ADDIS ABABA UNIVERSITY SCHOOL OF GRADUATE STUDIES DEPARTMENT OF INFORMATION SCIENCE

Constructing Subscription Fraud Detection Model Using Machine Learning Algorithms: The Case of ethio telecom

By

Hailemeskel G/Tsadik

January 2021

Addis Ababa, Ethiopia

ADDIS ABABA UNIERSITY
COLLEGE OF NATURAL SCIENCES
SCHOOL OF INFORMATION SCIENCE

Constructing Subscription Fraud Detection Model Using Machine Learning Algorithms: The Case of ethio telecom

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

By

Hailemeskel G/Tsadik

January 2021
Addis Ababa, Ethiopia

ADDIS ABABA UNIERSITY
COLLEGE OF NATURAL SCIENCES
SCHOOL OF INFORMATION SCIENCE

Constructing Subscription Fraud Detection Model Using Machine Learning Algorithms: The Case of ethio telecom

By

Hailemeskel G/Tsadik

January2021

Name and signature of Members of the Examining Board

Tibebe Beshah (Ph.D.)	_____	_____
Advisor	Signature	Date
Million Meshesha (Ph.D.)	_____	_____
Examiner	Signature	Date
Melkamu Beyene. (Ph.D.)	_____	_____
Examiner	Signature	Date

Declaration

I, H/Meskel G/Tsadik, hereby declare that the work which is being presented in this thesis entitled "Constructing Subscription Fraud Detection Model Using Machine Learning Algorithms: The case of ethio telecom" is an original work of my own and performed with the support of my supervisor Dr. Tibebe Beshah. It has not been presented for any scholastic achievement in any University. All the sources of the materials used in this research have been properly acknowledged.

Signature: _____

H/Meskel G/Tsadik

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

Advisor's Signature: _____

Tibebe Beshah (Ph.D.)

Dedications

This thesis is dedicated to my beloved families, Ms. Meseret Solomon, Nahom, Nathnael and Etsube, who have been encouraging me all the time, to make my dreams success. This work is also dedicated to those who are in the side of me by providing everything for my success in my life.

Acknowledgment

In the first place, I would like to thank Almighty God for all he did the best throughout my life. He gave me the opportunity and knowledge to start and finalize this study adequately. Nothing was done without his blessing.

I respectfully would like to thank my research advisor Dr. Tibebe Beshah for his unreserved support and guidance to be successful in my research. He was available all the time to support his advisee. Even, in the current bad situation COVID-19, he was always available online and providing unlimited support. I never forget his friendly approach, Thank you again.

I would also like to extend my heartfelt appreciation to my beloved wife Meseret and kids Nahom, Tutuye and Etsub for your support and encouragement. Also, special thanks to the moral man Ato Solomon who have been providing me a moral all the time to be successful in my profession.

I have a special thanks to my colleague and senior classmate Ato Mesfin Worku for his constructive advises and sharing his best experiences regarding all the courses and researches.

Abstract

Nowadays, the advancement of telecom services are becoming an essential communication means for people's day-to-day activities. However, this development provides some appearances that motivate fraudsters. Telecom fraud is a serious challenge in telecommunication industries. It is a threat for telecom companies to lose some percent of their annual revenue and to provide poor quality of services for their customers. Subscription fraud is one type of fraud in today's telecom business. It is a common and major types of telecom frauds in which the usage category is in contradiction with the initial subscription type. The main objective of the fraudsters is to make money illegally or getting telecom services with the intention of not to pay for the service they used.

The purpose of this study is to construct a model which uses machine learning Algorithms to detect subscription fraud calls by using Call Detail Records (CDR) data. The general approach used to perform this research was a quantitative laboratory experimental research method. Three classification techniques of machine learning algorithms have been applied; which are, Random Forest (RF), Support Vector Machine (SVM) and Artificial Neural Network (ANN) multilayer perceptron algorithms. WEKA data mining tool has been used to build a model for predicting fraudulent calls.

The experimentation results of the work show that RF classifier performs better among the three algorithms with an accuracy of 99.46%. The major finding of this research is that, ten interesting factors used to identify fraudulent subscribers from legitimate ones. Some of the attributes such as, Subscribers total number of calls, number of unique called numbers, number of incoming calls, total international calls and ration of international total are important factors for domain expert practically practicing protecting the telecommunication frauds.

There are misclassification results happened because of false positive and false negative. In telecom fraud detection, the cost of a false negative is more expensive than a false positive because a false positive can be classified correctly after further investigation, but a false negative means that the fraudster has managed to stay undetected and can continue committing fraud. Therefore further research needs to be done to reduce false negative in identifying subscription fraudsters.

KEYWORDS: **Subscription Fraud, Fraud Detection, Machine learning, Support Vector Machine, Random Forest, and Artificial Neural Network**

Contents

Declaration vii

Dedications viii

Acknowledgment vii

Abstract viii

LIST OF FIGURES xiii

LIST OF TABLES xiv

LIST OF ABBREVIATIONS vii

INTRODUCTION 1

1.1. Background 1

1.2. **Statement of the Problem** 4

1.3. Research questions 5

1.4. Objective of the Study 6

1.4.1. General Objective 6

1.4.2. **Specific Objectives** 6

1.5. Scope and limitation of the study 6

1.6. Significance of the study 7

1.7. Organization of the study 7

CHAPTER TWO 9

LITERATURE REVIEW 9

2.1. Overview 9

2.2. Telecommunication services and Fraud 10

2.2.1. Telecommunications Services 10

2.2.2. Telecommunications Fraud 11

2.2.3. Subscription Fraud 14

2.3. Machine learning 16

2.3.1. **Machine Learning techniques** 16

2.3.2. Supervised algorithms 18

2.4. Related works 25

Summary of related works 28

2.5. Research Gap 30

CHAPTER THREE 31

Research Methodology 31

3.1. Overview 31

3.2. Quantitative Research Method 31

3.3. Research Model 32

3.4. Data set and Sampling 33

3.4.1. Data set 33

3.4.2. Sampling technique and sample size 34

3.5. Data Preprocessing 34

3.6. Classification methods for Learning 35

3.7. Evaluation Methods 36

CHAPTER FOUR 39

Business understanding and Data preparation. 39

4.1. Overview 39

4.2. Business Understanding 39

4.3. Data collection 40

4.4. Understanding the data 42

4.5. Data selection 43

4.5.1. Attribute selection 43

4.6. Data preprocessing 45

4.6.1. **Data cleaning** 45

4.6.2. Data Integration 46

4.6.3. Data Aggregation 46

4.6.4. Data Formatting 47

CHAPTER FIVE 49

Experimentation and Modeling 49

5.1. Overview 49

5.2. Model Building 51

5.2.1. Building a model using Random Forest algorithm 51

5.2.2. Building a model using SVM Algorithm 52

5.2.3. Building a model using ANN Algorithms 53

5.3. Model Evaluation 58

5.4. Discussion of the result 61

CHAPTER SIX 65

Conclusion and Recommendation 65

6.1. Conclusion 65

6.2. Recommendation 67

6.2.1. Recommendation for practice 67

6.2.2. Recommendation for future work 67

References 69

Appendix 73

a) Sample data snapshot 73

b) Weka screenshot 74

c) Experiment output for Training 74

d) Experiment output for Testing 87

LIST OF FIGURES

FIGURE 2. 1 SUBSCRIPTION FRAUD SCENARIO [5] 15
FIGURE 2. 2 DIFFERENT MACHINE LEARNING TECHNIQUES AND THEIR REQUIRED DATA [21] 17
FIGURE 2. 3 SUPPORT VECTOR MACHINE CLASSIFICATION [25] 19
FIGURE 2. 4 AN ENSEMBLE CLASSIFIER FOR RANDOM FOREST [28]. 21
FIGURE 2. 5 A PERCEPTRON WITH MULTIPLE INPUTS AND SINGLE OUTPUT [29]. 24
FIGURE 2. 6 A MULTILAYER PERCEPTRON NEURAL NETWORK [31] 25
FIGURE 3. 1 THE OVERALL RESEARCH DESIGN & METHODOLOGY [42] 33
FIGURE 3. 2 CONFUSION MATRIX PERFORMANCE MEASURE [7] 37
FIGURE 4. 1 SCREENSHOT OF RAW CDR DUMP FILE 41
FIGURE 4. 2 SCREENSHOT OF SAMPLE DATA IN CSV FORMAT. 48
FIGURE 5. 1 A VISUALIZATION OF THE DATA SET SPLITS 49
FIGURE 5. 2 MODELS COMPARISON USING 10- FOLD CROSS VALIDATION METHOD. 57
FIGURE 5. 3 MODELS COMPARISON USING PERCENTAGE OF SPLIT METHOD. 58
FIGURE 5. 4 MODEL SUBJECTIVE EVALUATION RESULT 61

LIST OF TABLES

TABLE 2. 1 SUMMARY OF RELATED WORKS 28
TABLE 4. 1 CDR FIELDS DESCRIPTION 43
TABLE 4. 2 SUBSETS OF SELECTED ATTRIBUTES 45
TABLE 4. 3 DESCRIPTION OF AGGREGATED AND DERIVED ATTRIBUTES 47
TABLE 5. 1 PERFORMANCE METRICS OF RF CLASSIFIERS 51
TABLE 5. 2 CONFUSION MATRIX FOR RF CLASSIFIERS 52
TABLE 5. 3 PERFORMANCE METRICS OF SVM CLASSIFIERS 53
TABLE 5. 4 CONFUSION MATRIX FOR SVM CLASSIFIERS 53
TABLE 5. 5 PERFORMANCE METRICS OF ANN/MLP CLASSIFIERS 54
TABLE 5. 6 CONFUSION MATRIX FOR ANN/MLP CLASSIFIERS 54
TABLE 5. 7 SUMMARY OF PERFORMANCE METRICS OF ALL ALGORITHMS 55
TABLE 5. 8 SUMMARY OF CONFUSION MATRIX OF ALL ALGORITHMS 56
TABLE 5. 9 EVALUATION RESULT OF BEST MODEL USING TEST DATASET. 59
TABLE 5. 10 SUBSCRIPTION FRAUD DETECTION MODEL EVALUATION QUESTIONNAIRES. 60

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
CFCA	Communications Fraud Control Association
CDR	Call Detail Record
CRISP-DM	Cross Industry Process for Data Mining
CV	Cross Validation
FMS	Fraud Management System
FFNN	Feed Forward Neural Network
IMEI	International Mobile Equipment Identity
MNO	Mobile Network Operators
MLP	Multi-Layer Perception
PRS	Premium Rate Service
RNN	Recursive Neural Network
RF	Random Forest
SIM	Subscriber Identity Module
SMS	Short Message Service
SVM	Support Vector Machine
TSP	Telecom Service Provider

CHAPTER ONE

INTRODUCTION

1.1. Background

The fundamental changes in the expansion of telecommunication industries have made it challenges to manage and detect telecom fraud activities. Therefore, to achieve encouraging outcomes, the problem of fraud needs to be managed with thoughtful and efficient attention. The introduction of new communications technologies and convergence of telephony with the Internet has added to its complexity [1]. Telecommunications or telecom industry has growing dramatically as the result of communication technology advancements. The number of telecom service users increased with the development of affordable telecom technologies. Similarly, Telecommunications or telecom fraud is fast growing field of criminal activity [1]. The expansion in telecommunication industries provides certain characteristics that motivate fraudsters. Fraud method and techniques are increased in parallel to this dramatic expansion [1]. This increase in fraudulent activity brings hard challenges to telecom operators. However, it is a must task to telecom operators to protect their customers and their revenue from fraudsters.

Telecom fraud is a serious issue in telecommunications sector, with incomes and services used by prearranged crime and terrorist networks [2]. Whilst reliable statistics are difficult to come by industry association, Communications Fraud Control Association (CFCA) evaluations are the total worldwide fraud loss estimates $29 billion yearly, according to a CFCA 2017 Global Fraud Loss Survey. Of this, over $7 billion is attributed to subscription fraud and account takeover. The worldwide telecommunication industry is worth $2.2 trillion [2]. This is 1.7% of the total net worth of the worldwide telecommunication industry. The percentage is relatively low, but the amount is high and that is why fighting fraud inside an operator has priority for a mature telecom operator [3].

In today's dynamic and often turbulent business environment and complex technologies, telecom operators have been under great pressure to seek out techniques to manage, control and protect their customers and revenues from fraudsters. Ethio telecom is fighting the fraud activities by having separate business units, which are Revenue Assurance (RA) and Fraud Management (FM). These units are part of the Information security division and typically reports to the chief

Information Security officer (CISO) and chief finance officer (CFO) of the company. Fraud management units fight frauds by deploying Fraud Management System (FMS) and by hiring experts of fraud detection field. Revenue assurance and members of fraud management team have full access to most of ethio telecom systems and network infrastructure. These units are directly accessing the following infrastructure of the company [4].

- Core (Backbone) network: this network represents the technical core of a company. It is a composite of systems that provide both wired and wireless access to different network services. They perform the authentication of their subscribers, call switching, and interact with international gateways and similar elements. Examples of the systems that are part of the core network include: Home Location Register (HLR), which serves as the primary database of the subscribers, and Mobile Switching Center (MSC), which routes different services between the subscribers, such as voice calls, Short Message Service (SMS), and conference calls.
- Billing and charging systems: these systems function under two different schemes, either prepaid or postpaid. Prepaid is when a subscriber is charged and billed in advance for a service. Postpaid is when a subscriber is using different services, then gets billed for them at the end of the billing cycle.
- Customer Relationship Management (CRM): These systems contain subscription details of the subscribers. Examples are name, address, Kebele ID number, and activation and deactivation dates. Also, these systems can be used to assist the compliance of the subscribers and check their legitimacy.
- Operations and business support systems (OSS and BSS): these systems mediate between the previously mentioned systems and are vital to run and monitor the company business correctly.

Telecommunication fraud is any activity by which telecom service is obtained legally/illegally with the intention of not to pay the required amount of money for the service acquired, abuse of the services or to gain finance [5].

Subscription fraud

Subscription fraud is the source of most fraud types [5]. It characterized by a fraudster using own, Stolen or false identity to get service. It includes getting the subscriber profile to signing up for the purpose of new service from telecom companies with a legal authorization. But the main intention of the fraudster is not to pay for the services used. Subscription fraud is often considered as uncollectable revenue instead of fraud. Telecommunication companies generally estimate that almost 40% of all uncollectable revenue are essentially subscription fraud [6].

In most cases, telecom subscription fraud is a contractual fraud [5]. This type of fraud proceeds is made through the normal use of a service with the intention of not to pay. In this situation, the fraudster works at level of each telephone numbers where all communications from this number is fraudulent and all actions in such cases are more irregular throughout the active period of the subscriber.

Subscription fraud is the starting point for many other telecoms frauds and as such it is known as the most harmful fraud types of all non-technical frauds [7]. Fraudsters do not just stop with obtaining legitimate service illegitimately; they usually use it as a precursor for many other fraud types such as International Revenue Share Fraud (IRSF) and Premium Rate Fraud (PRF) [7]. It is very risky in their individual right. It is very difficult to measure the real impact of this kind of fraud. This is because, the impact will not stop with profits loss only. The effects can be disastrous in terms of escalating complaints, poor customer experience, dissatisfaction among support staff, and declining vendors' confidence [5].

Subscription fraud can be divided into two categories [5]. These are Subscription fraud for getting finance and Subscription fraud for the purpose of self-usage by the fraudster. In the first category, the fraudster opens a small group where they start up a call center. The intention of the fraudsters is not paying for the service they used; instead, they sell the airtime in cash to individuals who intend to do a call with big discount rate for long-distance.

Telecommunication companies have common interest to extract an accurate profile for each customer based on customer call detail record (CDR) patterns. Customer profiles are not only valuable for detecting irregular behaviors of the subscribers but also mostly used for marketing

purposes and customer relationship management (CRM). These customer profiles are based on either CDR (e.g., number of outgoing and incoming calls, call duration time, call type) or customer demographic properties (e.g., age, gender, location) or both. Subscriber fraud detection methods divide telecom subscribers into two main categories of genuine and fraudulent customers [7].

Many researchers studied customer fraud detection in telecommunication companies using data mining techniques. Detecting insolvent customers [9], detecting subscription fraud [7], detecting sim-box fraud [14, 34] and detecting fax lines from telephone lines are some examples of recent studies. Fraud detection in telecommunication business is mostly done for mobile services. Methods like rule mining, clustering, Bayesian network, Neural network and decision tree are some examples in this context [5].

Detecting fraudulent customers is a classification problem that may be resolved by several data mining methods. These methods differ in terms of statistical techniques (e.g., regression family techniques), artificial intelligence techniques (e.g., decision trees, artificial neural networks), dimension reduction method (e.g., PCA, MDS), number of features included in the model, as well as feature-selection method (e.g., theory versus stepwise selection) [8]. In any case, a classifier should categorize each subscriber into genuine or fraudulent subscribers; though, the main problems in detecting telecom fraud are: (i) to deal with the large volume of call conversation in an effective method; (ii) to be efficient in classifying small percentage calls which are fraudulent; and (iii) to complete at a reasonably low cost [5].

Other research had been conducted on the construction of predictive model for subscription fraud detection using data mining technique in the case of ethio telecom in the year 2013 [8]. The research was conducted using 25,000 records only for the analysis and detection of subscription fraud in ethio telecom using data mining technique.

1.2. **Statement of the Problem**

Telecom industry in Ethiopia is a fast-growing sector of the economy with increasing numbers of subscribers. But telecom fraud has been a major challenge to the rapid growth of this industry as it has caused in telecommunication operators revenue loss and its subscriber's loss quality of service. Ethio telecom loses 52 million dollars from its annual revenue due to fraud [3]

Subscription Fraud (SF) is one of the toughest and most expensive revenue losses. This fraud is the leading fraud of all types in ethio telecom and 40 percent of revenue loss is caused by subscription fraud. Telecom fraud describes each attempt to use the telecom operator network without intention of paying for it. Currently, ethio telecom is using rule-based fraud management system to prevent and control all types of fraud. Fraud detection with a rule-based engine in telecom services causes a substantial loss of annual revenue, losses of customers, increase fraud as well as facing security gap. In this study, we intend to explore the effectiveness of machine learning based subscription fraud detection techniques grounded on subscriber usage data (CDR).

Efficient fraud detection and analysis systems can save telecommunication operators a lot of money and also help restore subscribers' confidence in the security of their transactions. Automated fraud detection systems enable operators to respond to fraud by detection, service denial and prosecutions against fraudulent users. The huge volume of call activity in a network means that fraud detection and analysis is a challenging problem.

In general, the more advanced a service, the more it is vulnerable to fraud. In the future, operators will need to adapt rapidly to keep pace with new challenges posed by fraudulent users. In addition, the number of actors involved in the provision of a service is likely to increase, making the possibilities for fraud to expand beyond the simple case of subscribers trying to defraud an operator. While conventional approaches to fraud detection and analysis such as rule-based systems based on thresholds for particular parameters may be sufficient to cope with some current types of fraud, they are less able to cope with the myriad of new possibilities. In addition, fraudsters can change their tactics fairly easily to avoid detection; for instance, systems based on thresholds can be fooled by keeping the call duration below that of the detection threshold.

Therefore, there is a need to consider dynamic and adaptive fraud detection and analysis approaches; machine learning techniques offer the promise to effectively address some of these challenges.

1.3. Research questions

To achieve the objective of this study, the following research questions are formulated to explore and answer.

1. Which attributes are best to identify and classify subscription fraud calls, and best to build fraud detection model?
2. What machine learning algorithm can be more suitable for the purpose of predicting subscription Fraud?
3. To what extent, the model can identify and predict the fraudulent calls from genuine calls?

1.4. Objective of the Study

1.4.1. General Objective

The general objective of this study is to construct a subscription fraud detection model by employing machine learning techniques so as to detect subscription fraud calls using both postpaid and prepaid subscribers Call Detail Records (CDR) data.

1.4.2. Specific Objectives

To achieve the general objective of this study, the following specific objectives are targeted.

1. To understand the domain area through reviewing various literatures and identify methods used to detect and prevent subscription fraud.
2. To prepare quality dataset with attributes needed for the development and validation of the predictive model.
3. To identify suitable machine learning algorithms for detecting subscription fraud.
4. To design a model which classify fraudulent calls from genuine calls based on call detailed registered data.
5. To evaluate the performance of the proposed subscription fraud detection model.

1.5. Scope and limitation of the study

Frauds are the main challenge in the telecommunication industries. There are various types of fraud activities in the telecom sector such as, billing fraud, bypass fraud/GSM gateway fraud/SIM box fraud, over the Top (OTT) fraud, call forwarding/call diversion fraud, call transfer fraud, CDR manipulation fraud, Dealer fraud, interconnect/low-cost routes fraud, internal fraud, international

revenue sharing fraud and etc. However, this research is focused only on the detection of subscription fraud types using mobile subscribers CDR data from ethio telecom.

In telecommunication, there are different types of data like Network data, CDR data, customer data and Equipment Identity Registration data that can help for analysis and decision. In this research, we used only two-month CDR data for experimentation. The size of the data was more than 10 billion records, and it is sufficient for experimentation. Also, there are a number of techniques and tools which can be used to detect subscription fraud. However, in this work we are limited to use only three types of supervised machine learning algorithms (such as RF, SVM and ANN) and WEKA tool for data analysis.

Subscription fraud can be divided into two categories. These are: Subscription fraud for the purpose of personal usage and Subscription fraud for profit purpose. This study tried to address subscription fraud detection in general, and it has a limitation of identifying subscription fraud in specific purpose.

1.6. Significance of the study

The outcome of this research is helpful for ethio telecom in general to manage, control and protect their customers and revenues from fraudsters. Specifically, Fraud Management and Revenue Assurance departments can use the outcome of the research. The fraud predictive model enables the departments to identify subscription frauds and take appropriate measures. The fraud specialist collects the customer data, CDR data and network data from Customer Relation Management system, Convergent Billing System and, HLR and MSC systems, respectively. The collected data processed, and the detection processes are implemented to generate alarms on situations that deserve closer investigation by fraud analysis specialists.

This study can also motivate other studies to conduct on similar and/or related topics and it may also be used as a reference to related future works.

1.7. Organization of the study

This thesis report is organized into five chapters. Chapter one gives a general overview of the thesis. It starts with a background review of the thesis and describes the telecommunication fraud.

The chapter further describes the problem statement and the research questions. The general and specific objectives of the research are presented, and the scope and limitation of the study are also discussed in this chapter.

The rest of this thesis is organized as follows: In Chapter two, relevant literatures have been reviewed. The chapter is mainly focused on telecommunication fraud in general, subscription fraud, machine learning techniques and review of previously conducted related works. In Chapter three, the research methodology used to conduct this study has been presented. The chapter briefly describes the general research strategy and quantitative experimental research methods. The data analytics techniques, classification algorithms used in this research and the research validity techniques are also concisely presented. Chapter four explains the detail of data preparation. In this section we have discussed data collection, understanding of data, data selection and data preprocessing activities have been discussed. Chapter five focuses on experimentation, analysis, and modeling. The last chapter covered conclusions and recommendations part of the study.

CHAPTER TWO

LITERATURE REVIEW

2.1. Overview

Following to the enormous telecommunication technology growth in the end of the past century, the telecom operators face a new challenge on telecom fraud. It is a continuously changing, multi-facetted creature [10]. Telecommunications fraud are many (PBX/Voice mail systems, subscription/identity (ID) theft, International revenue share fraud (IRSF), credit card fraud, and so on) and new types of fraud evolving every now and then. Telecommunications fraud is a big issue to all telecom operators around the world, and it is an important factor in their annual revenue losses. Issues of telecommunication frauds and their detection have been studied over the years all over the world by the telecom operators and researchers.

Telecommunications fraud continues to be a big problem in the industry today. Advancements in technology have made life easier and more convenient for most people today, but not without a price [11]. These advancements not only bring innovation for good, but they also bring about increasingly sophisticated practices in which fraudsters can breach a company systems and networks [11]. Communication service providers are faced with enough challenges from competition, declining revenue, lower margins and other growth-related challenges. While paying more attention to these other areas, it can leave them vulnerable to unsuspecting attacks [5]. With fraud continuing to be a big problem, fraud management has evolved from a defensive and reactive strategy focused on prevention to a more proactive, revenue generating and innovative approach. Goals have shifted from simply detecting fraud to achieving higher customer satisfaction and creating new revenue streams [12].

Telecommunication fraud is defined as the theft of telecommunication services or the use of telecommunication service to commit other forms of fraud [5]. This type of fraud happens on a daily basis, sometimes without anyone knowing until the damage has already been done.

Fraud primarily occurs to a company with a weak defense system. Billing systems and network vulnerabilities are easily exploited to gain access, when if proper procedures were put in place, could have easily been prevented. With new voice technologies becoming more attractive, improperly installed systems can be penetrated easily and put a small company out of business in a short period of time.

2.2. Telecommunication services and Fraud

2.2.1. Telecommunications Services

Telecommunications, also known as telecom, is the exchange of information over significant distances by electronic means and refers to all types of voice, data and video transmission [11]. It includes a wide range of information transmitting technologies such as telephones (wired and wireless), microwave communications, fiber optics, satellites, radio and television broadcasting, the internet and telegraphs [11].

In many countries like Ethiopia, telecom service providers were primarily government owned and operated, but that is no longer the case, and many have been privatized. The International Telecommunication Union (ITU) is the United Nations agency that administers telecommunications and broadcasting regulations, although most countries also have their own government agencies to set and enforce telecommunications guidelines.

Telecom operators offer various services to satisfy their subscribers because of the market competition. In this regard, ethio telecom provides many services to its subscribers. Mainly the company provides voice and data services to its customers in different technologies. The following are some of the service provided by ethio telecom, which are communicated on the company official web sites [13].

- Hybrid SIM accounts service which enables customers to use postpaid and prepaid accounts with a single SIM Cards.
- Virtual private network (VPN): it is a network service that is constructed using public wires to connect remote users or regional offices.
- Machine to machine (M2M) services: It refers to direct communication between devices using any communications channel, including wired and wireless.

- Fixed Line Service: It is landline telephone connected to the public network by cables. A fixed line is wired to a telephone jack on the wall.
- Mobile internet is a wireless Internet access through a tablet, smart-phone or other mobile device to access the Internet while moving.
- Roaming refers to the ability for a cellular customer to automatically access home services when travelling outside home.
- Very Small Aperture Terminal (VSAT): it is a satellite communications system that serves business users.
- Business Mobile is a bundled postpaid mobile service that allows all postpaid mobile customers to make calls at a discounted rate.
- Evolution Data Optimized (EVDO). It is a service that relies on signal from a wireless tower rather than a physical connection like a phone line.
- Postpaid mobile is a mobile service, which is billed after the fact according to their use of mobile services at the end of each month.
- Mobile Broadband VPN is an extension of VPN to the mobile broadband access. The service is provided through 4G, 3G & 2G technologies.
- Business Internet service is wired and wireless connections with a speed starting from 256 Kbps to the business community.
- Fixed wireless CDMA is a fixed wireless phone, which can work in CDMA network; it provides high quality voice and stable SMS service.
- Asymmetric digital subscriber line (ADSL). It is a data communications technology that enables faster data transmission over copper lines.
- Prepaid mobile service is a mobile service which credit is purchased in advance of service use.

2.2.2. Telecommunications Fraud

The Concise Oxford Dictionary [14] defines fraud as “criminal deception; the use of false representations to gain an unjust advantage.” Fraud is as old as humanity itself and can take an unlimited variety of different forms. However, in recent years, the development of new technologies, which have made it easier for users to communicate and helped increase our spending power; this has also provided yet further ways in which criminals may commit fraud

[15]. Traditional forms of fraudulent behavior such as money laundering have become easier to commit and have been joined by new kinds of fraud such as mobile telecommunications fraud and computer intrusion [15].

In many of the existing literature, the intention of the subscriber plays an important role in the description of fraud. Telecom fraud is defined as the communication of data or voice over a telecommunications infrastructure where the intention of the caller or sender is not to pay or with very discount rate for the service used [11]. Likewise, fraud is defined as obtaining unbillable services and undeserved fees [16]. Fraudster seriously sees themselves as an entrepreneur, admittedly utilizing illegal methods, but motivated and directed by essentially the same issues of cost, marketing, pricing, network design and operations as any legitimate network operator [11]. Fraud is attractive from the fraudsters' point of view, since detection risk is low, no special equipment is needed, and the product in question is easily converted to cash [9]. It is important to state that although the term fraud has a particular meaning in legislation, this established term is used broadly to mean misuse, dishonest intention or improper conduct without implying any legal consequences.

In general, telecommunication fraud is classified in to four groups [12]; these are discussed as follows.

- **Contractual Fraud**- in this fraud group, proceeds is produced through the regular use of a service while having no intention of paying for the service used. Examples of such fraud are Premium Rate fraud and Subscription fraud.
- **Hacking fraud** – with this group, proceeds is produced for the fraudster by breaking into insecure systems and abusing or selling on any obtainable functionality. Examples of such fraud are Network attack and PABX fraud.
- **Technical fraud** - all frauds in this group include attacks in contradiction of weaknesses in the mobile system technology. These kinds of frauds specifically require some initial technical knowledge and skill, even though once a weakness has been exposed this information is often rapidly distributed in a form that unprofessional people can use. Examples of such fraud are Technical Internal fraud and Cloning.

- **Procedural fraud -** all frauds with this group include attacks against the procedures deployed to reduce exposure to fraud, and frequently attack the weaknesses in the business procedures used to provide access to the system. Examples of such fraud are Voucher ID duplication, Faulty vouchers and Roaming fraud.

Fraud is different from revenue leakage. Revenue leakage is characterized by the loss of revenues resulting from operational or technical loopholes where the resulting losses are sometimes recoverable and generally detected through audits or similar procedures. Fraud is characterized with theft by deception, typically characterized by evidence of intent where the resulting losses are often not recoverable and may be detected by analysis of calling patterns. Telecommunications is an attractive target for fraudsters. In terms of volume, it is now measured in the billions worldwide. Recent highly sophisticated schemes are employed by organized crime using hackers and self-learning; Estimated that telecommunications fraud is more attractive than the drug trade [17].

Telecom frauds can be grouped into several different classes, these classes show the method in which telecom companies was defrauded, for example, subscription exploiting wrong behavior. Each method can be exploited to fraud the system for revenue-based goals or non-revenue-based goals. A large percentage of these frauds are performed either by the fraudster copying another person or in fact deceiving the network systems [18].

Telecommunication fraud grouped into two: fraud types and fraud methods. The fraud types contain arbitrage, call and SMS spamming, local income share fraud, international income share fraud, phishing, premium rate service fraud, roaming fraud, and so on. The fraud methods include cramming/slamming, PBX Hacking, SMS Phishing, Subscription Fraud, Wangiri Fraud, and so on [13].

According to Michaux [19], Telecom fraud can also be grouped into three streams, these are technical fraud (boxing, clip-on fraud, payphones, tele card fraud), not-technical fraud (audio text scams, comfort services abuse, cramming, PABX-hacking, slamming, social engineering) and not so technical fraud (calling card fraud, premium rate service fraud, subscription fraud).

Fraud has a harmful effect on everybody, including individual and business customers. The fraud increases the telecom companies' operational costs [19]. In the current situation, new fraud types

and methods are appeared, and there is a need for experts in the area of telecommunications to update their expertise, and also know the trend and losses affected due to these fraud behaviors. It will notify telecom managers to know the urgency of installing fraud management systems whatever of the cost incurred.

2.2.3. Subscription Fraud

Business companies including telecom operators often receive significant losses from customers fraudulent behaviors. Subscription fraud is one of the common types of fraud in which usage type is in paradox with subscription type [15].

Subscription fraud - fraudsters obtain legitimate customer account to get telecom service. But they are not willing to pay for the service they used [11]. In this situation, irregular usage of data or voice calls occurs until the customer account to be disconnected. Cases for uncollectable or unmatched, where customers who do not essentially have fraudulent aims not pay for the bill amount, also categorized into this group. Such types of cases are not always considered as fraud. But it is very interesting and should be known.

Call details alone are not enough to establish cases of fraud. A certain call may be perfectly normal in one situation but indicate fraud in another. For example, a call to an Enterprise customer may be normal if the customer usually makes such calls, but suspicious otherwise. Usage volume, like total numbers of calls, duration or rated value of calls over a certain period are also crucial in establishing fraud cases.

Subscription fraud is mostly preferable by the fraudsters for digital roaming fraud. The modus operandi of subscription fraudster is posing as a credit worthy person or company; the fraudster can gain access to any network, anywhere 2G, 3G or 4G. Typically, the first step for fraudsters is to use subscription fraud to gain access to the home network [17]. Subscription fraud call has the following properties [20].

- Fraudsters make outgoing call frequently.
- Mainly the destination of outgoing call is International.
- After frequently calling, the fraudsters disconnected from the network.

- The incoming calls are much smaller than the outgoing calls.
- Sending frequently international SMS.
- High data usage.
- Longer duration per calls mainly International calls

Subscription fraud occurs when a fraudster uses their own, a stolen or a fake identity to acquire mobile devices and services with the intent of not paying for the service used [20]. In fact, the selling price of mobile devices has increased over time to time, a grey market has been formed and exploited by fraudsters and getting mobile devices to resell for excess returns. The mobile device with less price model is still demanded across many countries, whereby customers have little to no outlay upfront and receive an expensive device that can be resold, lends itself to a high-margin business model for the unscrupulous [5].

The fraud scenario presented in Figure 2.1 indicates that, genuine subscribers agree for a service from telecom service providers and entered into a contract with the service provider and accessing the telecom service provider infrastructure and pay for the service they used. But the fraudsters theft legitimate customer identity, cloning the SIM cards or stolen communication device then accessing the telecom service provider infrastructure with the intention of not to pay the service used but billing will be charged by genuine subscribers.

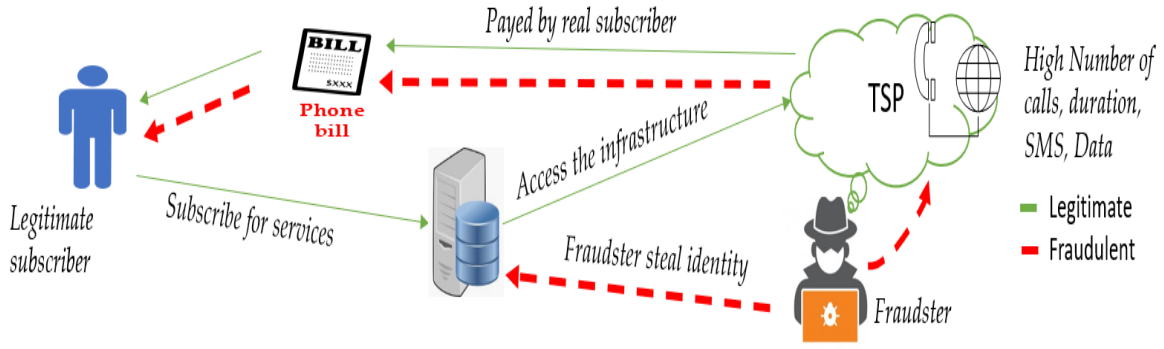


Figure 2. 1 Subscription Fraud Scenario [5]

2.3. Machine learning

Machine learning (ML) is a branch of artificial intelligence that aims at enabling machines to perform their jobs skillfully by using intelligent software [21]. Machine Learning is the study of algorithms that automatically improve their performance, with experience enrich their performance through learning, which is attained by an iterative process [22]. It provides tools by which large quantities of data can be automatically analyzed. Machine learning algorithms have been used to build classification rules from large datasets. Machine learning algorithms require data to learn, the discipline must have connection with the discipline of database. Similarly, there are familiar terms such as Knowledge Discovery from Data (KDD), data mining, pattern recognition and so on.

2.3.1. **Machine Learning techniques**

Machine learning (ML) techniques enable systems to learn from experience and it refers to a system's ability to acquire and integrate knowledge through large-scale observations and to improve and extend itself by learning new knowledge rather than by being programmed with that knowledge [54]. These techniques organize existing knowledge and acquire new knowledge by intelligently recording and reasoning about data. Learning systems have achieved a variety of results, ranging from trivial memorization to the creation of entire new scientific theories and have the potential to continuously self-improve enabling their systems to become increasingly efficient and effective.

ML techniques are used in intelligent tutors to acquire new knowledge about students, identify their skills and learn new teaching approaches. They improve teaching by repeatedly observing how students react and generalize rules about the domain or student. They use past experience to inform present decisions, enable tutors to adapt to new environments, and infer or deduce new knowledge [54].

Based on the desired outcome of the algorithm, Machine learning algorithms are organized into four categories (see figure 2.2). These are: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning [21].

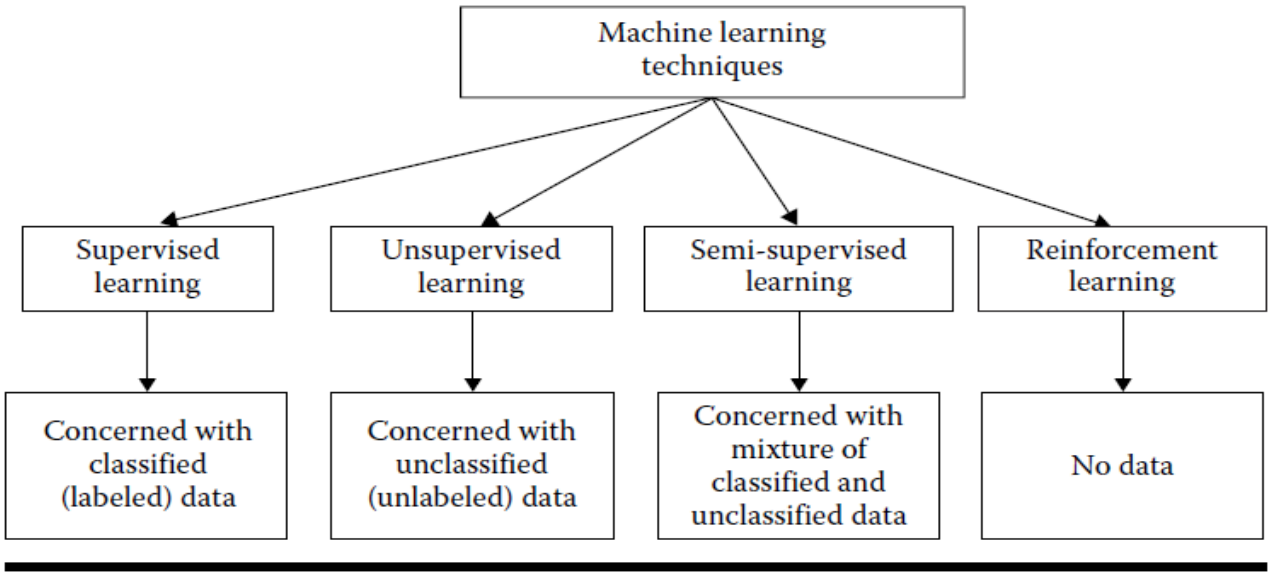


Figure 2. 2 Different machine learning techniques and their required data [21]

In a supervised learning model, the algorithm learns on a labeled dataset, providing an answer key that the algorithm can use to evaluate its accuracy on training data. An unsupervised model, in contrast, it uses unlabeled data that the algorithm tries to make sense of by extracting features and patterns on its own. But it lacks supervisors or training data. The idea is to find a hidden structure in this data [21]. The aim of unsupervised learning is to identify patterns in the data that extend our knowledge and understanding of the world that the data reflects [23]. Even though they are difficult to evaluate, unsupervised models have advantage over supervised models, that new types of fraud may be identified.

Semi-supervised learning takes a middle ground. It uses a small amount of labeled data supporting a larger set of unlabeled data. This combination of labeled and unlabeled data is used to generate an appropriate model for the classification of data. The target of semi-supervised classification is to learn a model that will predict classes of future test data better than that from the model generated by using the labeled data alone [21].

Reinforcement learning is another technique of machine learning. It trains an algorithm with a reward system, providing feedback when an artificial intelligence agent performs the best action in a particular situation.

2.3.2. Supervised algorithms

In supervised learning, the target is to infer a function or mapping from training data that is labeled. The training data consist of input vector X and output vector Y of labels or tags. A label or tag from vector Y is the explanation of its respective input example from input vector X. Together they form a training example. In other words, training data comprises training examples. If the labeling does not exist for input vector X, then X is unlabeled data [21].

Supervised learning models are trained with data that have been pre-classified. The examples of input/output functionality are referred to as the training data. Care needs to be taken in order to ensure that the training data is correctly classified.

The supervised learning methods are categorized based on the structures and objective functions of learning algorithms. Popular categorizations include ANN, SVM, and decision trees [23]. In the case of fraud detection, since legitimate calls occur more often than fraudulent calls, the training data will mostly contain legitimate calls, leading to a misclassification of the model. This needs attention in the supervised learning models. There are different types of supervised learning algorithms. The following are commonly used algorithms which are selected for this work.

- Support vector machine (SVM)

Support vector machines (SVMs) are powerful methods for solving classification problems on large datasets. They combine reliable techniques from linear learning with the intriguing theory of kernel-induced spaces [24]. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N - the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e., the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. So, the best hyperplane is chosen from several possibilities according to some optimization criteria (typically training set performance). Support vector machine algorithms discover the function (hyperplane)

that gives the highest minimum distance to the examples and we call it this distance a margin, and the examples closest to the margins are called support vectors. In the below Figure 2.3, show that all the points put on the margin lines are support vectors, and the distance between these margin lines is width of the margin; because the solution depends only on the support vectors.

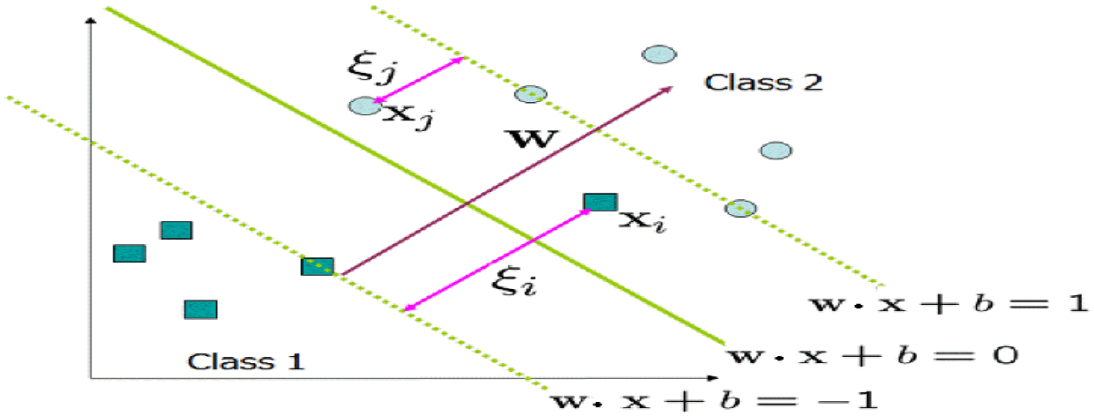


Figure 2. 3 support vector machine classification [25]

As shown in the above Figure 3.2, any hyperplane can be written as the set of points X satisfying wTX + b = 0 , where the vector w is a normal vector perpendicular to the hyperplane and b is the offset of the hyperplane, wTX + b = 0 from the original point along the direction of w. Given labels of data points X for two classes (class 1 and class 2), we present the labels as yi € {1, −1}. Meanwhile, given a pair of (wT, b), we classify data X into class 1 or class 2 according to the sign of the function f (X) = sin (wTX + b). Thus, the linear separability of the data X in these two classes can be expressed in the following equation 2.1 and equation 2.2.

xi.w+ b x +1, f or yi = +1 (2.1)

xi.w+ b w −1, f or yi = −1 (2.2)

The above two equations can be combined and form the following equation 2.3.

yi(xi.w+ b) x 1 (2.3)

In addition, the distance from data point to the separator hyperplane, wT x + b = 0 can be computed as r = (wT x + b)/||w||, and the data points closest to the hyperplane are called support vectors. As

denoted in Figure 3.2, the distance between support vectors is called the margin of the separator, which is simply 2/||w||. Linear SVM is solved by formulating the quadratic optimization problem in the following equation 2.4 [6].

$$\underset{w,b}{Minimize}\ \left(\frac{1}{2}\|w\|^2\right)$$
$$subject\ to\ \ y(w^T x + b) \geq 1 \tag{2.4}$$

In the case of linearly separable data, once the optimum separating hyperplane is found, data points that lie on its margin are support vector points and the solution is represented as a linear combination of only these points, other data points are ignored. Therefore, the model complexity of SVM is unaffected by the number of features encountered in the training data (the number of support vectors selected by the SVM learning algorithm is usually small). For this reason, SVMs are well suited to deal with learning tasks where the number of features is large with respect to the number of training instances.

- Random forest algorithm

Random forests (RF) are a combination of tree predictors that each tree depends on the values of a randomly selected vector samples and it distributes equal values of vector samples to all of the trees in the forest. The strength of individual trees in the forest and the correlation between them determines the generalized error of a forest and its tree [26].

Random forest algorithm builds many decision trees, the number of trees to build is a parameter, which can be selected during the learning phase. Every decision tree is learned with a random subset of features from a sampled training set with replacement. The output is decided by the votes given by all individual trees. Each decision tree is built by classifying the samples of the input data using a tree algorithm. Then, every tree will be used to classify testing data. Each tree has a decision to label any testing data. This label is called a vote. Finally, the forest decides the classification result of the testing data after collecting the votes, and the most popular class is returned [6]. This scenario is illustrated in Figure 2.4 diagrammatically.

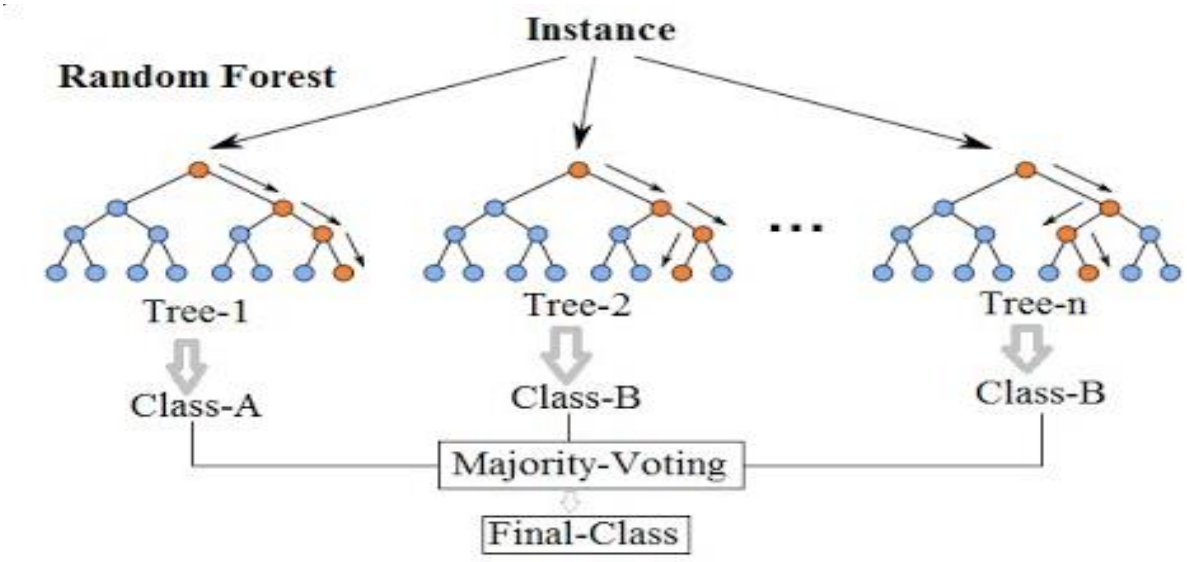


Figure 2. 4 An Ensemble Classifier for random forest [28].

In building a single decision tree, the model builder may select a random subset of the observations available in the training dataset. Also, at each node in the process of building the decision tree, only a small fraction of the available variables is considered. This Significantly reduces the computational requirement. It is also suitable when there are many input variables and little observations, and it can handle large dataset with higher dimensionality. RF algorithms build patterns and detect outliers [6]. The strength of individual trees in the forest and the correlation between them determines the generalization error of the forest and its trees [27].

RF Algorithm performs the following sequence of steps to accomplish its classification [27].

Step:1 Choose T number of trees to grow

Step:2 Choose m number of variables used to split each node. m ~ M, M (input variables)

Step:3 Grow T trees.

When growing each tree do

- Construct a bootstrap sample of size n sampled from Sn with the replacement and grow a tree from this bootstrap sample

- At each node select m random variables and use them to find the best split

Step:4 Grow the tree to a maximal extent and there is no pruning

Step:5 To classify point X collect votes from every tree in the forest and then use majority voting to decide on the class label

- Artificial Neural Network

In its simplest form, an artificial neural network (ANN) is an imitation of the human brain. A natural brain has the ability to learn new things, adapt to new and changing environment. The brain has the most amazing capability to analyze incomplete and unclear, fuzzy information, and make its own judgment out of it [30]. Artificial Neural Network (ANN) is gaining prominence in various applications like pattern recognition, weather prediction, handwriting recognition, face recognition, autopilot, robotics, etc.

There are many varieties of learning algorithms for the design of ANN. They differ from each other in the way in which the adjustment to a synaptic weight of a neuron (node) is formulated.

Learning algorithms can be described as a prescribed set of well-defined rules for the solution of a learning problem. Error-correction, memory-based, competitive, and Boltzmann learning are among the learning algorithms for ANN. ANN learning paradigm is either supervised (associative learning) or unsupervised (self-organizing). In the case of supervised, there is a need to train or teach the input and output pattern. But unsupervised neural network, only the input pattern is needed and from which it develops its own representation of the input stimuli.

Artificial neural network is classified in to three. These are: Feed forward Neural Network, Recurrent Neural Network and Self- Organizing Map [29]. In Feed forward Neural Network, activation is piped through the network from input units to output units. Sometimes they are also referred as static networks. It contains no explicit feedback connections. Conventional Feed

forward Neural Network are able to approximate any finite function as long as there are enough hidden nodes to accomplish this.

Feed forward Neural Network is the first and simplest type of ANN. Recurrent Neural Network on the other hand, are dynamical networks with cyclic path of synaptic connections which serve as the memory elements for handling time-dependent problems. Self-Organizing Map mainly used for cluster analysis.

The big developments in ANN during the past few decades have motivated human ambitions to create intelligent machines with human-like brain. Nowadays, ANN are considered one of the most efficient pattern recognition, regression, and classification tools [29].

When ANN is used as a supervised machine-learning method, efforts are made to determine a set of weights to minimize the classification error. One well-known method that is common to many learning paradigms is the least mean-square convergence. The objective of ANN is to minimize the errors between the ground truth Y and the expected output f (X; W) of the network as $E(x) = (f(X;W) - Y)^2$. The behavior of ANN depends on both the weights and the transfer function, which are specified for the connections between neurons. ANN models implicitly define the relationships between input and output, and, thus, offer solutions for tedious pattern recognition problems, especially when users have no idea what the relationship between variables is.

- **Perceptron**

Perceptron is the simplest kind of ANN, which consists of a single neuron that can receive multiple inputs and produces a single output. Perceptron's are used to classify linearly separable classes. As illustrated in Figure 2.5, a perceptron takes a vector of real-valued inputs, calculates a linear combination of these inputs, then outputs a 1 if the result is greater than some threshold and -1 otherwise using the selected function. The precise learning problem is to determine a weight vector that causes the perceptron to produce the correct output for each of the given training examples. The most common way that the perceptron algorithm is used for learning from a batch of training instances is to run the algorithm repeatedly through the training set until it finds a prediction vector which is correct on all of the training sets. This prediction rule is then used for predicting the labels on the test set.

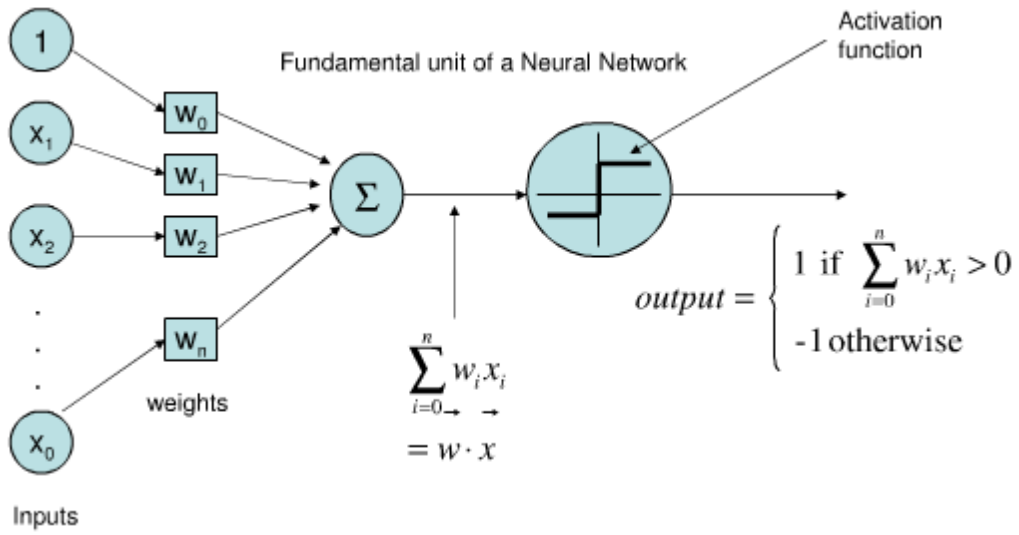


Figure 2. 5 A Perceptron with Multiple Inputs and Single Output [29].

- Multilayer Perceptron (MLP)

A single perceptron can solve any classification problem for linearly separable classes. If given two nonlinearly separable classes, a single layer perceptron network will fail to solve the problem. Such a nonlinearly separable problem is solved by using most popular types of ANN Multilayer Perceptron (MLP) [31]. In an MLP neural network, each perceptron receives a set of inputs from other perceptron's, and according to whether the weighted sum of the inputs is above some threshold value, it either fires or not. The ANN/MLP is ideally composed of three layers, the input layer, the hidden layer, and the output layer as show in Figure 2.6. The input layer consists of input nodes which represent the system's variable. The hidden layer consists of nodes which facilitate the flow of information from the input to the output layers. The flow is controlled by weight factors associated with each connector. The output layer consists of nodes which represent the system's classification decision. The values of the output nodes are compared with Limits to determine the output and classify each case.

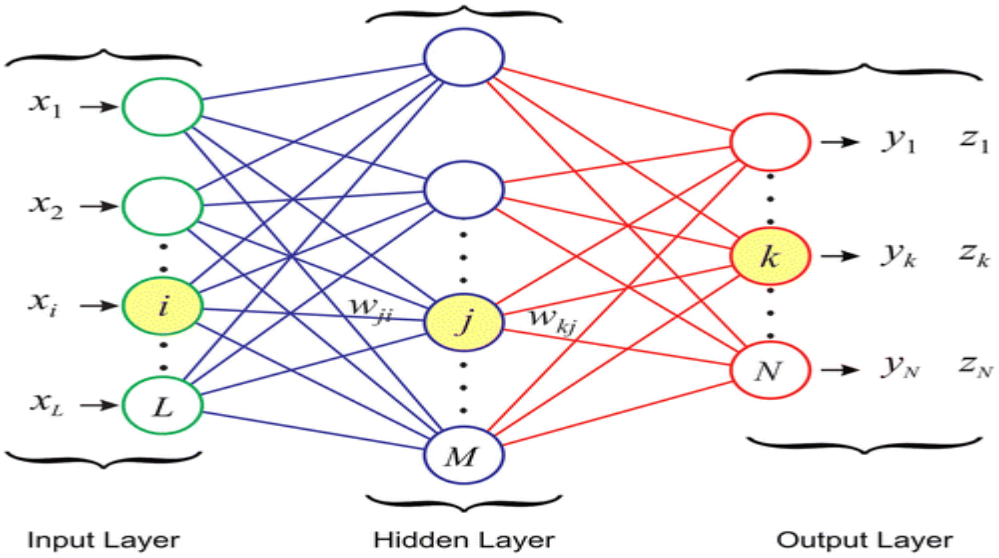


Figure 2. 6 A Multilayer Perceptron Neural Network [31]

2.4. Related works

Recently, telecommunications fraud detection has become a hot research area gradually. A research conducted on exploring the potential applicability of the data mining technology in developing models that can detect and predict pre-paid mobile subscription fraud in ethio telecom service provision [8]. Some researchers have used blacklisting and white listing methods to prevent telecommunications fraud [32].

System was proposed to prevent subscription fraud in fixed telecommunications with high impact on long-distance carriers. This system consists of a classification module and a prediction module. The first module classifies customers based on their past historical behavior into four different groups, which are subscription fraudulent, otherwise fraudulent, insolvent and normal. The second module permits to classify possible fraudulent subscriber at the time of initial subscription. The first module was employed using fuzzy rules. It was functional to a database containing records of more than ten thousand real customers of potential telecom operators in Chile [33].

On another research, a probability-based research technique was used for fraud identification in telecom industries by using Naïve-Bayesian classification to analyze the probability and an adapted version of KL-divergence system to detect the fraudulent subscribers on the basis of subscription [7].

A research conducted on sim-box fraud detection using data mining techniques in the case of ethio telecom. In this work, A models were developed to classify Call Detail Records (CDRs) to propose a model that differentiate fraudulent subscribers from legitimate subscribers with better performance. Classification methods of data mining are applied using J48, PART and multilayer perceptron algorithms on data collected from ethio telecom. Also, WEKA data mining tool had been used to come up with a model for predicting fraudulent activities. On the study, pre-paid sampled voice CDR data had been used along with SMS, GPRS and other data such as pre-paid wallet recharge log from OCS and CCB data warehouse in ethio telecom. The experimentation result of the study shows that the model from the PART algorithm exhibited 100% accuracy level followed by J48 algorithm with 99.98%. The researcher suggested that the rules generated from PART and J48 algorithms enable telecom operators in general and ethio telecom in particular to locate the whereabouts of SIM-boxes as well as other critical information.

Telecommunications Companies are facing a lot of problems due to fraud; hence the need for an effective fraud detection system for the telecommunications companies is unquestionable. A research that presents a design and implements of a subscription fraud detection system using Artificial Neural Networks was conducted. Neuro solution for Excel was used to implement the Artificial Neural Network [36].

More researchers use machine learning techniques to determine the calls are malicious or not. They extract different features for fraudulent call identification, and most of the features contain service number, domain names, call duration, network information, and the actions of sender and receiver, etc. [34][20].

Other research on telecom fraud using the features of a phone communication as the input and Quarter-Sphere SVM to differentiate fraudulent calls had been conducted [24]. The input features include call time, call frequency, call type, location, and time, and it has achieved better recognition accuracy. Later on, the research conducted again using another clustering technique C-means for telecommunication fraud detection and got a good result as well [35].

Other research had been conducted on the construction of predictive model for subscription fraud detection using data mining technique in the case of ethio telecom in the year 2013 [8]. The

research was conducted using 25,000 records only for the analysis and detection of subscription fraud in ethio telecom using data mining technique. The researcher had been used WEKA software for data processing and four different classification techniques, which are J48, PART, Random forest and Multilayer perceptron of artificial neural network. The research was conducted only by using prepaid customs CDR data. In the study, the researcher had been suggested in his future work that the study was limited on prepaid subscription of telecommunication fraud and indicated that other research could be done on postpaid subscription of telecommunication fraud.

Summary of related works

Table 2. 1 Summary of related works

Author & Year	Objective/ Purpose	Approaches/ Methodologies	Key Findings	Recommendation & Future Work
Tesfaye H, (2013)	Constructing predictive model for pre-paid subscription fraud detection using data mining technique at ethio telecom	experimental method using CDR Data and weka Tool, DM Algorithm used include: J48, PART, Random forest, and ANN	Developed a model that can detect and predict pre-paid mobile subscription fraud.	The need of research on postpaid subscription of telecommunication fraud
P. A. Estévez, C. M. Held, and C. A. Perez, (2006)	Identifying telecommunication subscription fraud prevention at the time of new fixed line application for long distance service	experimental method using CDR Data and PostgresSQL1 and weka tools. Algorithm used include fuzzy rules Classification: hierarchical tree structure	The feasibility of significantly preventing subscription fraud in telecommunications.	The research was conducted on wire line technology, but the techniques can be used to subscription fraud in mobile technology
P Saravanan, V Subramaniyaswamy, N Sivaramakrishnan, M Prakash, and T Arunkumar, (2014)	subscription-fraud detection in telecommunication sector using Data mining approach	Naïve-Bayesian classification was used	Identifying patterns for classifying legitimate customer from fraudulent customers.	Further work can be suggested using other probability-distribution based algorithms for a greater accuracy on results.

L. G. Kabari, D. N. Nanwin, and E. U. Nquoh(2016)	Telecommunications subscription fraud detection using naïve bayesian network	Used CDR Data: and Naïve Bayesian Network algorithm	Classifies the different subscription services provided by the telecom industries and detect subscription fraud	The need to develop prediction module to identify potential fraudulent customers at the time of subscription.
H. Farvaresh and M. M. Sepehri, (2011)	Detecting subscription fraud in telecommunication by employing data mining techniques	Hybrid of supervised (using decision tree (C4.5), neural networks, and support vector machines) and Unsupervised; clustering (using and K-means were combined)	Identifying the true fraudulent customer Alone	The effective use of combined data mining technique to identify telecom fraud
L. G. Kabari, D. N. Nanwin, and E. U. Nquoh, (2016)	Telecommunications subscription fraud detection using artificial neural networks,	Used CDR Data and Neuro solutions for Excel tool with ANN Algorithm.	Design and Implementation of subscription fraud detection system based on subscriber's data usage.	LSTM RNN trained with other unsupervised learning objective function applied for call pattern analysis in mobile telecommunication network.
K. Hagos, (2018)	SIM-Box Fraud Detection Model	Used CDR Data and weka tool with	proposed a model that differentiate fraudulent from	The need to conduct study on other telecom fraud types

	Using Data Mining Techniques:	Classification algorithms (SVM, RF NN)	legitimate subscribers with better performance.	

2.5. Research Gap

Based on the review of relevant literatures on telecom fraud detections and machine learning techniques used for the detection of various types of telecommunication fraud, we have understood some important gaps in the literatures.

The first gap in the reviewed literature was inadequate number of research related to subscription fraud detection in the case of ethio telecom. Even though there have been several studies which have examined in telecom fraud detection in different contexts (as briefly summarized and presented in Table 2.1), there is limited number of research, offering a complete and exhaustive investigation and analysis for the detection of subscription fraud.

Secondly, this review of the literature indicates that even if different studies have been examined in telecom fraud detection, telecommunication fraud is increasing dramatically time to time and resulting in loss of huge amount of money all over the world.

The other gap which has been observed in the detection of subscription fraud is the accuracy of classifier models to predict fraudulent calls. An important tool for the detection of fraud is the modeling of telecom subscribers' behaviors. Hence, the modeling and characterization of subscriber's behavior in telecommunications can be used to improve the detection of fraud.

Moreover, the traditional time-based and rule-based approaches of analyzing CDR data are quickly becoming outdated. It must be improved by intelligent technique like machine learning methods to have efficient fraud detection solution.

CHAPTER THREE

Research Methodology

3.1. Overview

Research Methodology is a way that deals with data collection, analysis and interpretation that shows how researcher achieves the objectives and answers the research questions [38]. Hence, in order to achieve the general and specific objectives of this study, a quantitative laboratory experimental research method is used to conduct this research. This method is selected based on previously conducted research recommendations on telecom fraud detection [2,4,6].

3.2. Quantitative Research Method

Different researchers and scholars give various definitions to quantitative research methods. Quantitative research method is the numerical representation and manipulation of observations for the purpose of describing and explaining the phenomena that those observations reflect. It is used in a wide variety of natural and social sciences [40].

In addition, other researcher concisely defined that quantitative research as a type of research that is explaining phenomena by collecting numerical data that are analyzed using mathematically based methods (in particular statistics) [39].

In short, quantitative research generally focuses on measuring social reality. Quantitative research and/or questions are searching for quantities in something and to establish research numerically [37]. Quantitative researchers view the world as reality that can be objectively determined so rigid guides in the process of data collection and analysis are very important.

There are several types of quantitative research methods and they are mainly classified in to four categories, such as: 1) survey research, 2) correlational research, 3) causal-comparative research and 4) experimental research. Each type has its own typical characteristics [37].

- Survey Research is defined as the process of conducting research using surveys that researchers send to survey respondents. The data collected from surveys is then statistically analyzed to draw meaningful research conclusions.

- Correlational research is a type of non-experimental research method in which a researcher measures two variables, understands, and assesses the statistical relationship between them with no influence from any extraneous variable.
- Causal-comparative research is an attempt to identify a causative relationship between an independent variable and a dependent variable. The relationship between the independent variable and dependent variable is usually a suggested relationship (not proven) because the researcher does not have complete control over the independent variable.
- Experimental research is a scientific approach to research, where one or more independent variables are manipulated and applied to one or more dependent variables to measure their effect on the latter. The effect of the independent variables on the dependent variables is usually observed and recorded over some time, to aid researchers in drawing a reasonable conclusion regarding the relationship between these two variable types.

This research is used a quantitative experimental research method to answer the research questions and to address the general and specific objectives of the study. Experimental research is one of the founding quantitative research methods. It is any research conducted with a scientific approach, where a set of variables are kept constant while the other set of variables are being measured as the subject of experiment [41]. The experimental research is used to prove or disprove the statement. In this research, it is used to classify ethio telecom subscribers CDR data/records either categorized as a subscription fraud or genuine.

3.3. Research Model

For the purpose of conducting this research, Experimental process model is defined based on the extensive literature review [5, 13, 22]. The model mainly adapted from data analytics life cycle with few modifications [42]. As presented in figure 3.1, the process model illustrates data collection (Raw CDR Data), data preprocessing, algorithm learning, evaluation of the algorithm and prediction.

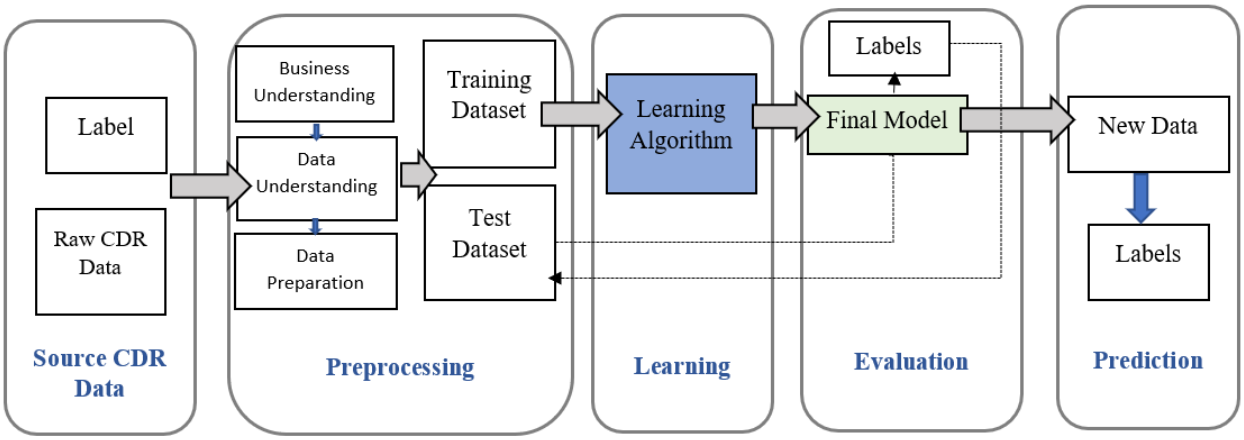


Figure 3. 1 The overall research design & methodology [42]

3.4. Data set and Sampling

3.4.1. Data set

Telecom companies might have millions of subscribers and they have different databases to maintain information on these subscribers and detail of call transaction records. Whenever a call is placed on telecom network, detail information about the call is stored as call detail record data. The dataset used for this research is taken from ethio telecom. The data collected from different databases of ethio telecom information system division, like CBS, CRM and FMS systems. The CRM system database contain customer information like, subscriber name, address, service plan, contract information, credit score and payment history. The CBS and FMS databases contain information about detail call records for the purpose of billing.

The CDR data has been collected from ethio telecom active mobile subscribers, which includes both the prepaid and postpaid customers. Prepaid services are almost common in all services delivered by telecom operators. In prepaid service, all business in this service is acquired whenever the subscriber made a payment for the service. Service provider can easily maintain a fraud made by prepaid customers and it is less vulnerable to fraud as compared to postpaid services. In postpaid services, the customer will receive a bill at the end of the month based on the usage and settle the bill amount accordingly. It is the most conventional service provided by telecommunication companies in worldwide. As the name indicates, postpaid service is a credit facility, which is given for services used for specific period of time, commonly from one to six months.

3.4.2. Sampling technique and sample size

Sampling is the process of selecting a portion of the population to represent the entire population. Telecommunication companies generate a huge volume of data every second and the types of the data are bulky. It is very important to take sample of the data so that the data size could be more manageable. There are different types of sampling techniques. In this study, random sampling technique is used to select records for the experimental analysis in order to get proportional representative sample. Proper sampling is important in machine learning. Specially for fraud detection problems, there are many more legitimate than fraudulent samples. Hence, appropriate classification approaches are needed to classify the unbalanced data. Decisions about how large of a sample to use must be made rationally. Reviewing literatures on telecom fraud detection [1,5,8,14] and by consulting ethio telecom domain experts, A researcher decided and incorporated a proportionality of 60% legitimate subscribers CDR data and the remaining 40% fraudulent CDR data in the dataset; which means, 16,000 fraudulent numbers and 24,000 legitimate customer numbers. Totally we used 40,000 subscriber numbers for this thesis. The legitimate subscribers' numbers are selected using random sampling technique from about 35 million active target mobile subscriber numbers.

3.5. Data Preprocessing

Data preprocessing is a critical task for data analysis and to build a model. Currently, huge volume of data is generated and stored frequently. Organizations are collecting and storing massive volume of data because of using fast and less expensive computers. New analytics technique and tools are required when organization databases size is increased due to accessibility of powerful and affordable database systems [43].

The raw CDR data are not used directly for data analysis, since the goal of this research is to build a predictive model to identify subscription fraud. The call detail records associated with subscribers must be summarized into a single record that describes subscriber calling behavior. All relevant data for each subscriber was constructed in the form of columns and row dataset and it represents a single subscriber with all the data associated to this subscriber includes the choice of features is critical in order to obtain a useful description of the subscriber.

3.6. Classification methods for Learning

This part of the study is concerned about a technique used for analyzing preprocessed data in order to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption. Data analytics techniques can reveal trends and metrics that would otherwise be lost in the mass of information. This information can then be used to optimize processes to increase the overall efficiency of a business or system.

Data analytics and machine learning are growing at fast rate and companies are now looking for professionals who can sift through the goldmine of data and help them drive swift business decisions efficiently. There are different types of data analytics techniques used for machine learning. The most used methods are Regression, Clustering, Visualization, Decision Trees/Rules and Random Forests [13].

In this study, research questions that can be answered using classification or prediction tasks like which call is fraudulent, which one is legitimate call or what kind of behavior fraudulent calls have.

Classification techniques and its algorithms are respectively used in this research, to detect subscription fraud by identifying the behavior of the legitimate and fraudulent customer numbers. As we have mentioned in the literature review part, in selection of algorithms, three classification algorithms had been selected and used in this research. The algorithms are selected for reason explained here. Firstly, the SVM algorithm is chosen for its simplicity and easy interpretability of the result generated by it [24]. Secondly, Random Forest (RF) is an ensemble classifier that contains numerous decision trees and outputs the class. That is, the mode of the class's output by individual trees and because of the advantage of no need for pruning trees, Accuracy and variable importance generated automatically, Overfitting is not a problem, not very sensitive to outliers in training data and easy to set parameters [41]. Finally, we applied artificial Neural Network techniques because of many researchers used this algorithm in the areas of telecommunication fraud detection and so that the final algorithm is the Multilayer perceptron of artificial Neural Network techniques. One of the basic targets of data analytics is to compare different models and to select the better classification accuracy accordingly.

The study is conducted using WEKA software as a tool for data classification. It is easy to use for novice user due to the graphical user interfaces it contains. It is very portable because it is fully implemented in the Java programming language and thus runs on almost any computing platform. The tool contains a comprehensive collection of data preprocessing and modeling techniques and it is freely available under the General Public License (GNU).

3.7. Evaluation Methods

The correct use of model evaluation, model selection, and algorithm selection techniques is vital in academic machine learning research as well as in many industrial settings [44]. Model evaluation aims at estimating the generalization error of the selected model, i.e., how well the selected model performs on unseen data. Obviously, a good machine learning model is a model that not only performs well on data seen during training (else a machine learning model could simply memorize the training data), but also on unseen data. Hence, before shipping a model to production we should be fairly certain that the model's performance will not degrade when it is confronted with new data [44]. The model needs to be evaluated in order to demonstrate its quality and efficiency. This helps to improve it in an iterative manner to ensure the quality of the proposed solution so that it can solve the real problem identified initially.

The model evaluation can be conducted in an objective and subjective methods [39]. The objective evaluation can be conducted using different validation strategies like holdout method, random subsampling, Cross-validation, Bootstrapping and so on, which are commonly used in machine learning model and algorithm evaluation.

In this study various classification models are developed and evaluated by applying training and testing dataset. The experimental output of the classification models is analyzed and evaluated the performances accuracy using confusion matrix. which is a common and recommended method of choice for the evaluation of different algorithm output for small to moderately sized datasets.

- **Confusion matrix**

The confusion matrix is used to measure the performance of two class problems for a given dataset. True positive (TP) and True negative (TN) means that correctly classify instances as well as false positive (FP) and False negative (FN) means incorrectly classify instances. Figure 3.6 shows the

confusion matrix (correctly classified instances (TP, TN) and incorrectly classified instances (FP, FN).

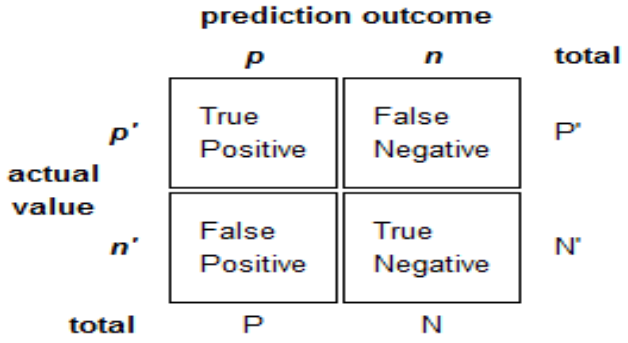


Figure 3. 2 Confusion matrix performance measure [7]

Total number of Instance = correctly classified instances + incorrectly classified instances.

- **Cost matrix**

A cost matrix is similar to confusion matrix, but minor difference is with finding the value of cost accuracy through misclassification error rate.

Misclassification error rate =1- accuracy

- **Recall**

Recall is the ratio of modules correctly classified as fault prone to the number of entire fault modules.

$$Recall = \frac{TP}{TP + FN}$$

Precision

Precision is the ratio of modules correctly classified to the number of entire modules classified fault prone. It is proportion of units correctly predicted as faulty.

$$Precision = \frac{TP}{TP + FP}$$

- **F-Measure**

It is a combination of recall and precision. It is defined as harmonic mean of precision and recall

$$F - Measure = \frac{2 * Recall * precision}{Recall + precision}$$

- **Accuracy**

It is defined as the ratio of correctly classified instances to total number of instances in the dataset.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Besides on the objective evaluation of the model, subjective evaluation is appropriate and crucial in order to assess the output of the study. The result of this research is also subjectively evaluated by discussing with the domain experts of ethio telecom fraud management section, revenue assurance and internal IT technical audit team and others at information systems division experts. This step contains understanding the outcomes, checking whether the result is novel and interesting, checking the effectiveness and efficiency of the model, measuring the degree of satisfaction on the result of this research and checking the impact of the discovered knowledge.

There are different techniques to conduct subjective evaluation. In this study, expert validation method is used to evaluate the proposed model. Accordingly, discussion and survey questionnaires are the two ways used to gain expert validation.

The evaluation was helpful to gain the different views and valuable inputs of the domain experts who work in ethio telecom in various positions for many years. Moreover, most of the experts have good experience in the case of telecom fraud management which adds some value to their multi directional view of the proposed subscription fraud detection model. The domain experts experience in the company was also helping to evaluate the model whether it fits to ethio telecom fraud management process or not.

CHAPTER FOUR

Business understanding and Data preparation.

4.1. Overview

One of the primary purposes of data preparation is to ensure that information being ready for analysis should accurate and consistent, so the results of the analytics will be valid. Data is often created with missing values, inaccuracies, or other errors. Additionally, data sets stored in separate files or databases often have different formats that need to be reconciled. The process of correcting inaccuracies, performing verification, and joining data sets creates a big part of the data preparation process. This chapter focuses data preparation process starting from business understanding, initial data collection, data understanding, data selection and data preparation.

4.2. Business Understanding

The goal of business understanding is identifying the key variables that are to serve as the model targets and whose related metrics are used to determine the success of the project and to identify the relevant data sources that the business has access to or needs to obtain.

Business understanding is an initial phase used to understand the objectives of data analysis project and the converting of these requirements from the perspective of the subject area, and the problem formulated into a definition of data analysis problem. Here, the main objective of this phase is a thorough understanding of what the customer really needs to accomplish. The purpose of this phase is to discover important factors that influence the success of the project [45].

Telecom operators suffer a lot by fraudsters who use telecom services without paying. The estimated losses amount to some billions of dollars in uncollectible bad debts per day. Even though this is a small percentage comparing to the telecom operator's revenue, it is still a significant loss. The mobile telecommunication industry stores and generates tremendous amounts of raw and heterogeneous data that provides rich fields for analysis. To understand the telecommunication, the researcher observed ethio telecom product sales and marketing and communication departments. As initial step the researcher involves working closely with domain experts to define the problem and determine the research goals. Domain experts were consulted to have brief understanding on the problem area.

Ethio telecom is transforming its infrastructure and services to world class standard. Thus, the company is born from this ambition in order to bring about a paradigm shift in the development of the telecom sector to support the continuous growth of the country.

In line with its ambitious mission, ethio telecom has impressive goals of being a customer centric company, offering the best quality of services, meeting world-class standards and building a financially sound company.

4.3. Data collection

Telecom data is growing at a rapid rate, all because of the deep penetration of mobile phones in our life. In telecommunication field there are some critical data on which any decision can be made. These are mobile phone usage, mobile location, server logs, call detail records, network equipment, social networks, and various others. Most of the time CDR data includes enough information describing the call. CDR is generated in real-time every time when a call is placed on the telecommunication network. The size of CDRs that are generated and stored is very large. The CDR should be kept online for several months and billions of CDRs stored. In the case of ethio telecom, CDR details stored for six months in the live systems and then it moves to data warehouse system.

After the approval of data access request from ethio telecom, CDR data collection started. Due to the bulkiness of the record and resource limitation, storing all CDR for an extended period is challenging. Moreover, storing these records in a separate machine simplifies the data preparation process and guaranties the safety of the business operations. Taking this into consideration, two-month CDR data stored in a dedicated server allocated for this thesis work.

The two-month data collection is decided upon the storage and computational resource capacity limitation of the server, as well as sufficiency of the data for the intended experiment. CDR dump files in a text file format, screenshot of the file illustrated in figure 4.1 are stored into the server. These revived files have been imported to a database prepared in advance in this server via automated data loading tool.

In addition, a sample of 16,000 fraudulent subscriber service numbers was collected from fraud management section. These fraudulent numbers were proven and blocked their access to the network due to their fraudulent activities. These sample fraudulent subscriber numbers have been imported into the database as well.

Figure 4. 1 Screenshot of Raw CDR Dump File

The CDR data contains each call information related to telephone call, such as billing number, Time of call initiation, duration of call in second, mobile number initiating call, mobile number receiving the call, online charge system, recharge ID number, call type (local or international), the amount charged or to be charged for the call duration, subscriber ID number used for billing and subscriber line status are some of the details of the call. In the application of machine learning technology and in developing a model that can support subscription fraud detection, the goal of this research is to developing model so as discover the presence of illegitimate calling activity of telecommunications customers. The illegitimate calling activity may not be able to be observed directly, but they are reflected in the calling behavior. The calling behavior is collectively described by the call detail record, which in turn can be observed. Therefore, it is reasonable to use call detail record to apply machine learning technology and to formulate model using training and testing dataset and evaluate the accuracy of the model.

4.4. Understanding the data

After the collection of the CDR data, the researcher discussed with domain experts and understood the content of each attributes in relation with the problem domain. Furthermore, confirming the importance of the data, redundancy, missing values, completeness, and rationality of the attributes.

As shown in Table 4.1, the collected CDR data, contains a total of 33 fields. Some fields are without values like Calling IMEI; others contain duplicate values like Billing Number and Calling number.

Most of them are generated for billing purpose like CHARGE, Call Fee, Account Item ID, Rate ID, Billing Date, Billing Offering ID and Billing Cycle ID. Upload traffic and Download traffic contains the Internet usage. Each CDR data is uniquely identified by the CDR ID sequence number. SMS, Voice, and internet data usage are identified by RE ID number. The CDR_TYPE is used to identifying mobile call initiating, ending, or transferring call types. Some attributes like caller, receiver and bill charging numbers are excluded from the attribute list for confidentiality reasons.

NO	Field Name	Data type	Description
1	CDR_ID	Numeric	CDR Sequence Number
2	RE_ID	Numeric	CDR type ID for voice, SMS and Data
3	BILLING_NBR	Numeric	Billing Number
4	CDR_TYPE	Numeric	Call type ID
5	CALLING_NBR	Numeric	Mobile number initiating or originating the call. It is the same as billing number.
6	CALLED_NBR	Numeric	Mobile number receiving the call.
7	CALLING_IMSI	Numeric	It is calling for International Mobile subscriber Identification (IMSI)
8	CALLING_IMEI	Numeric	It is calling for International Mobile Equipment Identification (IMEI)
9	START_TIME	Date	Time of call initiation (calling time).
10	THIRD_NBR	Numeric	It is forwarded call for third party number
11	DURATION	Numeric	Call duration in seconds
12	END_TIME	Date	Time of call terminated (ending time).
13	CALL_FEE	Numeric	Amount paid in cents (same as CHARGE).
14	CALLED_COUNTRY	String	It is the name of called country
15	CALLING_CARRIER	Numeric	It is the calling carrier number

16	CALLED_CARRIER	Numeric	IT is the called carrier number
17	CELL_A	String	Mobile BTS cell sector A number (BTS-ID) where the call is originated
18	CELL_B	String	Mobile BTS cell sector B number (BTS-ID) where the call is destined.
19	STATE_DATE	Date	It is the billing date
20	CALLING_SUB_ID	Numeric	It is the subscriber ID of the caller
21	BILLING_CYCLE_ID	Numeric	It is bill preparation time.
22	CHARGE1	Numeric	It is the bill amount that the customer charged
23	CHARGE2	Numeric	It is the bill amount that the customer charged
24	PRICE_ID1	Numeric	Rate ID
25	ACCT_ITEM_ID1	Numeric	Account item ID
26	TRAFFIC_UP	Numeric	Upload traffic
27	TRAFFIC_DOWN	Numeric	Download traffic
28	BILLING_OFFERING_ID	Numeric	Billing offering ID
29	ERROR_CDR_TYPE	Numeric	Error CDR Indicator
30	CALL_FORWARD_INDICATOR	Numeric	Call Forward Indicator
31	CALLED_TRUNK_ID	Numeric	It is called number trunk ID
32	CALLING_TRUNK_ID	Numeric	It is calling number trunk ID
33	HOT_LINE_INDICATOR	Numeric	It is hot line indicator for voice mail service

Table 4. 1 CDR Fields Description

4.5. Data selection

Data selection is a process, which requires domain knowledge to choose useful features that capture the variability and essentiality of the data for the target machine learning algorithm to learn patterns from the data successfully. In addition, it has a vital role in reducing complexity of learning process and increase fraud detection effectiveness. The behavior of subscription fraud discussed in Section 2.2.3 is an input for attribute selection. In this research, we are not used all attributes from the collected CDR data.

4.5.1. Attribute selection

Attribute selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model. Some predictive modeling problems have a large number of variables that can slow the

development and training of models and require a large amount of system memory. Additionally, the performance of some models can degrade when including input variables that are not relevant to the target variable.

Most machine learning algorithms are designed to learn appropriate features to use for making their decisions. Adding distracting features often confuses machine learning systems. In practical situations there are many attributes for learning process, some of them feasibly significant, and some are irrelevant or redundant. The problem is identifying a representative set of features from which to construct a classification model. For that reason, the dataset must be preprocessed to select useful attributes. Even though, many learning schemes can select features appropriately and ignore irrelevant ones, but in practice their performance might be affected. Because of the negative effect of irrelevant attributes on most ML algorithms, it is common to precede learning with an attribute selection.

There are different feature selection criteria. Filter method makes an independent assessment based on general characteristics of the data; attributes filtered to produce the most promising subset. Wrapper method is to evaluate feature subset using the machine learning algorithm that will be employed for learning. Making an independent assessment of an attribute subset would be easy if there were a good way of determining when an attribute was relevant.

In this thesis work we applied Correlation based Feature Selection (CFS) method is used to select relevant attributes. It evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. The central idea here is that, subsets of features which are highly correlated with the class while having low inter-correlation are preferred.

However, this method has limitation of selecting features that have locally predictive values when they are overshadowed by strong, globally predictive features. While a single feature may account for only a very small proportion of a dataset, a number of such features may cumulatively cover a significant proportion of the dataset. Having this into consideration, three subsets of features are proposed. In this preprocessing stage we have tested the three proposed feature subsets in small sample datasets. Out of which the subset contains all the proposed features performed better in all

the sample datasets. Then the subsets containing the entire features were taken for this experiment. The designed subsets of futures are presented in Table 4.2. Some of the features are not performed in all subset and they are presented in the table as "X". These attributes are dropped and not included in the selected features for experimentation.

Table 4. 2 Subsets of Selected Attributes

Subset 1	Subset 2	Subset 3
TOT_CALLS	TOT_CALLS	TOT_CALLS
DIST_CALLS	DIST_CALLS	DIST_CALLS
CALLED_COUNTRY	X	X
OUT_CALL_DURATION	OUT_CALL_DURATION	OUT_CALL_DURATION
CELL OUT	X	X
TOT_CALL_FEE	TOT_CALL_FEE	TOT_CALL_FEE
INT_OUT_CALL	INT_OUT_CALL	INT_OUT_CALL
TOTAL OUT CALL	X	TOTAL OUT CALL
INC_CALL	INC_CALL	INC_CALL
TOTAL_SMS	TOTAL_SMS	TOTAL_SMS
CALL_FORWARD_INDICATOR	X	X
DATA_USAGE	DATA_USAGE	DATA_USAGE
BILLING_CYCLE_ID	X	X

4.6. Data preprocessing

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. It is conducted because of the real-world dataset tends to be incomplete, noisy and inconsistent, thus the data preprocessing stage is one of the major tasks in the knowledge discovery process. In this study, in order to improve efficiency, reduce computational time and ease of the data mining process, we have performed the following major data preprocessing tasks, which are, data cleaning, integration, aggregation and finally the data formatting process has been conducted.

4.6.1. **Data cleaning**

Data cleaning is to fill the vacancy value of the data, eliminate the noise data and correct inconsistencies in the data. It is time-consuming and labor-intensive procedure, but one that is

absolutely necessary for successful data analysis. It needs domain experts support and understanding the data in depth. It is used for making sure that the data is free from different errors.

Applying tools that show the distribution of values of fields are very helpful in cleaning data. These tools simplify identifying missing values, outliers, and errors in the data. When collecting data, it is not possible to ensure it is perfectly collected, there will be errors in the data, we need to address data quality issues. The collected data for this thesis, as any big data collection, have some errors, such as incomplete values, missing value, duplicate records. In cleaning this data, the process started by selecting only calls generated by mobile devices, records of other than mobile subscriber numbers are removed. Records containing any missing value in any of their fields similarly excluded from the target data. Records with incomplete or invalid values (such as Calling Number length different from 12 characters) are removed. Records missing country code (251) prefix are modified by concatenating a prefix to those records. Duplicate records are removed by keeping only a single instance of the duplicates. In addition, quality and validity of the target data is checked in accordance with the intended machine learning techniques.

4.6.2. Data Integration

Data integration is the process of combining data from different sources into a single dataset to have a unified view. Before beginning work on machine learning process, it is necessary to bring all the data together into an instance. Integrating data from different sources usually presents many challenges.

However, voice outgoing call, SMS, Internet data usage, voice incoming calls and subscription age records of the selected fraudulent and non-fraudulent subscriber samples are stored in different tables. So, to bring all these records into a single instance record, it should be integrated into a single table based on the unique value of the records.

4.6.3. Data Aggregation

It is the process of gathering data and presenting it in a summarized format. The data may be gathered from multiple data sources with the intent of combining these data sources into a summary for data analysis. This is a crucial step since the accuracy of insights from data analysis

depends mainly on the amount and quality of data used. It is important to gather high-quality accurate data and a large enough amount to create relevant results. We used selected relevant attributes for aggregation based on customers' service usage patterns of call incidence, call fee, call duration, usage data and total incoming call. We used two derived attributes (RATIO_DISTCALL_TOTAL and RATIO_INT_TOT) which are new to detect subscription fraud. The RATIO_DISTCALL_TOTAL attribute was derived from distinct call over the total call and the RATIO_INT_TOT attribute was derived from international call over total calls. A total of ten attributes are used for experimentation. This collected data property of a customer gives a better fraud detection ability. Table 4.2 shows the aggregated result of voice call, SMS and internet data usage in a sing customer level.

Table 4. 3 Description of aggregated and derived attributes

Attribute	Description
TOT_CALLS	It is the number of total calls conducted by customers.
DIST_CALLS	It is the number of uniquely called numbers.
OUT_CALL_DURATION	It is the total time taken to conduct outgoing call.
RATIO_DISTCALL_TOTAL	It is the ration of distinct calls over total calls.
TOT_CALL_FEE	It is the total amount of call fee.
INT_OUT_CALL	It is total international calls.
RATIO_INT_TOT	It is the ration of international calls over total calls.
INC_CALL	It is the number of received calls.
TOTAL_SMS	It is the total number of SMS sent.
DATA_USAGE	It is total data usage.
FRAUD_STATUS	Identifying fraudulent or non-fraudulent call.

4.6.4. Data Formatting

Data format plays a key role in understanding of data, representation of data, space required to store data, data I/O during processing of data, intermediate results of processing, in-memory analysis of data and overall time required to process data [46]. Different data mining and machine learning algorithms require input data in specific types and formats. In this study we used Weka for data analysis. Weka prefers to load data in the ARFF or CSV format, and we have converted our data into CSV formats. Figure 4.2 shows that screenshot of sample data in CSV format.

	A	B	C	D	E	F	G	H	I	J	K
1	Total_Call	DIST_CALL	OUT_CALL	RATIO_DI	TOT_CALL	TOT_SMS	USAGE_DA	INT_OUT	InC_Call	Ratio_INT	F_STATUS
2	47	13	20.17	0.28	5.662	1	3.5	28	7	0.6	Y
3	189	1	1.33	0.01	0.667	1	25	187	0	0.99	Y
4	85	11	95	0.13	27.36	1	8	65	5	0.76	Y
5	122	10	34.67	0.08	12.47	7	14.5	95	6	0.78	Y
6	246	20	117	0.08	48.4365	9	19.5	197	0	0.8	Y
7	1016	5	6	0	3.0017	22	134.5	1008	0	0.99	Y
8	156	16	37.33	0.1	16.24	2	20.5	128	0	0.82	Y
9	78	12	133.67	0.15	47.27	1	5.5	45	25	0.58	Y
10	41	3	10	0.07	4.5023	7	3.5	33	1	0.8	Y
11	172	3	6.83	0.02	3.3432	12	22	165	0	0.96	Y
12	48	9	20	0.19	10.0053	7	3.5	36	0	0.75	Y
13	94	10	60.33	0.11	25.6801	1	10.5	76	9	0.81	Y
14	12	9	77.67	0.75	38.853	0	0	0	0	0	N
15	9	6	51	0.67	18.8361	0	0	0	1	0	N
16	36	18	83	0.5	33.06	1	1	7	7	0.19	N
17	4	4	30	1	15.0076	0	0	0	0	0	N
18	7	3	37	0.43	11.8864	0	0	0	3	0	N
19	12	7	73	0.58	25.084	0	0	0	6	0	N
20	3	2	1.33	0.67	0.6671	0	0	0	0	0	N
21	41	20	88	0.49	41.6662	1	1	7	2	0.17	N

Figure 4. 2 Screenshot of sample data in CSV format.

CHAPTER FIVE

Experimentation and Modeling

5.1. Overview

In this study, the experimentation part is performed by using WEKA data mining tool to identify subscription fraudulent numbers and their behavior. Different experiments are made by using ten attributes. The experimentation is performed using three machine learning algorithms which have been discussed in detailed in section 2.5. They are, Random forest, Artificial neural network (multilayer perceptron) and Support vector machine (SVM) algorithms.

The dataset is labeled as fraudulent subscribers CDR data as (Y) and non-fraudulent subscribers CDR data as (N) and it contains 172, 822 (39.68%) and 262,753 (60.32%) records, respectively. Training data is used for building classification model and testing data is used for measuring the performance of the model.

In machine learning, three types of data sets are used for training, validation, and test purpose [46]. Training and validation dataset used for model building. These datasets are not the same, they are mutually independent and created by random sampling. Test dataset is used for model evaluate. Figure 5.1 shows the split of dataset into Train, Validation and Test sets.

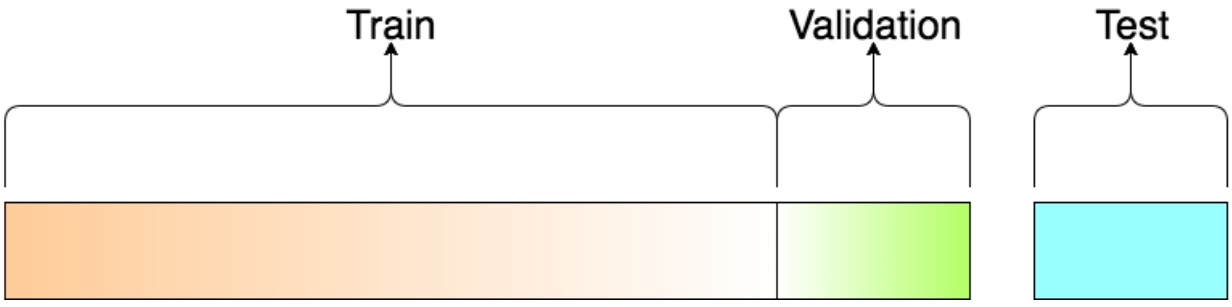


Figure 5. 1 A visualization of the data set splits

In this study, all experiments have been used 10-fold cross-validation and percentage split training and testing mode because it reduces the variance of estimate. In addition to this, to ensure a balanced level of algorithms comparison, recommended default parameters have been used. In order to show the effect of training dataset size on the output performance, this study has taken two training and testing options. The 10-fold cross validation testing method is used. This method

works first the initial dataset are arbitrarily divided into ten equally exclusive subgroups or folds, 1,2,3,4,5 .....10. The training and testing process have been conducted ten times. In the first iteration, the first subset is kept for testing purpose, and the other nine subsets are collectively used to train the classifier.

Train classifier on folds: 2 3 4 5 6 7 8 9 10; Test against fold: 1
Train classifier on folds: 1 3 4 5 6 7 8 9 10; Test against fold: 2
Train classifier on folds: 1 2 4 5 6 7 8 9 10; Test against fold: 3
Train classifier on folds: 1 2 3 5 6 7 8 9 10; Test against fold: 4
Train classifier on folds: 1 2 3 4 6 7 8 9 10; Test against fold: 5
Train classifier on folds: 1 2 3 4 5 7 8 9 10; Test against fold: 6
Train classifier on folds: 1 2 3 4 5 6 8 9 10; Test against fold: 7
Train classifier on folds: 1 2 3 4 5 6 7 9 10; Test against fold: 8
Train classifier on folds: 1 2 3 4 5 6 7 8 10; Test against fold: 9
Train classifier on folds: 1 2 3 4 5 6 7 8 9; Test against fold: 10

The classifier of the second iteration is trained on folds 1, 3, 4... 10 and tested on the 2nd fold etc. The final output of the training and testing estimate the accuracy level of the model with the complete number of precisely classified datasets [8].

The other test option is percentage split training and testing method. This option is used 66% of the total dataset for training to build the model and the remaining 34% used for testing of the model, it means two-third of dataset for training and one-third for testing.

In the first experiments, we applied random forest algorithms using 10-fold cross validation and percentage of split techniques for training and testing purpose. The algorithm consists many decisions trees and It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. In the second experiments, support vector machine algorithm is used with the same test methods (cross validation and percentage of split technique). It uses a kernel trick technique to transform our data and then based on these transformations it finds an optimal boundary between the possible outputs. In the third experiments, multi-layer perceptron neural

network algorithm with the same test methods is used. The algorithm consists three types of layers—the input layer, output layer and hidden layer. The input layer receives the input data to be processed and the required task such as prediction and classification is performed by the output layer. An arbitrary number of hidden layers that are placed in between the input and output layer are the true computational engine of the MLP.

5.2. Model Building

The main objective of this thesis work is to build a model which uses machine learning algorithms to detect subscription fraud calls by using subscribers call detail (CDR) data. Therefore, to achieve the main objectives and answer the research questions of the study, the collected dataset has been preprocessed and selected relevant features. In addition to this, in order to show the effect of training dataset, the researcher used two training approaches and three classification algorithms. Therefore, we have six models of experimental analysis that presented in the next subsection. The experimentation results are attached in the appendix part. The experimentation results are attached in Appendix C.

5.2.1. Building a model using Random Forest algorithm

As it is thoroughly discussed in section 2.5, Random Forest is a supervised classification algorithm consists a collection of trees structured classifiers. This classifier grows independent identically distributed random vectors and each vector casts a unit vote for the most popular class at the input. The experimentation of this model has been done by employing two training approaches, which are 10-fold cross-validation and percentage split approach. Table 5.1 presents the result of RF experimentation.

Table 5. 1 performance metrics of RF classifiers

Training Approach	Result					
	Time to build model	Time to test model	Precision	Recall	F-measure	Accuracy
10-Fold Cross Validation	153.09	42.01	0.995	0.995	0.995	99.46
Percentage Split (34:66)	164.82	2.43	0.936	0.932	0.934	99.37

The highest classification accuracy that achieved in this experiment is 99.46%, from which developed applying 10-Fold Cross Validation method. This method also achieves better precision recall, and F-Measure results. The time taken to build the model is 153.09 seconds and it is a slightly (11 second) less than the model built by Percentage Split which are 164.82 seconds. In general, we can say that the model built by 10-Fold Cross Validation approach is measured a better model than Percentage split.

The confusion matrix is used to evaluate the performance of the two class problems for a given dataset. Table 5. 2 shows both correctly and incorrectly classified instances.

Table 5. 2 confusion matrix for RF classifiers

Training Approach	Correctly Classified		Incorrectly Classified	
	TP	TN	FP	FN
10-Fold Cross Validation	172308	260916	1837	514
Percentage Split (34:66)	58516	88657	624	298

The confusion matrix supports easily to visualize the algorithms accuracy level on a set of test data for which the true positive values are identified. As it is presented in RF classifier confusion matrix on table 5.2, 10-Fold Cross Validation method achieve relatively smaller incorrectly classified instances which are 0.53% or 2,351 from the total of 435,575 instances.

5.2.2. Building a model using SVM Algorithm

Support Vector Machine has promising result to telecom fraud detection problem, it is deeply discussed in Section 2.5. In this section building models using SVM algorithm performed in the same way to steps followed in Section 5.1. Following arranged settings of training methods and training Dataset, two SVM models were developed using the two training approaches. The models are presented in table 5.3.

Table 5. 3 performance metrics of SVM classifiers

Training Approach	Result					
	Time to build model	Time to test model	Precision	Recall	F-measure	Accuracy
10-Fold Cross Validation	324.68	91.43	0.949	0.947	0.947	94.66
Percentage Split (34:66)	334.55	0.36	0.947	0.945	0.945	94.5

10 – Fold cross validation method achieved the highest classification accuracy in this experiment which is 94.66%. This method also achieves relatively better precision, recall and F-measure results. The time taken to build the model is 324.68 seconds and it is also smaller than the model built by Percentage Split method. In general, we can say that the model built by 10 – Fold cross validation method is measured a better model than the other.

Similarly, SVM is used confusion matrix to measure the performance of the two class problems for a given dataset. Table 5.4 shows both correctly and incorrectly classified instances of SVM classifiers.

Table 5. 4 confusion matrix for SVM classifiers

Training Approach	Correctly Classified		Incorrectly Classified	
	TP	TN	FP	FN
10-Fold Cross Validation	167202	245139	17614	5620
Percentage Split (34:66)	56725	83234	6047	2089

The confusion matrix for SVM classifiers in table 5.4 shows that 10 – Fold cross validation method achieves relatively smaller incorrectly classified instances which are 5.34% or 23,234 from the total 435,575 instances.

5.2.3. Building a model using ANN Algorithms

Artificial Neural network or Multilayer Perceptron has promising performance to telecom fraud detection problem, it is discussed in detailed in section 2.5. In this section building models using ANN algorithm is implemented in the same way that followed in the previous Sections 5.1 and 5.2. Following arranged settings of training methods and training dataset, two ANN models were

developed using the two-training approach. Table 5.5 shows the model developed by using ANN/MLP classifiers.

Table 5. 5 performance metrics of ANN/MLP classifiers

Training Approach	Result					
	Time to build model	Time to test model	Precision	Recall	F-measure	Accuracy
10-Fold Cross Validation	235.37	65.91	0.975	0.974	0.974	97.44
Percentage Split (34:66)	229.74	0.42	0.97	0.97	0.97	96.98

Cross Validation method achieved the highest classification accuracy in this experiment which is 97.44%. This method also achieves better precision, recall and F-Measure results. But the time taken to build the model is greater than the model built by Percentage Split methods.

Similarly, with RF and SVM classifiers, ANN/MLP classifier also used confusion matrix to measure the performance of the two class problems for a given dataset. Table 5.6 shows both correctly and incorrectly classified instances of ANN/MLP classifiers.

Table 5. 6 confusion matrix for ANN/MLP classifiers

Training Approach	Correctly Classified		Incorrectly Classified	
	TP	TN	FP	FN
10-Fold Cross Validation	168604	255827	6926	4218
Percentage Split (34:66)	56571	87062	2219	2243

The confusion matrix for ANN classifiers in table 5.6 shows that 10-Fold Cross Validation method achieves relatively smaller incorrectly classified instances which are 2.56% or 11,244 from the total 435,575 instances.

Different metrics have been used to measures the performance of the models which are accuracy, precision, recall and F-Measure. Table 5.7 summarizes the entire algorithms performance results. As we can understand from the table, the highest classification accuracy scored from RF algorithm in both test options (10 – Fold cross validation and supplied test options are by RF algorithm) result of 99.46% and 99.37% respectively. Both the validation technique results are almost similar

in the case of RF algorithm. On the other hand, SVM classifier scored relatively the lowest accuracy of 94.5% from the total experimentation results using percentage of split test options. Relatively it is less than by 0.16% with the 10 – Fold cross validation test options. ANN is relatively the second better classifier in both 10 – Fold cross validation and percentage split test options with a score of 97.44% and 96.98% respectively.

Table 5. 7 Summary of performance metrics of all Algorithms

Performance Metrices	Algorithms with 10-fold cross validation			Algorithms with Percentage split		
	RF	SVM	ANN	RF	SVM	ANN
Time to build model	153.09	324.68	235.37	164.82	334.55	229.74
Time to test model	42.01	91.43	65.91	2.43	0.36	0.42
Precision	0.995	0.949	0.975	0.994	0.947	0.97
Recall	0.995	0.947	0.974	0.994	0.945	0.97
F-Measure	0.995	0.947	0.974	0.994	0.945	0.97
Accuracy	99.46	94.66	97.44	99.37	94.5	96.98

The other metrics that we used to measure the performance of the models are the time taken to build and evaluate the models. As it is presented in table 5.7, RF classifier using 10-Fold cross validation option took relatively best time (153.09 seconds) to build the model, while SVM classifier took 334.55 seconds using percentage of split option is the extended time to build the model over the other algorithms. The model evaluation is also the other metrices used to test the performance of the models. SVM classifier using percentage split test option took better evaluation time of 0.36 seconds to test the model. ANN algorithm using percentage split test option took the second better evaluation time of 0.42 seconds to test the model. However, relatively RF with percentage split test option took longer evaluation times of 2.43 seconds to test the model.

The confusion matrix supports easily to visualize the algorithms accuracy level on a set of test data for which the true positive values are identified. The confusion matrix of all experiments performed in the previous section has been summarized in Table 5.8. As it is presented in the summery table 5.8, RF algorithm using 10-Fold cross-validation method is relatively better in incorrectly classifying instances of 0.54% or 2351 from the total 435,575 instances. whereas SVM

algorithm using both percentage of split and cross validation technique are incorrectly classified 5.5% or 8,136 and 5.34% or 23, 234 instances, respectively. It is relatively the highest numbers of incorrectly classified instances.

Table 5. 8 summary of confusion matrix of all Algorithms

Test Approach	Algorithms	Correctly classified		In correctly classified	
		TP	TN	FP	FN
10-Fold cross Validation	RF	172308	260916	1837	514
	SVM	167202	245139	17614	5620
	ANN	168604	255827	6926	4218
Percentage of split	RF	58516	88657	624	298
	SVM	56725	83234	6047	2089
	ANN	56571	87062	2219	2243

Once we understand the four parameters of confusion matrix (TP, TN, FP and FN) values, it is very important to compare the performance of each algorithm in terms of precision, recall and F-measure results using both 10-fold cross validation and percentage of split test methods. As it is presented in Figure 5.1, RF algorithm result using 10-Fold cross validation test option shows that all the metrices precision, recall and F-measure have scored relatively the highest result of 0.995 in each parameter. Also, ANN algorithms show the second better results in all metrices; whereas, SVM algorithms show relatively the lowest results in all metrices from the three algorithms. When we observe the precision value in each model, relatively better results have been scored, and it means the lower false positive instances are detected. In fraud detection, the smaller false positive detection is preferred because it reduces the risk of blocking genuine customers.

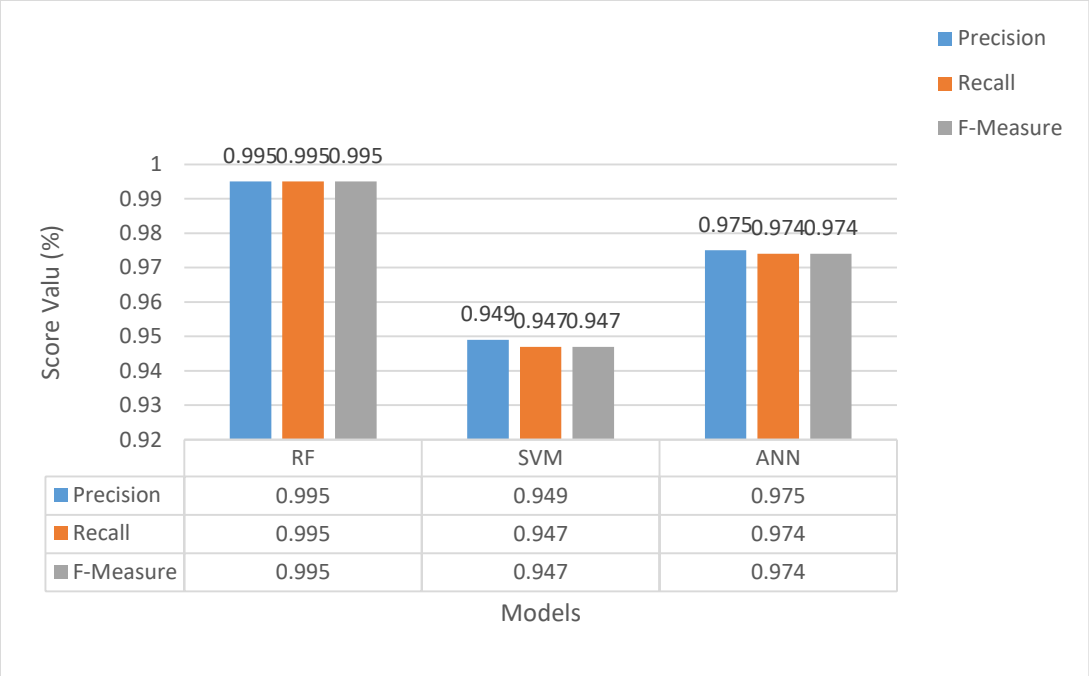


Figure 5. 2 Models Comparison using 10- fold cross validation method.

In the same way, we have compared the performance of the entire algorithms using percentage of split test options with similar performance measurement metrics (precision, recall and F-measure). As it is presented in figure 5.2, from the graph, Again the RF algorithm using percentage of split test option shows that all the metrices precision, recall and F-measure scores the highest result of 0.994. Also, ANN algorithms show the second better results in all metrices; whereas, SVM algorithms show relatively the lowest results in all metrices from the three algorithms. Similarly, when we observe the precision value in each model, relatively higher results have been scored. It means, the lower false positive instances are detected. In fraud detection, the smaller false positive detection is preferred because it reduces the risk of blocking genuine customers.

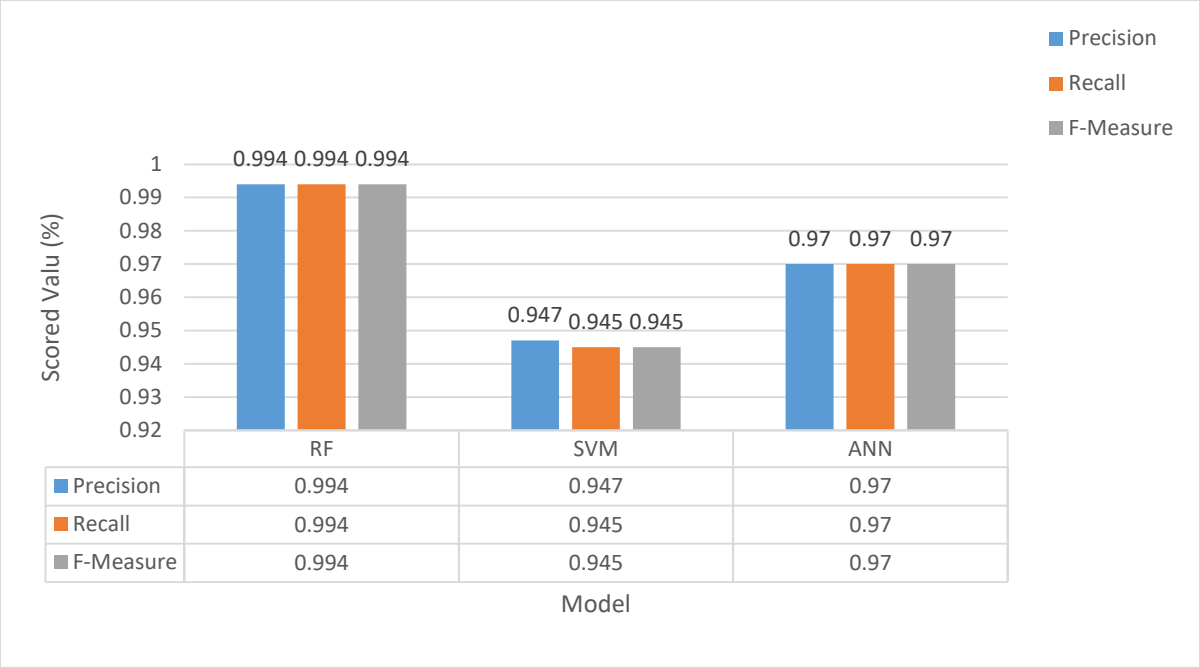


Figure 5. 3 Models Comparison using percentage of split method.

5.3. Model Evaluation

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and overfitted models [47].

In section 5.2, we have built different models and selected the best one with best performance, which is RF algorithms. In this section we used test set (new data) as presented in process model and evaluate the performance of the selected model in detecting subscription fraud. It is used to provide an unbiased evaluation of the final model fit on the training dataset. Many a times the validation set is used as the test set, but it is not good practice. The test set is generally well curated. It contains carefully sampled data that spans the various classes that the model would face, when used in the real environment. The new test dataset used in this section to evaluate the performance of the selected model have a data set of 291,204 records. Table 5.9 shows evaluation result of the selected best model using new test dataset. The detail experimentation result is attached in Annex D.

Table 5. 9 Evaluation result of best model using test dataset.

Training Approach	Result					
	Time to build model in second	Time to test model in second	Precision	Recall	F-measure	Accuracy
10-Fold Cross Validation	81.19	32.04	0.999	0.999	0.999	99.0265

5.3.2. Subjective evaluation

Besides the objective evaluation, user acceptance test is subjectively conducted in order to demonstrate the functionality and effectiveness of the model. This helps to improve the accuracy of the model and ensure the quality of the proposed solution on the detection of subscription fraud. The output of this research had been discussed individually with domain experts of ethio telecom fraud management section, revenue assurance, internal IT technical audit team and others at information systems division experts. A total of 10 experts have been selected and evaluated the models with different criteria.

The domain experts are certain that the selected attributes used in this research are relevant for the detection of subscription fraud. Some of these attributes are also important for domain expert practically working to protect the telecommunication frauds. The analysis results also discussed with the experts and achieved a good result. The domain experts also noted that, this research is helpful and play an important role to control and prevent the current threat on the mobile communication of ethio telecom.

The model was evaluated by the domain experts with some criteria weather the proposed model can satisfy their needs to achieve specific goals. The model evaluation result is summarized for each of the characteristics and sub-characteristics as shown in table 5.10.

Table 5. 10 Subscription fraud detection model evaluation questionnaires.

Evaluation Criteria	Performance values				
	5	4	3	2	1
1. Effectiveness					
How do you evaluate the accuracy and completeness of the model with which users achieve specified goals	8	1	1		
2. Satisfaction					
Degree to which a user or other stakeholder has confidence that the model will behave as intended	7	2	1		
Overall rating of the proposed model	6		4		
3. Freedom from risk					
Rate degree to which the model mitigates the potential risk to financial status, efficient operation, commercial property, reputation or other in telecom fraud	2	1	7		
4. Context coverage					
Degree to which the model can be used with effectiveness, efficiency, freedom from risk and satisfaction in subscription fraud detection	6	1	3		
Degree to which the model can be used with effectiveness, efficiency, freedom from risk and satisfaction in contexts beyond those initially specified in the requirements	7	1	2		
The implementation of the proposed model can fit with the organization process	7	1	2		

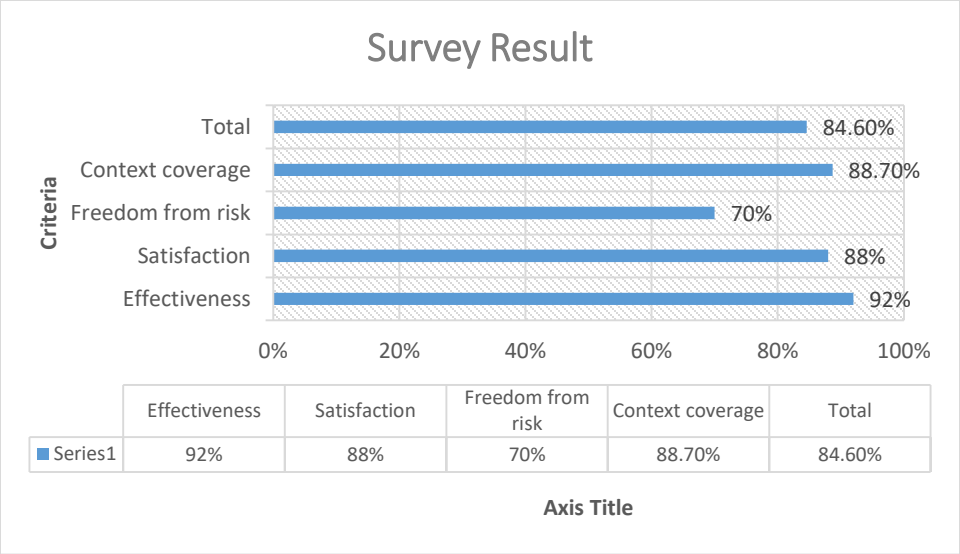


Figure 5. 4 Model Subjective Evaluation Result

The performance evaluation results of table 5.10 is computed and presented in figure 5.4 above. The subjective evaluation result of the model shows that it is 92% effective, 70% freedom from risk, 88% satisfaction and 88.7% context coverage. Overall, the model has 84.6% quality.

5.4. Discussion of the result

Fraud detection problems are found in many sectors of lives endeavor and the telecoms sector is not an exception. Hence fraud detection is referred to as the attempt engaged in discovering illegitimate usage of a communication network by identifying fraud as quickly as possible once it has been confirmed. The radical changes in the telecommunications sector have made it difficult to control and detect fraudulent activities. Thus, to achieve positive results the problem of fraud requires to be handled with serious and effective attention. Machine learning technique has been found to be very useful in fraud detection.

This work thus identified ten attributes which are best to detect subscription frauds using machine learning techniques. Some of the attributes were also identified and used in some other related research. But in this research, we have been used two derived attributes (RATIO_DISTCALL_TOTAL and RATIO_INT_TOT) which are new to detect subscription fraud. The RATIO_DISTCALL_TOTAL attribute was derived from distinct call over the total call and the RATIO_INT_TOT attribute was derived from international call over total calls.

As discussed in chapter four in section 4.3, The source of data for this research has been collected from ethio telecom databases two-month postpaid and prepaid CDR data. Due to the fact that the CDR data is very huge and requires more space, the researcher used a separate server to store and preprocess the data. The researcher has used random sampling technique to select the sample data for experimentation. After the necessary data preprocessing was done, the researcher has been selected and used 435,575 datasets for the purpose of conducting this study. Other related research [8] on the detection of subscription fraud uses only prepaid customer CDR data with 25,000 records for experimentation.

The researcher evaluates the information content of the attributes in collaborated with domain experts, 10 most appropriate attributes have been selected and used for conducting this research. To extract and manage the huge CDR from the telecommunication server and to integration process of the data was very challenging times for the researcher and it was not possible without the support of ethio telecom domain experts. Therefore, the researcher took a lot of time for this process.

One of the basic targets of data analytics is to compare different models and to select the better classification accuracy based on the results. Therefore, detailed experimentation for different models has been conducted. Accordingly, the best classification algorithm which is appropriate for subscription fraud detection has been selected. The classification model used in this research is to predict subscription frauds. The attributes selected and used in this research for the prediction of subscription frauds are evaluated by domain experts. Some of the selected attributes are also important for domain expert practically practicing protecting the telecommunication frauds. The analysis, which was closely undertaken with domain experts are, achieved a good result.

The research will play an important role to controlling and preventive the current threat on the mobile communication of ethio telecom. Now a day the telecom fraud is very serious problem in telecommunication industries. Telecom fraud is offences, and it is a serious threat to national security beyond economic losses so that this research will open new research areas in the field of subscription telecom fraud detection and preventive mechanism.

During the discussion with ethio telecom domain experts, the researcher noted that the most difficult aspect of fighting fraud is how can detect subscription frauds in the ethio telecom. Because the major behavior of the fraudsters who made subscription frauds are calling frequently without the intention to paying and finally disappearing from the network. As discussed before from those experiments conducted in supervised approach, the Random forest algorithm with the 10-fold cross validation model gives a better classification accuracy of predicting newly arriving class category.

As it has been presented in chapter one, the main objective of this research is to build a model which uses machine learning techniques to detect subscription fraud by using subscribers call detail records (CDR) data. Accordingly, classification models building experiment have been achieved and summarized in Table 5.7 using the three algorithms and two test options. As a result, the models which have been developed with RF algorithms in both test options 10-fold cross-validation and percentage split are considered as relatively best and almost have similar results as compare to the other two algorithms for the detection of subscription fraud.

The goal of this study is subscription fraud detection using machine learning technique based on the customers call transaction patterns. In this research, we have identified best attributes which are relevant for the detection of subscription fraud. The following are best attributes used to identify subscription fraud from legitimate customers.

- Subscribers total number of calls
- Number of Unique called Numbers
- Number of Incoming Call
- Total outgoing calls duration
- Sum of Total call fee
- Ration of Dist_Call & Total_Call
- Total International calls
- Ration of INT_TOT
- Number of SMS sent
- Total data usage

There are different types of machine learning algorithms and test options that can be used for the purpose of predicting subscription fraud. In this research, we have used three machine learning algorithms (RF, SVM and ANN) and two test option (10-Fold cross validation and percentage of split) to build the models. As a result, RF algorithm with 10-Fold cross validation test option is more suitable for the purpose of predicting subscription frauds.

CHAPTER SIX

Conclusion and Recommendation

6.1. Conclusion

Telecommunication technologies are becoming dynamic and innovative from time to time. Telecom operators, who are engaged in managing and operating the complex telecommunication network and information system infrastructures for the purpose of voice and data transmission are in a serious challenge to control the telecom infrastructure from fraud and cyber-attacks. Because of fraudsters flexibility and enhancing themselves along with new technologies and behavioral changes, telecom operators are challenged by fraudsters activities. There are many types of telecommunication fraud. Subscription fraud type is one of the common and major types among telecom frauds in which the usage category is in contradiction with the initial subscription type. Subscription fraud is a contractual fraud type, and it is the starting point for many other telecom fraud types. Even though many security policies and systems have been implemented, subscription fraud is still prevalent and challenging every telecommunication companies.

Subscription fraud is becoming the most thoughtful threat for telecommunication companies all over the world. The main objective of the fraudsters is making profit illegally or getting telecom services with the intention of not to pay for the service they used. As a result, it is caused in telecommunication operators revenue loss and its subscriber's loss quality of service.

The behavior of subscription fraud is not statics and its behavior will change frequently. Therefore, detect subscription fraud type using traditional fraud management systems like rule-based engine system is not efficient.

The main objective of this research is to build a model which uses machine learning techniques to detect subscription fraud by using subscribers call detail records (CDR) data. The dataset used in this study has been taken from ethio telecom customer relation management (CRM) system. Additionally, blocked customer data were collected from ethio telecom fraud management systems (FMS).

This research was started by reviewing different literatures conducted in telecommunication fraud detection, which is essential to know the significance of the problem, identifying its constraints of existing fraud detection.

In order to enhance the performance of subscription fraud detection, we have selected and used three classification algorithms (RF, SVM and ANN) based on the reference of several literature and domain experts in the area of telecom fraud detection. The models were developed using WEKA data mining tool.

After the selection of machine learning algorithms and datasets collection, the dataset was preprocessed and became suitable for data mining experiment. Initially, the two-month data was collected and preprocessed. Finally, 262,754 (60.32%) legitimate call records and 172,821 (39.68) fraudulent call records have been selected and used for experimentation. The dataset was initially containing a total of 33 attributes. And 10 attributes, which are relevant for this work have been selected and used. The dataset is converted into ARFF format which is suitable for WEKA to data analytics. For each algorithm, we have used two test options which are 10-Fold cross validation and percentage of split for building the model and testing purpose. The second option has been used 66% of the total dataset for training to build the model and the remaining 34% for testing the model.

In general, the experimentation gives encouraging solution for the detection of subscription fraud and provide useful contribution to the field of fraud detection in telecom sectors. The evaluation results show that RF algorithm performs 99.46% of accuracy and the other algorithms ANN and SVM scores 97.49 % and 96.02% accuracy, respectively. The model with RF classification algorithm also attains better results in all performance evaluation metrics of precision, recall, F-measure, and time to build and evaluate the models.

In this research we used CDR data to identify fraudulent customer from legitimate. It is a valuable data source for telecom data analysts to gain information about mobile subscribers' behavior or develop features for predictive models. However, call detail records analysis can be used to extract even more profound insights about individual subscribers. Subscribers that can be derived from CDR data used to improve telecom customer segmentation and predictive model accuracy.

Detection and prevention of telecom frauds are the main targets of the telecommunication companies. In this research, we construct different subscription fraud detection models using machine learning techniques and proposed a single better performed model. It is found that when a number of models are combined, instead of using a single model in isolation, there is improved performance.

6.2. Recommendation

6.2.1. Recommendation for practice

The result of this research will help and guide ethio telecom information system security and revenue assurance departments to control and manage telecom frauds, specifically to detect subscription frauds and to realize the intended business benefits of the company. The models are supposed to provide insight for the possible solutions for the detection of subscription fraud types.

Ethio telecom should improve the fraud management techniques and approach that currently used to detect genuine customers from fraudsters. Rule-based telecom fraud detection technique has lots of limitations and fraudsters can easily pass the telecom infrastructures. Deploying the proposed model is recommended to improve the detection of subscription fraud.

6.2.2. Recommendation for future work

A potential research area can be the creation of meaningful initial behavior profiles for new customers so that fraud detection methods may be used from the first call data record. One way to attain this is to group the behavior profiles for existing customers based on information available in the first one or few calls for a customer. Customer behavior may change over time demanding a technique to keep the subscriber's behavior profile current, which is another area of potential study.

Another area of potential study is a combination of machine learning techniques that may be used to build a comprehensive fraud detection model that is capable of outperforming models based on a single machine learning methodology.

Telecommunication fraud can be detected either online or offline modes. This research is used offline CDR data to identify subscription fraudsters. But, in a fraud detection system that is

working in real time, it is crucial to minimize the gap between the time when the fraud happened and the time when it was detected. So, working in real time subscription fraud detection is recommended.

This study is focuses on subscription fraud detection. However, there are many other types of telecom frauds in the industry, hence, performing similar research on other fraud types and methods is advisable. Specially, Over the Top (OTT) is a serious problem in the telecom industries and it is recommended to work on this type of fraud.

References

[1] M. Sahin, A. Francillon, P. Gupta, and M. Ahamad, (2017) "Sok: Fraud in telephony networks," in Security and Privacy (EuroS&P), IEEE European Symposium on, IEEE, 2017, pp. 235–250.

[2] AlBougha, Mhd Redwan, (2016). "Comparing Data Mining Classification Algorithms in Detection of Sim-box Fraud" Culminating Projects in Information Assurance.

[3] A.-A. Ababa. (2017), "Ethiopia-telecom fraud. A.-A. Ababa, Ed., [Online]" Retrieved on 20 Jan 2019, from: http://apanews.net/en/news/ethiopia-loses-over-52mto-telecom-fraud-official.

[4] G. Yeshinigus (2013). "Predictive modeling for fraud detection in telecommunication" the case of ethio telecom, Unpublished Master's Thesis, Department of Information Science, AAU, Ethiopia

[5] Kabari, Ledisi & Nuka, Nanwin & Nquoh, Edikan. (2016), "Telecommunications Subscription Fraud Detection Using Naïve Bayesian Network". IIARD International Journal of Computer Science and Statistics, www.iiardpub.org. ISSN 2467-5832. 2467-5832.

[6] H. Kahsu (2018)," Sim-box fraud detection using data mining techniques": The case of ethio telecom, Unpublished Master's Thesis, Institute of Technology School of Electrical and Computer Engineering, AAU, Ethiopia.

[7] P. Saravanan, V. Subramaniyaswamy, N. Sivaramakrishnan, M. Prakash, and T. Arunkumar (2014), "Data mining approach for subscription-fraud detection in telecommunication sector," Contemporary Engineering Sciences.

[8] H. Tesfaye (2013)" constructing predictive model for subscription fraud detection using data mining technique for the case of ethio telecom", Unpublished Master Thesis department of information science AAU, Ethiopia.

[9] P. Hoath. (1998), "Telecoms fraud, the gory details". Computer Fraud and Security, vol. 20 no. 1 pp. 10–14.

[10] J. Hollmen (2000), "User Profiling and Classification for Fraud Detection in Mobile Communication Networks" PhD thesis, Helsinki University of Technology, Department of Cognitive and Computer Science and Engineering, Espoo, Finland.

[11] M. Johnson (1996)," Cause and effect of telecoms fraud", Telecommunication International Edition, vol. 30, no. 12, pp. 80–84.

[12] P. Gosset and M. Hyland (1999), "Classification, detection and prosecution of fraud in mobile networks", In Proc. ACTS Mobile Summit, Sorrento, Italy.

[13] Howells, V. Scharf-Katz, and P. Staple (2009), "TELECOM FRAUD 101: Fraud Types, Fraud Methods, & Fraud Technology", n.d. Retrieved on 14 Feb2019 from http://www.argyledata.com/files/Telecom-Fraud-101-eBook.pdf

[14] TY - BOOK T1, (1993) - The concise Oxford dictionary of English etymology AU - edited by T.F. Hoad PY - 1993 PB - Oxford; New York: Oxford University Press. AB - xiv, 552 pages.

[15] M. M. Sepehri, and H. Farvaresh (2011), "A data-mining framework for subscription fraud detecting in telecommunication," Engineering Applications of Artificial Intelligence, vol. 24, no. 1, pp. 182–194.

[16] B. Davis and S. K. Goyal (1993), "Management of cellular fraud: Knowledge based detection, classification and prevention". In Proceedings of the 13th International Conference on Artificial Intelligence, Expert Systems and Natural Language, Vol. 2, no. 11 pages 155–164.

[17] M. I. Akhter and M. G. Ahamad (2012), "Detecting telecommunication fraud using neural networks through data mining" Int. J. Sci. Eng. Res, vol. 3, no. 3, pp. 601–606.

[18] L. G. Kabari, D. N. Nanwin, and E. U. Nquoh (2016), "Telecommunications subscription fraud detection using artificial nueral networks," Transactions on Machine Learning and Artificial Intelligence, vol. 3, no. 6, p. 19.

[19] W. Moudani and F. Chakik,(2013) "Fraud detection in mobile telecommunication," Lecture Notes on SOftware Engineering, vol. 1, no. 1, p. 75.

[20] L. G. Kabari, D. N. Nanwin, and E. U. Nquoh (2016), "Telecommunications subscription fraud detection using naïve bayesian network," International Journal of Computer Science and Mathematical Theory, vol. 2, no. 2.

[21] L. G. Kabari, D. N. Nanwin, and E. U. Nquoh (2016), "Telecmmunications subscription fraud detection using artificial nueral networks," Transactions on Machine Learning and Artificial Intelligence, vol. 3, no. 6, p. 19.

[22] M. R. AlBougha (2016), "Comparing data mining classification algorithms in detection of sim-box fraud," St. Cloud State University the Repository at St. Cloud State.

[23] S. Pan, T. Morris, and U. Adhikari (2015), "Developing a Hybrid Intrusion Detection System Using Data Mining for Power Systems," IEEE Trans. Smart Grid, vol. 6, no 1

[24] S. Subudhi and S. Panigrahi (2015) Quarter-sphere support vector machine for fraud detection in mobile telecommunication networks. Procedia Comp Sci 48:353–359.

[25] G. Williams (2011), "Data mining with Rattle and R": The art of excavating data for knowledge discovery. Springer Science & Business Media.

[26] Breiman, (2001), "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32.

[27] Cutler, D.R. Cutler, J.R. Stevens (2012)- Ensemble machine learning, DOI: 10.1007/978-1-4419-9326-7_5 · Source: DBLP

[28] K. Mishra, S. V. Ramteke, P. Sen, and A. K. Verma (2017), "Random forest tree based approach for blast design in surface mine," Geotechnical and Geological Engineering, vol. 36, no. 3, pp. 1647–1664.

[29] Fausett, Laurene (1994), Fundamentals of Neural Networks: Architectures, Algorithms and Applications, Prentice-Hall, New Jersey, USA.

[30] Shiruru, Kuldeep, (2016). AN INTRODUCTION TO ARTIFICIAL NEURAL NETWORK. International Journal of Advance Research and Innovative Ideas in Education. Vol 1. No. 2, pp.27-30.

[31] S. Zhang, C. Zhang, and Q. Yang (2003), "Data preparation for data mining," Applied artificial intelligence, vol. 17, no. 5-6, pp. 375–381.

[32] G. Zhang, S. Fischer-Hubner (2011) Detecting near-duplicate spits in voice ¨mailboxes using hashes. ISC. Springer, pp 152–167

[33] C. M. Held, C. A. Perez and P. A. Estévez (2006) "prevention of Subscription fraud in telecommunications using neural networks and fuzzy rules " Expert Systems with Applications, vol. 31, no. 2, pp. 337–344.

[34] M.A. Azad, R. Morla (2013) Caller-rep: Detecting unwanted calls with caller social strength. Comput Secur vol. 39, no. 1, pp.219–236

[35] S. Subudhi and S. Panigrahi (2017) Use of Possibilistic fuzzy C-means clustering for telecom fraud detection. In: Behera H, Mohapatra D (eds) Computational intelligence in data mining. Advances in intelligent systems and computing, vol 556. Springer, Singapore.

[36] G. Kabari, D. N. Nanwin, and E. U. Nquoh (2016), "Telecommunications subscription fraud detection using artificial neural networks," Transactions on Machine Learning and Artificial Intelligence, vol. 3, no. 6, p. 19.

[37] J.W. Creswell (2009). Research Design: Qualitative, Quantitative and Mixed Method Approaches (3rd Ed.). Los Angeles: SAGE Publications.

[38] W. Tarikua (2018), "Predictive Modeling for International Roaming Fraud Detection" the case of ethio telecom, Unpublished Master Thesis, department of information science, AAU, Ethiopia.

[39] J.W. Creswell (1998), Research Design: Qualitative, Quantitative and Mixed methods approaches (2nd ed.). Thousand Oaks, CA: Sage.

[40] Wikipedia Encyclopedia (2005), Retrieved on 19 Jan 2019, from: http://en.wikipedia.org/wiki/Socialscience.pdf

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

[42] El Arass, Mohammed & Khadija, Ouazzani Touhami & Souissi, Nissrine. (2020). Data Life Cycle: Towards a Reference Architecture. International Journal of Advanced Trends in Computer Science and Engineering. Vol. 9, pp. 5645 - 5653. 10.30534/ijatcse/2020/215942020.

[43] S. Sukamolson (2007), “Fundamentals of quantitative research”, Language Institute Chulalongkorn University, 1-20. Thailand https://www.academia.edu › Fundamentals_of_quantitative_research

[44] Sebastian Raschka (2018) Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. University of Wisconsin–Madison Department of Statistics.

[45] K. Cios, R. Wieniawski, W. Pederick & L. Kurgan (2007), “The Knowledge Discovery Process” In Data Mining, Springer US.

[46] S. Ahmed, M. U. Ali, J. Ferzund, M. A. Sarwar, A. Rehman and A. Mehmood (2017), "Modern Data Formats for Big Bioinformatics Data Analytics," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 8, no. 4.

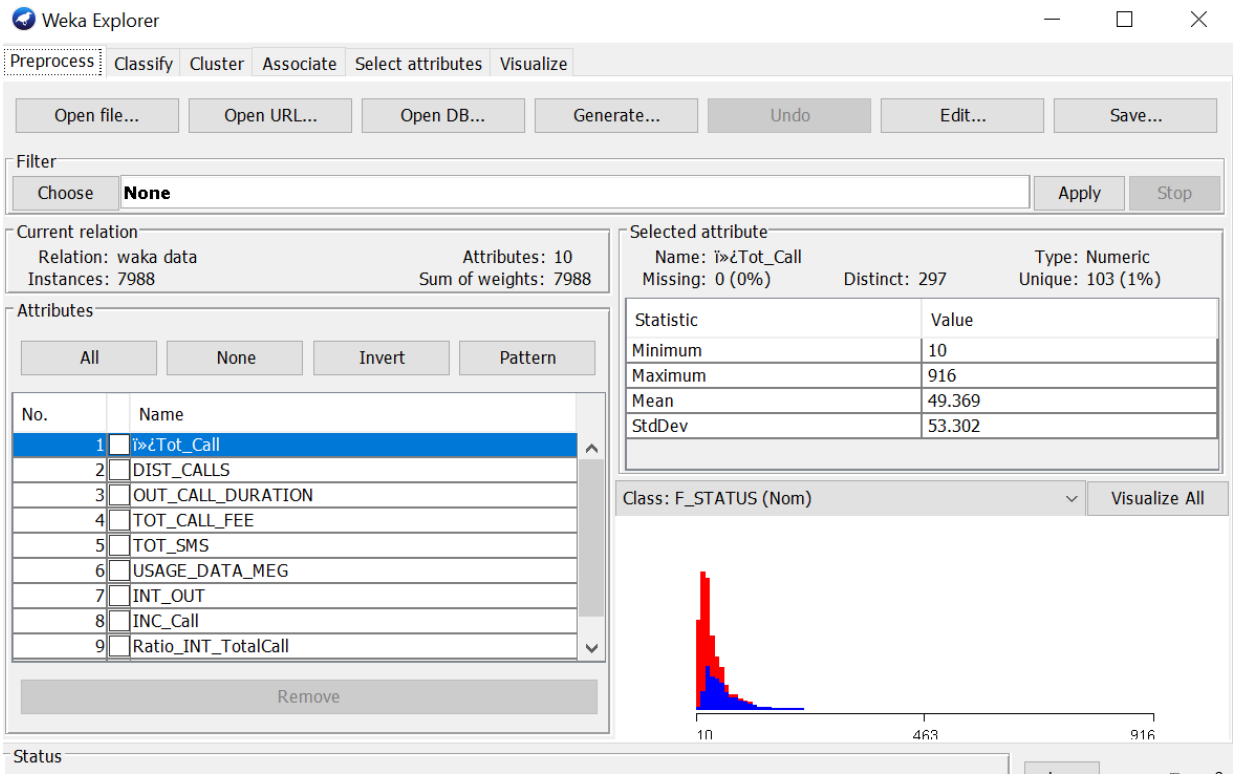
[47] F. Musumeci et al (2019), "An Overview on Application of Machine Learning Techniques in Optical Networks," in IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1383-1408, doi: 10.1109/COMST.2018.2880039.

Appendix

a) Sample data snapshot

	A	B	C	D	E	F	G	H	I	J	K
1	Total_Call	DIST_CALL	OUT_CALL	RATIO_DI	TOT_CALL	TOT_SMS	USAGE_DA	INT_OUT	InC_Call	Ratio_INT_	F_STATUS
2	47	13	20.17	0.28	5.662	1	3.5	28	7	0.6	Y
3	189	1	1.33	0.01	0.667	1	25	187	0	0.99	Y
4	85	11	95	0.13	27.36	1	8	65	5	0.76	Y
5	122	10	34.67	0.08	12.47	7	14.5	95	6	0.78	Y
6	246	20	117	0.08	48.4365	9	19.5	197	0	0.8	Y
7	1016	5	6	0	3.0017	22	134.5	1008	0	0.99	Y
8	156	16	37.33	0.1	16.24	2	20.5	128	0	0.82	Y
9	78	12	133.67	0.15	47.27	1	5.5	45	25	0.58	Y
10	41	3	10	0.07	4.5023	7	3.5	33	1	0.8	Y
11	172	3	6.83	0.02	3.3432	12	22	165	0	0.96	Y
12	48	9	20	0.19	10.0053	7	3.5	36	0	0.75	Y
13	94	10	60.33	0.11	25.6801	1	10.5	76	9	0.81	Y
14	12	9	77.67	0.75	38.853	0	0	0	0	0	N
15	9	6	51	0.67	18.8361	0	0	0	1	0	N
16	36	18	83	0.5	33.06	1	1	7	7	0.19	N
17	4	4	30	1	15.0076	0	0	0	0	0	N
18	7	3	37	0.43	11.8864	0	0	0	3	0	N
19	12	7	73	0.58	25.084	0	0	0	6	0	N
20	3	2	1.33	0.67	0.6671	0	0	0	0	0	N
21	41	20	88	0.49	41.6662	1	1	7	2	0.17	N

b) Weka screenshot



c) Experiment output for Training

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Relation: Final Data csv format

Instances: 435575

Attributes: 11

Total_Calls

DIST_CALLS

OUT_CALL_DURATION

RATIO_DISTCALL_TOTAL

TOT_CALL_FEE

TOT_SMS

USAGE_DATA_MEG

INT_OUT

InC_Call

Ratio_INT_Tot

F_STATUS

=== Test mode: split 66.0% train, remainder test===

Classifier model (full training set)

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 164.82 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 2.43 seconds

=== Summary ===

Correctly Classified Instances	147173	99.3774 %
Incorrectly Classified Instances	922	0.6226 %
Kappa statistic	0.987	
Mean absolute error	0.0092	
Root mean squared error	0.0711	
Relative absolute error	1.9154 %	
Root relative squared error	14.5324 %	
Total Number of Instances	148095	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.995	0.007	0.989	0.995	0.992	0.987	0.999	0.998	Y
	0.993	0.005	0.997	0.993	0.995	0.987	0.999	0.999	N
Weighted Avg.	0.994	0.006	0.994	0.994	0.994	0.987	0.999	0.998	

=== Confusion Matrix ===

a	b	<-- classified as

```
 58516   298 |     a = Y
   624 88657 |     b = N
```

```
=== Run information ===
```

Scheme: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Relation: Final Data csv format

Instances: 435575

Attributes: 11

- Total_Calls
- DIST_CALLS
- OUT_CALL_DURATION
- RATIO_DISTCALL_TOTAL
- TOT_CALL_FEE
- TOT_SMS
- USAGE_DATA_MEG
- INT_OUT
- InC_Call
- Ratio_INT_Tot
- F_STATUS

Test mode: 10-fold cross-validation

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

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 153.09 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	433224	99.4603 %
Incorrectly Classified Instances	2351	0.5397 %
Kappa statistic	0.9887	
Mean absolute error	0.0083	
Root mean squared error	0.0664	
Relative absolute error	1.7293 %	
Root relative squared error	13.5689 %	
Total Number of Instances	435575	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.997	0.007	0.989	0.997	0.993	0.989	0.999	0.998	Y
	0.993	0.003	0.998	0.993	0.996	0.989	0.999	0.999	N
Weighted Avg.	0.995	0.005	0.995	0.995	0.995	0.989	0.999	0.999	

=== Confusion Matrix ===

a	b	<-- classified as
172308	514	a = Y
1837	260916	b = N

=== Run information ===

Scheme: weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

Relation: Final Data csv format

Instances: 435575

Attributes: 11

Total_Calls

DIST_CALLS

OUT_CALL_DURATION

RATIO_DISTCALL_TOTAL

TOT_CALL_FEE

TOT_SMS

USAGE_DATA_MEG

INT_OUT

InC_Call

Ratio_INT_Tot

F_STATUS

Test mode: split 66.0% train, remainder test

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

Sigmoid Node 0

Inputs Weights

Threshold 2.5032738911534147

Node 2 6.438554819439504

Node 3 -3.037955356989566

Node 4 -2.394790524601822

Node 5 -4.283440039086201

Node 6 12.26574314765701

Node 7 -2.2064103183017787

Sigmoid Node 1

Inputs Weights

Threshold -2.503273891153407

Node 2 -6.438554819439502

Node 3 3.0379553569895674

Node 4 2.394790524601818

Node 5 4.2834400390861935

Node 6 -12.265743147657004

Node 7 2.206410318301776

Sigmoid Node 2

Inputs Weights

Threshold 203.32146423444433

Attrib Total_Calls 23.936903886376466

Attrib DIST_CALLS -9.598129551761563

Attrib OUT_CALL_DURATION 0.6459377652197166

Attrib RATIO_DISTCALL_TOTAL -12.040976873653037

Attrib TOT_CALL_FEE 9.20345122870259

Attrib TOT_SMS -364.9798025476071

Attrib USAGE_DATA_MEG -186.87537574536506

Attrib INT_OUT 577.3552269820432

Attrib InC_Call -11.165998281675947

Attrib Ratio_INT_Tot 186.37218677678882

Sigmoid Node 3

Inputs Weights

Threshold 36.892812152925444

Attrib Total_Calls -0.7224515387969671

Attrib DIST_CALLS 141.3304069772249

Attrib OUT_CALL_DURATION -27.46429603960688

Attrib RATIO_DISTCALL_TOTAL 7.179617361518608

Attrib TOT_CALL_FEE -67.82042978559357

Attrib TOT_SMS -19.479997931453376

Attrib USAGE_DATA_MEG -43.81948887937633

Attrib INT_OUT 68.27626960072385

Attrib lnC_Call 16.323711047427807

Attrib Ratio_INT_Tot -32.43139089139044

Sigmoid Node 4

Inputs Weights

Threshold -23.61778057729099

Attrib Total_Calls -1.2684812283880629

Attrib DIST_CALLS 32.52064511577796

Attrib OUT_CALL_DURATION -12.050330872338671

Attrib RATIO_DISTCALL_TOTAL 21.704119969257036

Attrib TOT_CALL_FEE -11.021608494680722

Attrib TOT_SMS 4.893499562306068

Attrib USAGE_DATA_MEG 3.709486421801124

Attrib INT_OUT -11.853282151729767

Attrib InC_Call 11.29749229292542

Attrib Ratio_INT_Tot -55.58086616817221

Sigmoid Node 5

Inputs Weights

Threshold -69.17282138972517

Attrib Total_Calls 4.3444390145675955

Attrib DIST_CALLS 49.01988655353867

Attrib OUT_CALL_DURATION 0.535263579689939

Attrib RATIO_DISTCALL_TOTAL 15.050660341552257

Attrib TOT_CALL_FEE 40.641470562165544

Attrib TOT_SMS -21.324110856953887

Attrib USAGE_DATA_MEG 16.92807354264477

Attrib INT_OUT -84.7589879148521

Attrib InC_Call -70.17143079580575

Attrib Ratio_INT_Tot -22.225622809676768

Sigmoid Node 6

Inputs Weights

Threshold 201.66993188002212

Attrib Total_Calls -29.55099771448358

Attrib DIST_CALLS 16.84270767594472

Attrib OUT_CALL_DURATION 4.969581187634541

Attrib RATIO_DISTCALL_TOTAL 1.501392356918009

Attrib TOT_CALL_FEE -185.65061329792167

Attrib TOT_SMS 33.28816834452932

Attrib USAGE_DATA_MEG -199.3687460017176

Attrib INT_OUT 604.3106999673156

Attrib InC_Call -4.865271542201331

Attrib Ratio_INT_Tot -14.859644929073902

Sigmoid Node 7

Inputs Weights

Threshold -167.37159202601595

Attrib Total_Calls 30.32130975116004

Attrib DIST_CALLS -9.212788166948433

Attrib OUT_CALL_DURATION 2.5774195038133376

Attrib RATIO_DISTCALL_TOTAL 15.28847137741082

Attrib TOT_CALL_FEE 62.7134893177351

Attrib TOT_SMS 2.0558138520793277

Attrib USAGE_DATA_MEG 101.34119750056078

Attrib INT_OUT -190.40378159968546

Attrib InC_Call 0.20513292284995646

Attrib Ratio_INT_Tot -183.66337247694233

Class Y

Input

Node 0

Class N

Input

Node 1

Time taken to build model: 229.74 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.42 seconds

=== Summary ===

Correctly Classified Instances	143633	96.9871 %
Incorrectly Classified Instances	4462	3.0129 %
Kappa statistic	0.9371	
Mean absolute error	0.0362	
Root mean squared error	0.1401	
Relative absolute error	7.5518 %	
Root relative squared error	28.6401 %	
Total Number of Instances	148095	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.962	0.025	0.962	0.962	0.962	0.937	0.997	0.995	Y
	0.975	0.038	0.975	0.975	0.975	0.937	0.997	0.997	N
Weighted Avg.	0.970	0.033	0.970	0.970	0.970	0.937	0.997	0.996	

=== Confusion Matrix ===

a	b	<-- classified as
56571	2243	a = Y
2219	87062	b = N

d) Experiment output for Testing

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Relation: Test Data

Instances: 291204

Attributes: 11

Total_Calls

DIST_CALLS

OUT_CALL_DURATION

RATIO_DISTCALL_TOTAL

TOT_CALL_FEE

TOT_SMS

USAGE_DATA_MEG

INT_OUT

InC_Call

Ratio_INT_Tot

F_STATUS

Test mode: 10-fold cross-validation

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

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 81.19 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	288369	99.0265 %
Incorrectly Classified Instances	2835	0.9735 %
Kappa statistic	0.9766	
Mean absolute error	0.0125	
Root mean squared error	0.0854	
Relative absolute error	3.0144 %	
Root relative squared error	18.7477 %	
Total Number of Instances	291204	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.993	0.015	0.994	0.993	0.993	0.977	0.998	0.999	N
	0.985	0.007	0.982	0.985	0.983	0.977	0.998	0.996	Y
Weighted Avg.	0.990	0.013	0.990	0.990	0.990	0.977	0.998	0.998	

=== Confusion Matrix ===

a	b	<-- classified as
203975	1521	a = N
1314	84394	b = Y