it is vulnerable to identity theft and masking [35]. There are other more organized methods to committing roaming fraud, such as subscription fraud committed by subscribers, or PBS and Wangiri fraud committed by third parties, or even cramming and slamming committed by the phone companies themselves [4]. Roaming is one of the main target of criminal organizations involved in subscription fraud. A number of roaming service subscriptions are made and the SIM cards are sent abroad within a few hours. Once registered on a foreign network, the SIM cards are used to resell traffic which will never be paid to the home operator, even if the HPMNO pays the invoices received from the VPMNO, including those related to fraudulent calls.

Roaming fraud can happen when a subscriber that used the services of the visiting network refuses to pay for them either by claiming ignorance, insufficient knowledge of additional costs, or by appealing that the service was never requested [4]. There are other more organized methods to committing roaming fraud such as subscription fraud, PBX Fraud, Wangiri fraud, IRSF fraud, PRS Fraud and many more [4, 20, 36]some of them are discuss under international roaming related fraud topic.

2.1.4. International Roaming Related Fraud Types

Various types of frauds are related to international roaming fraud the main reason is home operator do not have direct site of the roaming SIM. Roaming is more properly a gateway or entry point to the full range of mobile fraud threats, and the biggest priority, of course, is to ensure that this gateway is properly managed and cleansed and done so in a timely fashion and on a worldwide basis [35].

SIM Cloning

Cloning occurs when a customer's Mobile Identification Number (MIN) and Electronic Serial Number (ESN) are programmed into a cellular telephone not belonging to the customer. With the stolen MIN and ESN a cloned phone user can make virtually unlimited calls, whose charges are billed to the customer [26, 29]. Cloned phones are usually used to make international calls and in roaming, possibly abroad. False or stolen identities are used to acquire subscriptions that can never be properly billed to the defrauder. Technical developments like roaming are making the task of fraud avoidance even more difficult. In many instance, a fraud is committed with an intention to use the phone while roaming in a different network [36].

International Revenue Share Fraud (IRSF)

International revenue share fraud is a new type of fraud that started out on the internet and migrated to telecommunications. In the early days, fraudsters would commit international revenue share fraud by obtaining SIM cards and using them to call international revenue share numbers either in roaming or international areas [4]. Over time, technology has made it easier for fraudsters to artificially inflate traffic to these international revenue share numbers, and the attacks have become much more organized. Call forwarding and conference calling, for example, can extremely increase the cost of each call by having multiple international revenue share fraud as one of the biggest threats in telecommunication fraud.

Premium Rate Service (PRS) Fraud

Premium rate service (also known as audio text services) provide callers or SMS senders with information or content, which is charged at a premium rate. Typically this will related to adult chat lines, gambling, horoscopes, news, weather

etc. that may be provided via voice or SMS. Fraud is committed when the operator of the service stimulates or inflates calls into that service to attract higher settlement payments [4]. The most common occurrences of premium rate service fraud directly attack phone companies through the subscription fraud method [19].

SMS Fraud or Mobile malware:

People use smart phones for web browsing, social networking, online banking, and more. Smartphones also provide features that are unique to mobile phones, like SMS messaging, constantly-updated location data, and ubiquitous access. As a result of their popularity and functionality, smartphones are a burgeoning target for malicious activities. [37] Mobile phone spam is similar to email spam. Like with junk emails, mobile phone users receive unwanted texts and calls about special rewards or deals in the form of a simple message, a link to a number to call or text, or a link to a website. Call and SMS spamming victims may be charged a fee for every text message received, and the message could also be linked to a premium rate service line, which is not included in unlimited texting plans.

Call Selling

SIM cards are obtained to re-sell services, mainly to international destinations and / premium rate services. Calls are to users for a fraction of the true cost. Call selling is usually seen together with other types of fraud, such as subscription fraud or roaming fraud [29].

Call Conference Fraud

Call conference services enable a number of calls to be simultaneously set up and received by a mobile customer, international conference calling is sometimes offered as a service feature but it can be fraudulently abused of the phone being used to generate calls is fraudulently obtained or stolen. This feature is a common

method used by IRSF to artificially inflate revenue into their premium rate or other high cost services. [29, 27].

Call Divert/ Forwarding Fraud

Call forwarding services enable incoming calls to be automatically redirected to any national or international number. Four different call forwarding services are available (unconditional divert, when busy divert, when not reachable divert, and when not answered divert). While roaming, the conditional call forwarding features are set in the VLR of the VPMN. Calls towards the HPMN's subscriber are transferred from the HPMN to the VPMN, as no unconditional call forwarding is set. [29, 4]. Call forwarding features are setup on a stolen or fraudulently obtained SIM card. The destinations for forwarding the calls to are typically PRS or high termination fee destinations or other well-known IRSF destinations. But costs are charged to the subscriber of the SIM. Subscriber is not in HPMN Call is therefore forwarded to VPMN (RCF: roaming call forwarding) Subscriber is not answering the call, conditional call forwarding is set to target destination. More than 100 parallel calls can be made by the fraudsters using a single SIM card.

Subscription Fraud

Fraudsters obtain an account without intention to pay the bill. In such cases, abnormal usage occurs throughout the active period of the account. The account is usually used for call selling or intensive self-usage.

- Giving valid details, but disappearing without paying the bill
- Using false details
- Using the identity of another person(identity theft)

Fraudsters are still able to get connected to services, including purchasing mobile phones, by providing false details when subscribing. This can be particularly ex-

pensive for operators when a GSM mobile phone is purchased with roaming capabilities. This means that the phone can be used to make calls from foreign countries, and these calls obviously incur much higher charges. [29, 4].

Arbitrage Fraud

Telecom arbitrage fraud is the exploitation of the difference in settlement rates between countries. Phone carriers often charge different interconnection rates according to the type of call or service provider involved. International calls cannot be processed and completed through one phone carrier, so the originating carrier routes traffic via an intermediary phone carrier for an additional fee, called settlement rates. Settlement rates are what phone companies pay to each other for completing their calls and, until recent regulation, they were much higher than the actual cost of completing the calls. Different companies have different settlement rates according to each county. [29, 20, 27].

Internal Fraud:

These frauds are committed by personnel belonging to the telecommunications companies themselves, enabled by defective internal security systems of protocols. Two variants consist of the theft of SIM cards and their subsequent activation, and the use of test cards for roaming scenarios [29, 31]. Involves in technical configuration fraud, altered the routing configuration to favor a specific interconnect carrier's routes, outside of agreed and expected parameters.

The above mentioned fraud types are not the only frauds related to international roaming service. In addition to the above fraud types there are a lot more roaming service related fraud types including Roaming in LTE, IMEI reprogramming, Dealer Fraud, Handset Theft, Mobile Malware, Wingiri and many more.

2.1.5. The Effect of Fraud in Telecommunications

Communications Fraud Control Association (CFCA) one of globally known association which is working to fight communications fraud, annually publish results of worldwide telecom fraud loss survey. Based on the survey global estimation of annual losses are in billions of dollars. By 1999 it was estimated $ 12 billion, and in 2003 annual estimation was in the range of $35-$40 billion, in 2006 global fraud loss estimation was $54.4-$60 it is up by 52% from 2003 survey. In 2009 it became $72-$80 billion and by 2011 declined to $40.1 billion dollars and it was 33% down. From last estimation. In 2013 the estimated survey was 46.3 billion and it is up 15% from 2011, [9, 24] and the main reason for the relative increase in fraud is due to more fraudulent activity targeting the wireless industry and around $6.11 billion lose is because of roaming fraud [14] , 2015 communication fraud loss estimation survey announced $38.1 billion USD [6]. In addition to revenue loss there are losses service inaccessibility [missing genuine calls], Call Hijacking and lack of Lawful interception because of call redirection over unauthorized channels, and it leads to a failure in terms of national regulatory compliance [38] , in some fraud types due to lack of the original CLI, terminated /bypassed calls of the real called and calling party are not identifiable. Image loss due to bad quality of service and additional investment to battle fraud are other impacts

2.1.6. Telecom Fraud Detection Tools

It is probably true that it is impossible to totally eliminate fraud and the most difficult aspect of fighting fraud is identifying it. Fraud detection is an increasingly important and difficult task in today's technological environment, and required further investment to combat [22]. In 1999 there were thirty fraud management products on the market place [18]. Fraud Detection Tool (FDK) was one of the first fraud detection tool. Currently a vast number of antifraud systems and

tools are developed for intelligent fraud detection, and action. Fraud Management System (FMS) is considered as one of smart analytical fraud detection tool. FMS is using different data mining techniques and had built-in data mining engine. For international roaming fraud prevention Near Real-Time Roaming Data Exchange (NRTRDE) is mandatory between operators (visited network and Home network), NRTRDE Procedure is the sole GSMA approved roaming fraud procedure since 1 October 2008 [27].

Fraud detection has been implemented by a number of methods such as data mining, statistics and artificial intelligence [39] and there are a lot other method and technique but in this limited resource only data mining approach is used.

2.2.Data Mining

Data mining (DM) is becoming a mainstream technology used in business intelligence applications supporting enterprises such as financial services, healthcare, telecommunications and higher educations. The core phase of data mining is to mine or discover the novel information in terms of patterns or rules from the large volume of data. Now a days, it is becoming a common practice among business analysts, scientists and researchers to apply data mining. DM tasks can be classified into two categories that is Predictive and Descriptive mining. Dataset may be obtained from a variety of sources, including traditional relational databases, data warehouses, web documents or simple local textual files [11].

2.2.1. Data Mining and the KDD Process

DM also denoted to as knowledge Discovery in Databases (KDD), but both DM and KDD are diverse with each other. The term KDD is widely used to denote the overall process of extracting high-level knowledge from low-level data. Others also use the term data mining and KDD interchangeably. Data mining is the

process of finding patterns and relationships in the data [11]. Data mining consists developing a model form historical data and applying that model on new data. Data mining can also be seen as a combination of tools, techniques and processes in knowledge discovery. In other words, it is uses a variety of tools ranging from classical statistical methods to neural networks and other new techniques originating from machine learning and artificial intelligence.

Data Mining Processes

Data mining process has four major steps. These are: data selection, data transformation, data mining and result interpretation. 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 sample data selection is important. In data transformation step three things should be considered to perform the data transformation (the task, data mining operations, data mining technique). In Data mining Step the desired information is extracted by using one or more techniques on the transformed data. In the last step according to the goals seated and decision-support task the researcher finally analyze and interpret the minded information.

The KDD Process

The basic steps of data mining for KDD are: defining business problem, creating a target dataset, 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 [40]. All KDD processes are illustrated in Figure: 3.1. [41], and the shored description of the processes are listed below.

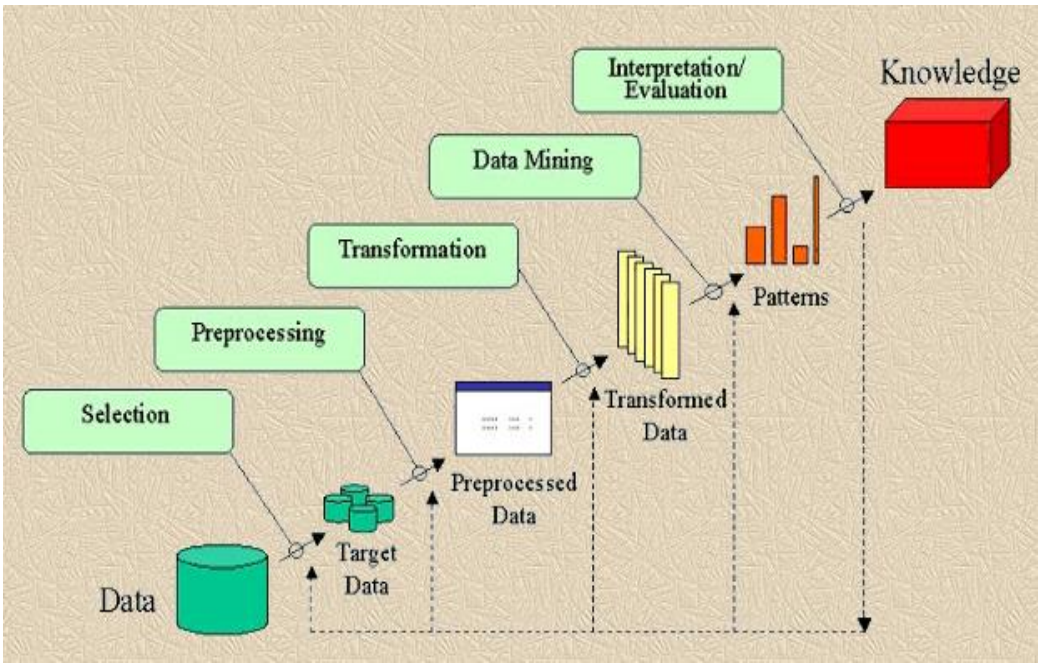


FIGURE 2-1 OVERVIEW OF THE STEPS ESTABLISHING THE KDD PROCESS [1]

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 objective and understanding.

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

Data cleaning and pre-processing: Tasks like removing noise or outliers if any, collecting the necessary information to model of account for noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes.

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.

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.

Choosing the Data Mining Algorithms: Selecting methods to be used for searching patterns in data and matching a particular data mining methods with the overall criteria of the KDD process are the major activities in this step.

Data Mining: It is all about searching for patterns of interest in a particular representational form of a collection of such representations. These representations include classification rules or trees, regression, clustering, sequence modeling, dependency, and line analysis.

Interpretation: In interpreting the discovered patterns and returning to any of the previous steps is a possibility.

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.

2.2.2. Data Mining Technologies

Data Mining Models

For this particular research, different data mining algorithms are used [42] and Random Forest is the best method for IRS fraud detection. ZeroR and RJ48 are also used for this research. The resulted models from the above algorithms are compared to propose the best algorithm for this study. Further discussion on the mentioned techniques and algorithms is the focus of this chapter.

CRISP-Data Mining Model

As described in introduction section CRISP-DM model is industry standard model and it has clear and easy follow stages. The sequence of the six stages of

CRISP-DM stages are not rigid. As it is characterized in Figure 3.2, it is extremely complete model. All the stages are well organized, structured and defined. CRISP-GM modeling toll is independent of any data mining techniques. The CRISP-DM model process relationship is illustrated in Fig: 3.2.

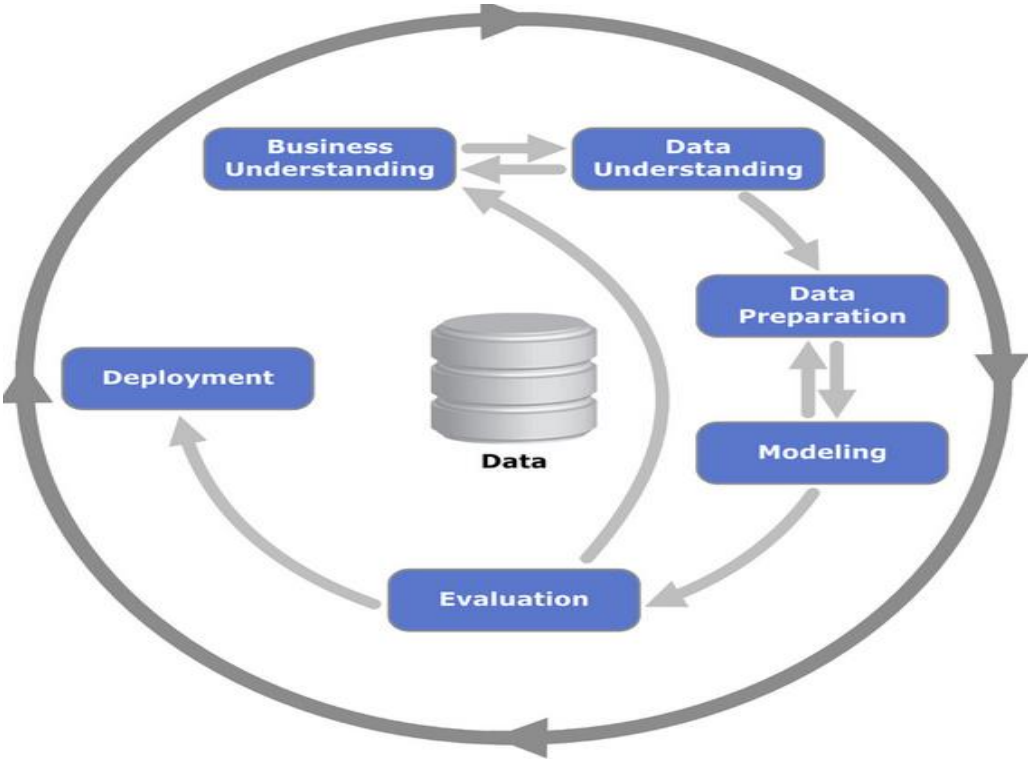


FIGURE 2-2 THE SIX STAGES OF CRISP-DM STAGES [40]

2.2.3. WEKA Data Mining Tool

Waikato Environment for Knowledge Analysis (WEKA) is a Java-based, open-source DM platform developed at the University Of Waikato, New Zealand. The software is free under GNU GPL 3 for non-commercial purposes [11]. WEKA offers four options for DM: command-line interface (CLI), Explorer, Experimenter, and Knowledge flow. The preferred option is the Explorer which allows the definition of data source, data preparation, machine learning algorithms and

visualization [16]. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

2.3.Related Works

In this section four selective related works are reviewed. These four related research papers are selected among several works. Two of them are ethio telecom cases studies, the first research is to detect SIMBOX fraud [43]. The second research works are to create datamining model for subscription fraud detection [7]. The last two related works are about international roaming fraud detection in telecommunication using datamining techniques [20, 36] but both of the cases are not ethio telecom related.

The First selected research paper "Predictive modeling for fraud detection in telecommunications: the case of ethio telecom" by Yeshnegus Getahe [43] has been conducted in SIMBox fraud detection. The objective of this paper is detect SIMBox fraud by using data mining technique. After deep statistical analysis of ethio telecom international incoming minutes of years 2003-2012, and based on the statistical result the writer found that around 40% international minutes are bypassed by SIMBOX and considered as local minuets. The writer using CRISP-DM process, decision tree and neural networks classification methods, and WEKA tool. The data is processed to find out the best predictive model to detect SIMBox fraud. Based on prepaid CDR, (DATA, SMS, GPRS service usage) and other fields. The researcher identified that PART algorithm from decision tree result has the best than neural network in accuracy. Therefore based on the analytical result decision tree is better predictive model for SIMBOX fraud detection. And the writer recommended other options like IMEI identification mechanisms for SIMBox device detection.

The second research work by the title "Constructing Predictive Model for Subscription Fraud Detection using Data Mining Techniques: the Case of ethio telecom" by Tesfaye Hadush [7]. The objective of this research is to create a model

to detect and predict Subscription fraud, and to improve fraud prevention and detection mechanism of ethio telecom

The researcher used WEKA Version 3.7.9 D.M. tool to create a model, and classification algorithms like J48, PART, Random forest and ANN are used, also oracle DB and excel are used for data preparation.

Based on 25,000 prepaid subscribers SIX months CDR the researcher found out that data mining technique is a proposed method for fraud identification and as the experiment based on the experiment result Random Forest algorithm is perform better. Subscription fraud is one of the cause of international roaming fraud but the researcher not consider international roaming situation.

The third research paper is "Fraud in roaming Scenarios, An Overview" [36] by Gebriel Macia and Jesus E. Diaz The objective of the paper is to develop a model to detect subscription fraud based on behavioral pattern of mobile customers' usage. By reviewing the most important notions about how roaming service function, main telecom service vulnerable areas are described, classifies and finally proposed a methodology by raise a series of questions that needs to be answered to combating roaming fraud.

Possible fraud protection strategies are opposed like, provision of subscription fraud, use of NRTRDE for fastest data collection, using fraud management tools (FMS), take timely action. Also a methodology has been proposed to examine the current (2010) status of the roaming fraud problem. This paper explain roaming related fraud types but data mining tools are not used tool to predict roaming fraud patterns.

The fourth research is also other international roaming fraud related work with a title "Roaming fraud: assault and defense strategies" [20] by Gabriel Macie University of Granada, Spain. The objective of the article is first presents the major concerns regarding telecom service and network security threats, and proposes a classification method for this type of attack and also highlighting the necessity for the different players involved to take joint action.

The writers using quantitative research methodology, by reviewing and by surveying fraud techniques and protection policies with the purpose of clarifying and identifying the main questions. The research findings are how the technologies in this fields are not mature, possibly because most work has been aimed towards data collection mechanisms, learning other aspects unconsidered.

Chapter Three

3. Data Mining Methods

3.1. Classification Methods

Classification is data mining task that maps the data into predefined groups and classes [42]. It is also called supervised learning. It consists of two steps which are model construction and model usage [42]. In model construction which consists of set of predetermined classes and for model usage the known label of test sample is compared with the classified result from the model.

For this particular research, data mining classification technique which are J48 Decision tree, Random Forest and ZeroR classification techniques are used.

3.1.1. J48 Decision Trees

Decision tree is one of mostly used data mining techniques for fraud detection [42], because its model is easy to understand. In decision tree technique, the toot of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to set of questions or conditions that help to determine the data so that can make final decision. Decision tree models are constructed in a top-down recursive divide-and-conquer manner [42]. J48 decision tee algorithm have adopted this approach. The training set is recursively partitioned into smaller subset as the tree is being built. J48 decision tree algorithm is a predictive machine learning model that decides the target value of a new sample based on various attribute values of the available data [42]. The algorithm performs the following sequence of steps to accomplish its classification task. First Check for base cases, then Find normalized information gain, after that Select the attribute with the highest normalized information gain and Create decision nodes.

Finally Recourses on the sub lists obtained by splitting and add those nodes as child node. J48 algorithm has various advantages [42]. Some of them are gains of balanced flexibility and accuracy, capability of limiting number of possible decision points and higher accuracy.

3.1.2. Random Forest (RF)

Random forest is a combination of tree predictors that each tree depends on the values of a randomly selected vector sample and it distributes equal values of vector samples to all of the trees in the forest [44]. The strength of individual trees in the forest and the correlation between them determines the generalized error of a forest and its tree. Random forest algorithm performs the following sequence of steps to accomplish its classification [44]. First it choose "T" number of trees to grow, then choose "m" number of variables used to split each node. "m<M" (Where "M" is the number of input variables and "m" hold constant while growing the forest, the next step is Grow "T" trees and the final step is classify points to collect votes from every tree "T" in the forest and then use majority voting to decide the class label [44].

Random forest algorithm has various advantages [44]. Some of them are its accuracy and ability of running on large data bases, Capability of handling thousands of input variables without variable deletion and fast learning, having an effective method of estimating messed data and highest maintenance accuracy and ability of saving generated forests for future use.

3.1.3. ZeroR (Zero Rule)

ZeroR classifier predicts the majority of class in training data. It predicts the mean for numeric value and mode for nominal class [42].It is the simplest classification method which relies on the target and ignores all predictors. ZeroR predict the majority category class correctly and useful for determining a baseline performance for other classification methods.

Chapter Four

4. 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 three, CRISP-DM process model is followed in order to come up with the desired output.

4.1. Ethical Standard

The required data is collected from ethio telecom Information Systems (IS) division. To accumulate the data, support of cooperative letter written from Addis Ababa University is used. The letter was directed to concerned department from chief executive officer (CEO) of the company and redirected to IS Chief Information Officer (CIO).

4.2. Business Understanding

As the researcher have the domain knowledge it is not difficult to understand the business area.

4.3. Understanding of the Data

The data is collected from different Departments of IS Division (IT operation and IT and Network Security). Roaming CDR data is collected from PRM (Partner Relation Management) database which is deployed in recent Telecom Expand Project (TEP) project by Huawei for interconnect and roaming data. IT Operation teams briefly describe the CDR attributes and there importance. In addition to

description of the data they give Partner Relation Management (PRM) documents. Since 2015, SYNVERS is a company working for ethio telecom as a data clearing house agent [31]. Data clearing house agents are working with telecom operators to collect International Roaming CDR data from all roaming partners and settle the interconnect fee. And sending high usage reports for the telecom operators. So that SYNVERS collect roaming CDR from all ethio telecom partners and transfer in less than two hours [20, 31]. There are CAMEL and NRTRDN agreements signed to get the data per the agreements. Getting Roaming CDR in a real time base minimize roaming fraud. It is also discuss on literature review section that NRTRDN and CAMEL agreements are originated to minimize roaming fraud.

The Call Detail Record (CDR) data holds a lot of attributes. It records each and every record of calls made by the customer with details like calling number, called number, calling time, duration, number of calls amount charged, cell site information, caller and receiver location information, service type (voice, SMS, Data, MMS), for prepaid mobile users balance related information. All the above descriptions are illustrated in the below figure Fig: 4.1.

The upper layer of the data distribution of PRM system shows that the dual link between different systems. Customer Relation Management, I two thousand (I2000) system, billing management system, partner system, business intelligent (BI) system and FMS. The lower layer shows how the clearing house, network devices and mediation center interact for data collection and synchronization.

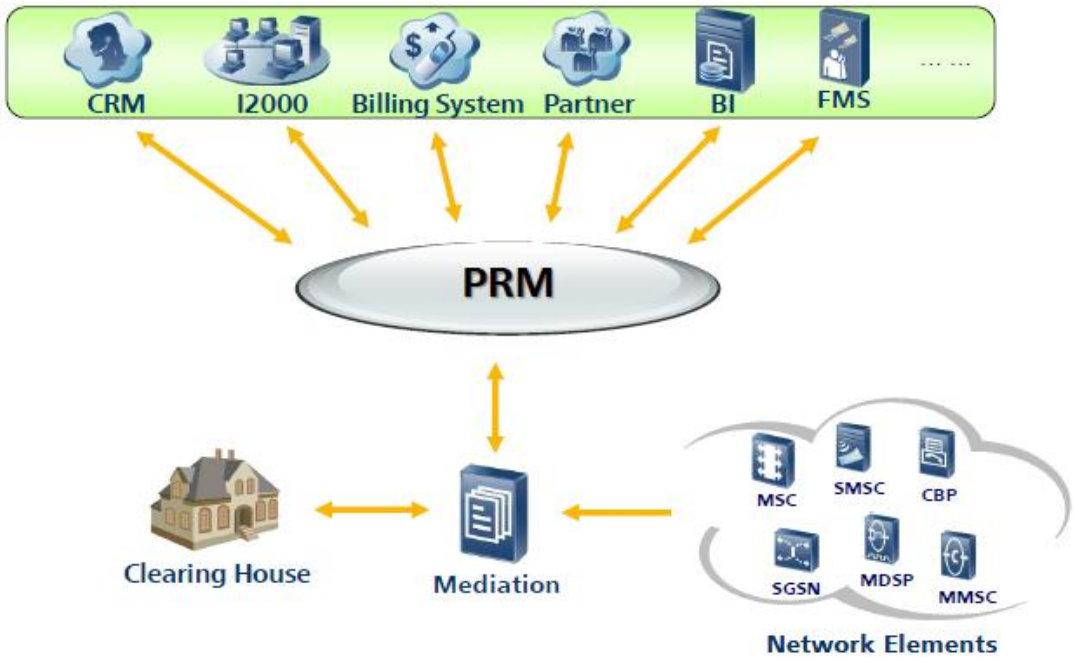


FIGURE 4-1 PRM NETWORK STRUCTURE [45]

PRM CDR holds all roaming each and every records of calls made by the customer with details like calling number, called number, date and time of call, duration, amount of charged, MSC number, location of calling number, roaming partner name, Roaming type, call type, called number region, Home operator, International Mobile Subscriber Identity Number (IMSI), MISIDN are some fields among the 117 fields of PRM Roaming CDR.

Electronic Customer Accusation Form (e_CAF) data base holds customer information such as customer name, address, number of SIM cards which are used with the same customer. Customer Relation Management (CRM) database holds customer information with service type with status. Convergent Billing System (CBS) handle all the billed CDR with all billing related attributes.

4.3.1. Initial Data Collection

Due to the size of the data, different query techniques are applied in order to reduce the size. The data is queried from the PRM database based on the requirement given by the researcher. From PRM active database holds six months CDR it is around 32.2 GB and have 117 attributes.

4.3.2. Data Description

As roaming CDR contain a lot of information, it was not easy to select the specific attributes and fraud may happen in every aspect. In addition to PRM database, Electronic Customer Accusation Form (e_CAF) database is used to get customer profile related data and CRM for product and service related information and Convergent Billing System (CBS) for billing related information are accessed. Four months one million Roaming CDR with 32 fields are extracted from the database. From these records 500,000 records with 12 attributes sampled carefully. Table 4.1 shows the selected attributes (12 fields) with datatype and description. e_CAF, CRM and CBS information is also used to identify the mandatory fields and records.

SN.	Field name	Data type	Description
1	IMSI	Number	International Mobile Subscriber Identity number
2	MSISDN	Number	Mobile Station International Subscriber Directory Number
3	HOMEMANAGER	Char	Home operator
4	HREGION	Number	Home region
5	OTHERTELNUM	Number	called number

6	THIRDTELNUM	Number	third called number/ divert number/ call forward
7	STARTTIME	Date	call start time
8	DURATION	Number	measured time
9	MSCID	Number	Mobile switching center ID
10	MSRN	Number	Mobile Subscriber Roaming Number (Roaming Test Number)
11	SERVICE_TYPE	Char	GPRS, MMS, SMS, Voice
12	Fraud	Char	Yes or No

TABLE 1 *SELECTED FIELDS WITH DESCRIPTION*

4.3.3. Data Quality Assurance

The CDR data which collected from PRM holds all services data such as SMS, VOICE, GPRS in as a receiver and sender. These different services have different behavior. For instance all cell related attributes for GPRS services are null value in the other hand all byte up and byte down attributes are empty for voice and SMS services. Also all charging, duration and fee related attributes are null value for SMS (SMS is not charge on International Roaming scenario) services. This and mixed data type behavior makes the data preparation difficult.

One million Records with 32 attributes data were collected but all are not relevant for this study. Regarding to the CDR attribute and data size the computer and Weka tool were not able to process all the data. Data proceeding tools like Microsoft excel, 010 editor and ultra-edit are used to check the data quality and for editing. But, it was not easy to use these tools for such data. Therefore, sample data 500,000 records with 12 attributes are sampling with care.

4.3.4. Data Preparation

Data quality is the main input to get relevant output. If the organization data has no quality all operational reports, managerial decisions, marketing predictions are wrong and in the worst case the business may collapse. To implement data mining technique, there are data preprocessing tasks like data cleaning, data integration, data transformation and data reduction techniques are important.

4.3.5. Data Cleaning

As mentioned earlier, the first data cleaning method is the way of reducing redundant roaming CDR attributes. Attributes with the dame information is eliminated. For instance, other telephone number and called number has the same information. The other Data cleaning is made by removing attributes having similar (static) values like currency (ALL in USD), zero and one's values , tax rate (the same for all).

4.3.6. Data Integration and Transformation

Data integration is made for start time, number of calls per day, sent and received SMS's, used GPRS service, called and caller country, call duration and number of cell ID records. The integration is made for such attributes from e_CAF, CRM and CBS data with PRM CDR.

4.3.7. Data Reduction

The size of the data was originally huge even the data queried by the researcher direction. The total data which are dumped to the computer is 1 million records with 32 fields. For this research only representative samples 500,000 records with 12 attributes are selected. As described in the previous section the size of the data was originally big and requiring a server for processing. Six months compressed data was 32 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 with criteria's such as roaming subscriber who call different countries, long call durations, High GPRS usage, send SMS messages to different destination and call to known International Revenue Share Fraud (IRSF) Numbers. By doing so manageable size 500.000 records and 12 attributes are selected.

4.3.8. Data Formatting

Before dealing with the data modeling the data set has to be formatted in a manner that suites the tool to be used for modeling. In this study WEKA 3.8.0 and MATLAB tools are used. MATLAB pattern recognition and classification tool is used by feeding different data. The MATLAB tool required binary format for the target data. The WEKA tool require the file format to be comma delimited CSV or 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. This selection method is applied for all algorithms used in this study.

Facts from Sample data

A total of 500,000 records with 12 fields are selected for this study. The different algorithms from classification model J48, RF, and ZeroR are used. The sampled data is used to build different models for detecting fraudulent calls. Out of 500.000 roaming CDR 29073 are for voice CDR. Voice call duration "Duration" field contains the duration of the call in seconds, the minimum, average and maximum values are 1, 124.78 and 1800 seconds (1800 seconds or 30 min is the maximum duration that one record can hold). On this voice CDR the maximum calling duration is 3600 seconds/ one hour which is hold 2 records.

Chapter Five

5. Experimentation and Modeling

5.1. Modeling and Discussion

In the modeling phase of this research, three classification algorithms are used. The algorithms include decision tree J48, rule based ZeroR and Random Forest (RF) tree based are used. The rules are generated from classification using decision tree algorithm.

The experimentation is made using WEKA data mining tool version 3.9 Different experiments are made using 12 attributes, using only non-determinant attributes. The experimentation is made also by gradually reducing the number of determinant attributes in the domain area. Such experiments are conducted using J48, ZeroR and Random Forest (RF). 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 above three algorithms. A print screen image that shows the selected attributes in WEKA data mining tool interface is attached in Appendix I.

5.1.1. Classification Modeling

In this study, the classification models are three types. The first model uses J48 and random forest algorithm from tree based, the other uses ZeroR algorithm from rule based. The algorithms for two decision tree and Rule based 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.

5.1.2. Decision Tree Modeling

As it is thoroughly discussed in chapter three of this research, decision tree is the most commonly used for prediction and classification purposes. On top if this, it does not require prior information about the data to be classified or predicted and the rules are also used for prediction purpose.

These experiment for decision tree and rule based are 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 to Table 5.6. But, the detail experimentation results using J48 algorithm, ZeroR, and Random forest with 12 attributes by varying test options and parameters are summarized in Appendix III shows the result of Random Forest in Training Set test. Which is the best output for this research.

In experiments the resulted models summaries using 12 attributes. The change in the test mode and number of attributes has effect on the accuracy of the model, time taken to build the model. The models resulted by changing the classifier test options, using training test, cross-validation and percentage split are used to compare the modes with each other.

J48 Experimentation

Experimentation 1 using J48 Algorithm with 12 attributes

Experiments	Algorithm	Number of Attributes	Test Modes	Time taken to build the model (sec)	Accuracy (%)
1	J48	12	Cross-Validation 10-fold	0.17	98.9899
		12	Percentage split 66%	1.02	99.7013
		12	Use training set	1.05	98.9796

TABLE 2 *J48 RESULT WITH DIFFERENT TEST MODES*

This experiment result of J48 with 12 attributes and three test mode shows that percentage split has better accuracy 99.7013 than cross validation and training set. J48 percentage split test confusion matrix classify fraud and non-fraud is as below.

```
    a    b   <-- classified as
 3646    1 |    a = No
   10   26 |    b = Yes
```

ZeroR Experimentation

Experimentation 2 using ZeroR Algorithm with 12 attributes

Experiments	Algorithm	Number of Attributes	Test Modes	Time taken to build the model (sec)	Accuracy (%)
1	ZeroR	12	Cross-Validation 10-fold	0.11	98.1392
		12	Percentage split 66%	0.11	98.1501
		12	Use training set	0.06	98.1392

TABLE 3 *ZEROR RESULT WITH DIFFERENT TEST MODES*

This experiment result of ZeroR with 12 attributes and three test mode shows that percentage split has better accuracy 98.1501 than cross validation and training set. And its confusion matrix classify fraud and non-fraud as bellow.

```
a      b   <-- classified as
0      3235 |   a = Yes
0  171644 |   b = No
```

Random Forest Experimentation

Experimentation 3 using RF Algorithm with all attributes

Experiments	Algorithm	Number of Attributes	Test Modes	Time taken to build the model (sec)	Accuracy (%)
1	RF	12	Cross-Validation 10-fold	1.11	99.8722
		12	Percentage split 66%	1.06	99.8154
		12	Use training set	1.01	99.9891

TABLE 4 *RANDOM FOREST RESULT WITH DIFFERENT TEST MODES*

This experiment result of RF with 12 attributes and three test mode shows that training set shows better accuracy 99.9891than cross validation and training set. Random forest get the best output in training set test, and the confusion matrix shows there is 100% accuracy because no data classified as not fraud while it is fraud.

```
    a    b   <-- classified as
 9114    0 |    a = No
    1   92 |    b = Yes
```

5.1.3. Evaluation

Evaluation of the model is made based on the objective of the research. To identify the best Roaming fraud detection and prevention model from different data

mining algorithms.to meet the objectives predicting fraudulent calls coming through International Roaming Fraud will help to minimize the international interconnect revenue lose.

The resulted classification models of both tree and rule based summarized as follows. Random Forest show the accuracy of 99.9891 % while the other two algorithms less accuracy than RF. Decision tree based algorithm J48 highs accuracy is 99.7013 % and rule based ZeroR algorithm maximum accuracy 98.1501 % both accuracies are based on Percentage split test condition. 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 international calls.

5.1.4. 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.

5.2.Discussion

On this section the research questions which are raised on the first chapter of this research are discussed.

RQ1. What possible fraud prevention and detection mechanisms can be setup in international roaming fraud?

Based on the literatures which are reviewed for this specific research there is no single fit all solution for IRF detection. The main detection methods are Proper Roaming Agreement, NRTRDE, Fraud Management

System (FMS), IRSF number blocking, HUR, Identify Fake Identity and machine learning approaches (Data mining techniques).

Proper Roaming agreement with other operators is crucial, because only roaming operator with proper agreement is enable in the network. So the network is block for other operators which has no agreement. It is also important to settle interconnect fee. **NRTRDE** is first introduce by GSMA to reduce IRF by exchange near real time data between operators [22]. It helps the operators to have the CDR data of its roaming subscriber for quick action based on the usage (Voice, and data usage). Prompt action is important because roaming is the main expansive telecommunication service.

Using **FMS** to analyze CDR data to detect fraud is preferable but update the FMS rules and conditions is important. Data mining methods can be integrated with FMS to improve FMS roaming fraud detection technique. For instance the model which is created on this research can improve Roaming fraud detection performance of FMS.

Blocking High risk International Revenue Share Fraud (IRSF) numbers at international gateway is also prevent IRF. The other method is exchange **High usage Reports (HUR)** in less than two hours with in roaming partners [22]. **Identify Fake Identity** is important to know the owner of Subscriber Identity Module (SIM) cards, subscription fraud through fake identity is the main roaming fraud enabler. Fraudsters get postpaid SIM Cards with roaming service to make international calls with no intention to pay. So early identification helps to detect the. A lot of researches are conducted to detected roaming fraud using Machin learning approaches (**Data mining techniques**) [46, 20].

RQ2. What are the methods used in ethio telecom to detect and prevent roaming fraud?

For roaming fraud prevention ethio telecom using proper roaming agreement with more than 500 roaming partners [8]. The agreement is made based on GSMA Roaming partner standard document [47]. The roaming data and other interconnect issues are exchange through SYNVERS [31]. SYNVERS is a data clearing House Company working with telecom operators. So that SYNVERS send near real time CDR data from and all operators including HUR alerts.

FMS is also used to detect roaming fraud in ethio telecom. FMS collect Roaming CDR from PRM database and data analyzes is made based on preconfigured thresholds (Rule based). The other method is blocking high risk IRSF numbers at core networks.

RQ3. Which Data Mining algorithm can be more suitable for the purpose of predicting International Roaming Fraud?

International roaming fraud is the highest cause of revenue lose for telecom operators because roaming service is the main expensive telecom service. Fraud detection using data mining is an effective method for many fraud types.

To predict the suitable data mining technique the researcher collect one million roaming records with 32 attributes. From these records 500,000 records with 12 attributes is carefully sampled. The sample data set has been preprocessed and prepared in a format suitable for the DM tasks and the WEKA 3.9 tool is used for the study. Three classification algorithms J48, ZeroR, and Random forest are compared with different test methods. The training set test result shows Random Forest algorithm registered

better performance of 99.9891% and the other algorithms J48 98.9796% and ZeroR result is 98.1392 %. Random forest get the better performance accuracy running with training set test environment. So that Random Forest (RF) algorithm is the suitable algorithm to detect international roaming fraud.

Chapter Six

6. Conclusion and Recommendations

6.1. Conclusion

Telecommunication fraud begins since the commercial telecommunication service started. Detecting and prevention of telecommunication fraud become the main task of telecom operators, because fraud techniques are dynamically changed. That means if the new detection method is implemented the fraudsters also change their behavior and techniques. So that telecom operators should also frequently update their detection mechanisms and try to identify the fraud loopholes in telecom ecosystems.

International Roaming Fraud is the gate for other frauds, because the service is vulnerable by different type of attacks because of geographic location difference of the mobile user and home operator. The objective of this study was to develop model for detecting and predicting of international roaming fraud using data mining techniques.

This study focused on introducing a new predictive model for roaming fraud detection using decision tree and rule based classification algorithms such as random forest, J48 and ZeroR. Hence, the result shows that the random forest meet the highest accuracy 99.8154% with around 0.0109% false positive rate. Therefore Random Forest algorithm is the best algorithm to detect international roaming fraud with that of J48 and ZeroR.

6.2.Recommendations

There are a lot of tools and techniques are used to detect international roaming fraud. Some of the methods are, making proper **wholesale inter-operator agreement**, exchange Roaming data through **Data Clearing House service, NRTRDE** which are the mechanism to exchange data within 30 minutes of after the roaming call, using **Fraud management tool** by applying different data analysis techniques and algorithms. Even if all the above mechanisms are practiced International roaming fraud is still the main door for fraudsters. A lot of fraud attacks are committed while roaming service is enabled. Some of them are International Revenue share fraud (IRSF), subscription fraud, internal fraud, premium rate service fraud (PRS). SIM cloning, SMS spam.

The following areas can be future research areas:

- Other research can made with more months data of Roaming-In CDR with more number of attributes and try with different data mining algorithms.
- One can conduct similar researches on outbound roamers (Roaming-Out fraud)
- International roaming related fraud types such as cloning, IRSF, and High usage fraud during Roaming are a possible researches areas.
- For this research only CDR Data from different systems are used but other researcher can conduct the same research using Signaling data.
- One of the limitation on telecom network and a gate for roaming fraud is international SS7 security leakage. So improving SS7 protocols can be a research area.

But, there is a need to check for data availability before propose any of the above research areas.

References

[1] B. Richard, V. Chris and R. W. Allan, "Fraud Detection in Telecommunications:History and Lessons Learned," *Technometrics,* vol. 52, no. 1, pp. 20-33, 2010.

[2] E. Sutherland, "International Mobile Roaming: Competition, Economics and Regulation," papers.ssrn.com, 10 Jun 2010. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1622759. [Accessed 20 10 2015].

[3] K. Souvik, J. Sanket and M. Vijay, "Evolving Early Combat Systems in Next Generation Telecom Fraud: Catch Them Young," in *IBM*, UK, 2013.

[4] D. Howells, D. V. Scharf-Katz and S. Padraig, "TELECOM FRAUD 101:," ARGYLE DATA, San Mateo, 2010.

[5] H. Abdikarim, I. Subariah and Roseli, "Detecting SIM Box Fraud Using Neural Network," *IT Convergence and Security 2012,* pp. 5-69, 24 4 2013.

[6] CFCA, "Fraud Loss Survey | CFCA - Communications Fraud Control Association," www.cfca.org, New Jersey, 2015.

[7] T. Haddish, "Constructing Predictive Model for Subscription Fraud Detection Using Data Mining Techniques," *Diss. AAU,* 2013.

[8] ethio telecom, "http://www.ethiotelecom.et/," ethio telecom, 28 1 2012. [Online]. Available: http://www.ethiotelecom.et/?q=aboutus. [Accessed 20 10 2015].

[9] COMMUNICATIONS FRAUD CONTROL ASSOCIATION (CFCA), "2011 Global Fraud Loss Survey," CFCA, new jersey, 2011.

[10] G. Prepizyk and Dawson, "Information security and privacy," in *12th Ausralasian conference*, Townsville, 2007.

[11] H. Mark, F. Eibe, . H. Geoffrey, P. Bernhard, R. Peter and W. Ian, "The WEKA Data Mining Software: An Update," *ACM SIGKDD explorations newsletter 11,* vol. 11, no. 1, pp. 10-18, 2009.

[12] D. Yuxiao, K. Qing, C. Yanan, W. Bin and W. Bai, "TeleDatA: data mining, social network analysis and statistics analysis system based on cloud computing in

telecommunication industry," in *International workshop on Cloud data management*, New York, 2011.

[13] G. Phil and H. Mark, "Classification, Detection and Prosecution of Fraud on Mobile Networks," in *Proceedings of ACTS mobile summit,* , Sorrento, 2012.

[14] Communication Fraud Control Association (CFCA), "CFCA 2013 Global Fraud Loss Survey," CFCA, Roseland, 2013.

[15] O. Ogundile, "Fraud Analysis in Nigeria's Mobile Telecommunication Industry," *International Journal of Scientific and Research Publications,* vol. 3, no. 2, pp. 2250-3153, 2013.

[16] I. Mohammad, M. Akhter and A. Gulam, "Detecting Telecommunication Fraud using Neural Networks through Data Mining," *International Journal of Scientific & Engineering Research,,* vol. 3, no. 3, pp. 2229-5518, 2012.

[17] K. Souvik, J. Sanket and M. Vijay, "Evolving Early Combat Systems in Next Generation Telecom Fraud: Catch Them Young," in *www.academia.edu*, Chicago, 2013.

[18] K. H. John Shawe-Taylor, "Detection of Fraud in Mobile Telecommunications," Information Security Technical Report, London, 1999.

[19] K. Ledisi, N. Domaka and U. Edikan, "Telecommunications Subscription Fraud Detection using Artificial Neural Networks," *Society for Sc ience and Education United Kingdom,* vol. 3, no. 6, pp. 19-33, 2015.

[20] Gabriel Maciá-Fernández, "Roaming fraud: assault and defense strategies," in *CITEL Workshop on International Roaming Services*, CITEL, 2008.

[21] FEDERAL DEMOCRATIC REPUBLIC O'F ETHIOPIA, *TelecomFraud Offenc.eProclamation,* Addis Ababa: Federal Negarit Gazeta, 2012.

[22] GSMA , "Information Paper Overview of International Mobile Roaming," GSMA Press Office, London, 2012.

[23] W. Moudani and C. Fadi, "Fraud detection in mobile telecommunication," *Lecture Notes on SOftware Engineering,* vol. 3, no. 4, pp. 2319-8753, 2013.

[24] COMMUNICATIONS FRAUD CONTROL ASSOCIATION (CFCA), "2009 Global Fraud Loss Survey," www.cfca.org/fraudlosssurvey, New Jersy, 2009.

[25] K. Yufeng, L. Chang and S. Sirirat, "Survey of Fraud Detection Techniques," in *IEEE international conference*, 2004.

[26] GSM Association, "https://infocentre2.gsma.com," 16 Dec 2014. [Online]. Available: https://infocentre2.gsma.com/gp/wg/FSG/OfficialDocuments. [Accessed 24 April 2016].

[27] GSMA FASG, "Fraud Manual," www.GSMA.com, London, 2012.

[28] TM-forum, "Tmforum.org," 27 11 2014. [Online]. Available: https://www.tmforum.org/resources/standard/gb954-fraud-classification-guide-v2-4/. [Accessed 7 6 2016].

[29] TM Forum, "GB954 Fraud Classification Guide V2.4 - TM Forum," TM Forum, New York, 2013.

[30] Synverse, "www.syniverse.com," 10 4 2014. [Online]. Available: https://www.syniverse.com/Synivers. [Accessed 6 7 2016].

[31] Syniverse, "Syniverse Risk Management," Syniverse Proprietary, UK, 2016.

[32] argyledata.com, "www.argyledata.com," 26 10 2010. [Online]. Available: https://www.argyledata.com/company/. [Accessed 7 6 2016].

[33] ITU, "INTERNATIONAL MOBILE ROAMING REGULATION – AN INCENTIVE FOR COOPERATION," in *International Telecommunication Union*, Thailand, 2008.

[34] J. Stewart, Interviewee, *Roaming Fraud: The Importance of Real-Time Data Exchange and Analysis.* [Interview]. 13 September 2011.

[35] D. Baker, "Synergy The SYNIVERSE BLOGAZINE," 15 November 2014. [Online]. Available: http://synergy.syniverse.com/2014/11/stopping-mobile-fraud-roaming-gateway/. [Accessed 23 04 2016].

[36] F. Macia, P. Gabriel and T. Garcia , "Fraud in roaming scenarios: an overview." *IEEE Wireless Communications,* vol. 16, no. 6, p. 88, 2009.

[37] P. Adrienne, M. Felt and E. Finifter, *A Survey of Mobile Malware in the Wild,* California: University of California, Berkeley, 2011.

[38] Subex, "Bypass Fraud- Are you getting it right?," www.subex.com, Paris, 2010.

[39] Y. Kou, L. Chang and S. Sirirat, "Survey of Fraud Detection Techniques," *IEEE international conference,* vol. 2, pp. 749-754, 2004.

[40] U. Fayyad, S. Gregory and Padh, "From data mining to knowledge discovery in databases," *AI magazine,* vol. 3, no. 17, p. 37, 1996.

[41] *Process diagram showing the relationship between the different phases of CRISP-DM,* 2017.

[42] S. Aher and L. Lobo, "Comparative study of classification algorithms.," *International Journal of Information Technology,* vol. 5, no. 2, pp. 239-43, 2012.

[43] Y. Getaneh, "Predictive Modeling for Fraud Detection In Telecommunications," *Diss. AAU,* 2013.

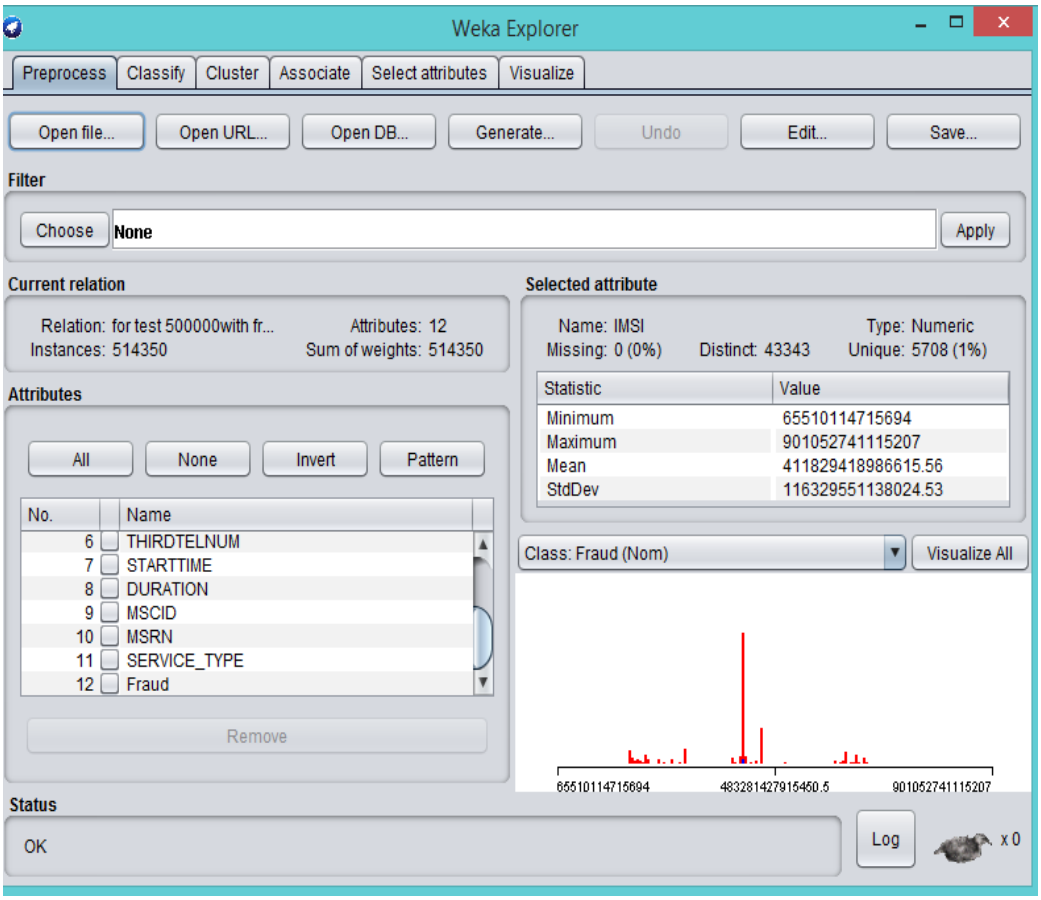
[44] J. Ali, K. Rehanullah, A. Nasir and Imra, "Random forests and decision trees," *International Journal of Computer Science ,* vol. 5, no. 9, pp. 272-278, 2012.

[45] *PRM System Overview Huawei Technologies,* 2014.

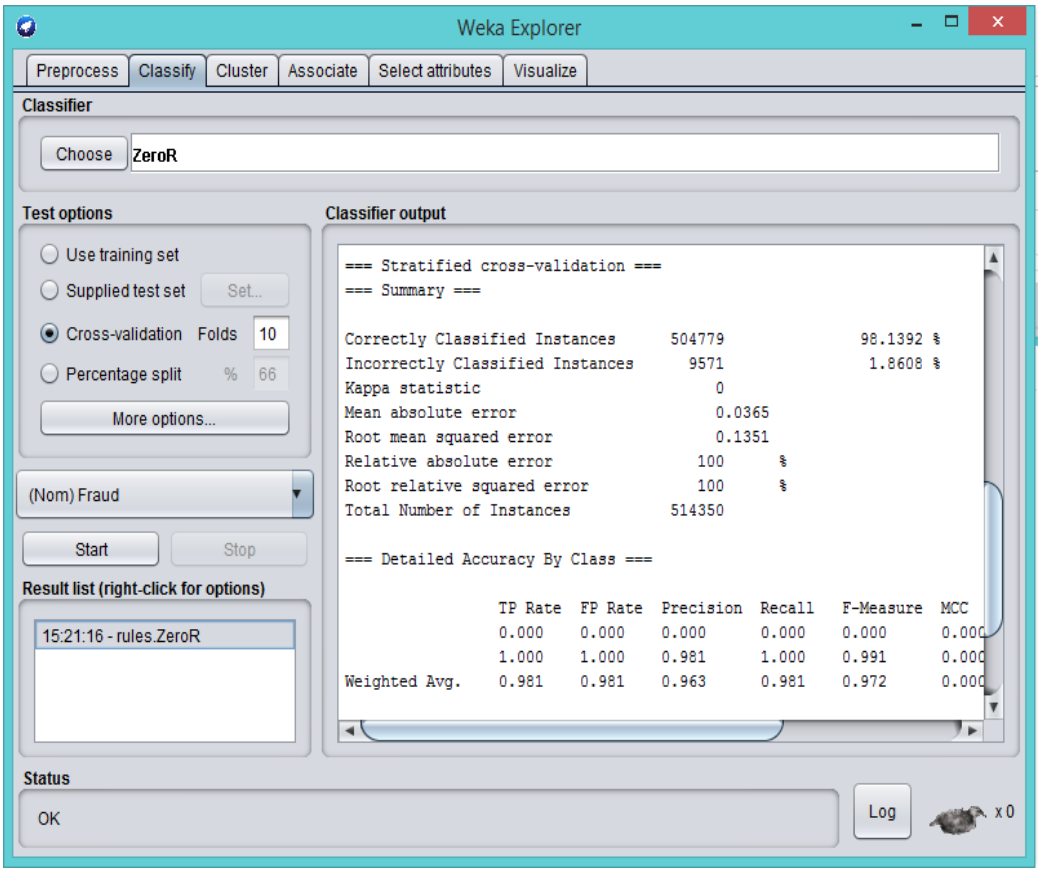
[46] A. H. Elmi, I. Subariah and S. Roselina , "Detecting SIM Box Fraud Using Neural Network," *IT Convergence and Security ,* pp. 575-582, 2013.

[47] GSMA assosation, "Official Document BA.20-Fraud Prevention Procedures," Feb 2015, 2015.

Appendices



APPENDIX 1 ***UPLOADED DATA 500,000 RECORDS WITH 12 ATTRIBUTES TO WEKA TOOL***



APPENDIX 2 *ZEROR RUNNING WITH CROSS-VALIDATION TEST METHOD*

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

Random forest of 10 trees, each constructed while considering 4 random features.
Out of bag error: 0.0022



Time taken to build model: 1.01 seconds

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

Correctly Classified Instances        9206               99.9891 %
Incorrectly Classified Instances         1                0.0109 %
Kappa statistic                          0.9945
Mean absolute error                      0.0015
Root mean squared error                  0.0152
Relative absolute error                  7.5258 %
Root relative squared error             15.2415 %
Total Number of Instances             9207

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.011      1         1         1          1        No
                 0.989     0          1         0.989     0.995      1        Yes
Weighted Avg.    1         0.011      1         1         1          1

=== Confusion Matrix ===

    a    b   <-- classified as
 9114    0 |    a = No
    1   92 |    b = Yes
```

APPENDIX 3 ***RANDOM FOREST RESULT WITH TRAINING SET TEST METHOD***