A Comparative Analysis of Machine Learning Algorithms for Subscription fraud Detection: The case of ethio telecom

PREPARED BY: DEREBE TEKESTE

ADVISER: EPHREM TESHALE (PHD)

A Thesis submitted to
School of Electrical and Computer Engineering
Addis Ababa Institute of Technology

In Partial Fulfillment of the Requirements for the Degree of Master of Telecommunication Engineering (TIS)

ADDIS ABABA UNIVERSITY
Addis Ababa, Ethiopia
February 21, 2020

Declaration of originality

I, the undersigned, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research. I trully acknowledged and referred every materials which used in this thesis work.

DEREBE TEKESTE

Name

Signature

ADDIS ABABA UNIVERSITY

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Thesis on

A Comparative Analysis of Machine Learning Algorithms for Subscription fraud Detection: The case of ethio telecom

By: **DEREBE TEKESTE**

Signed by :

Adviser	Ephrem Teshale (PhD)	Signature	_____	Date	_____
Evaluator	_____	Signature	_____	Date	_____
Evaluator	_____	Signature	_____	Date	_____

Dedication

This thesis is dedicated to the memory of my beloved father Tekeste Birhanu, for making me be who I am. I still miss him every day.

ABSTRACT

In these days due to the development of affordable technologies, the number of subscribers and revenue-generating increased over the past few years in the telecommunication industry. However, advancements of the telecom industry provides certain appearances that stimulate fraudsters. One of the common and predominant fraud types is subscription fraud. It is usually the precursor to other fraud types. Since 2013 subscription fraud is listed as a top-five predominant fraud type. Subscription fraud alone causes billions of dollar losses of telecomm companies.

This thesis is conducted on comparative performance of three supervised machine learning algorithms Artificial Neural Network (ANN), Support Vector Machine (SVM) and J48, done using two classification techniques. Before analyzing and comparing the algorithms Call Detail Record (CDR) data were collected, relevant features were selected and various preprocessing techniques such as feature selection, data cleaning, shaping of data frame and feature types were performed.

As a result, J48 algorithm using Cross Validation (CV) options is found to be the best classifier algorithm by scoring 99.3% accuracy followed by the two algorithms highest scores of ANN (CV) and SVM (ST) with 97.51% and 96.0% respectively. This result happens because of J48's capable of learning disjunctive expressions in addition to it reduced error pruning. Pruning decreases the complexity in the final classifier, so that improves predictive accuracy from the decrease of over fitting.

KEYWORDS

Keywords: telecommunications, CDR, fraud detection, ANN, SVM, J48, Machine learning, accuracy, CV , Supplied Test (ST)

ACKNOWLEDGMENTS

First and foremost, I would like to thank my Almighty God " YAHWEH " for all of things happened in my life and for giving me beautiful wife Yodit Gudeta and beloved kids Fraol, Mesgana and Cute lady Maya.

Secondly, the success of this thesis is credited to the extensive support and assistance from my advisor Ephrem Teshale [PhD]. I would like to express my grateful gratitude and sincere appreciation to him for his guidance, valuable advice, constructive comments, encouragement and kindness to me throughout this study. Thank you !

Thirdly, my special thanks go to ethio telecom as a company who gives me a chance to move one step ahead in my educational carrier. Addiitonally special thanks to security staffs Mehari Gebreegziabher and Selamawit Assefa; who helped me in providing necessary information and materials which are crucial in this study.

Finally, I would like to thank all of you who support me to complete this thesis work even if I didn't mention your name here.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AI	Artificial Inteligent
ANN	Artificial Neural Network
AUC	Area Under the Curve
CDR	Call Detail Record
CFCA	Communications Fraud Control Association
CV	Cross Validation
FFNN	Feed Forward Neural Network
FMS	Fraud Management System
IQR	Interquartile Range
ML	Machine Learning
MLP	Multi Layer Perception
MNO	Mobile Network Operators
PRS	Premuim Rate Service
ROC	Receiver Operating Characteristic
RMSE	Root Mean Squared Error
RNN	Recurcisve Neural Network
SIM	Subscriber Identity Module
ST	Supplied Test
SVM	Support Vector Machine
TSP	Telecom Service Provider

1

INTRODUCTION

The development of the telecommunication industry is rapidly increasing with one innovation replaces another in a matter of years, months, and even weeks [1]. Due to the development of affordable technologies, telecommunications are turning the world into a glolbal village. Though, there is enormous growth in terms of the number of subscribers and generating revenue over the past few years, it is highly vulnerable to fraudsters. Fraudsters' methods and techniques are advanced in corresponding to this expansion of the telecommunation industry posing a severe treat to the industry..

Even if the development of telecommunication industry shows a dramatical growth interms of revenue and technological capacity, telecommunication companies often receive substantial damage from customers' fraudulent behaviors which causes loses of the company's image, brand and trust between Telecom Service Provider (TSP) and customer and decreases its profitability.

There are many different definitions of telecommunications fraud. However, there seems to be a general consensus that telecommunications fraud involves the theft of services or deliberate abuse of telecommunication networks [2]. Fraudulent use of the telecom service is with the intention of not paying. Furthermore, it is accepted that the perpetrator's intention is to avoid completely or at least reduce the charges that would legitimately have been charged for the services used.

Telecommunication frauds have different categories on their nature and characteristics. According to Kang and Yang [3] it is categorized into two as subscription and superimposed whereas Becker, Volinsky, and Wilks [4] classifies in to seven groups as superimposed, Subscription, Technical, Internal Fraud, Fraud based on loopholes in technology, Social Engineering and Fraud based on new technology.

One of the common types of fraud is subscription fraud in which usage type is in contradiction with subscription type [5].

Communications Fraud Control Association (CFCA) is the non-profitable organizations which conducts a survey of annual global fraud loss in every two years. Based on the CFCA report [6] telecommunationmunation companies worldwide suffer with Fraudulent illegal activities by losing a huge amount of money specially the one who subscribe a telecommunation service and served without payment which is subscription fraud [7]. According to the report subscription fraud included on the top five ranked lists since 2013. Table 1.0.1 shows how subscription fraud damages telecommunation companies.

Table 1.0.1: CFCA Summary report of fraud losses $Billion [8]

year	Top five total fraud loses	subscription fraud losses alone
2013	35.8	11.3
2015	43.78	6.69
2017	29	6.95

Observing from Table 1.0.1 it is possible to verify the dimension of the "financial hole" generated by fraud in the telecommunications industry and the impact of fraud in the revenue of the telecom operators specifically subscription fraud it has a big impact on telecommunication companies revenue.

1.1 STATEMENT OF THE PROBLEM

Telecomm fraud is a major source of revenue loss for TSP and their customers. Recent (2017) CFCA survey shows that Global Fraud Loss Estimated $29 billion [6]. Experts agree that most telecom providers are losing 3 to 10 percent of their income to fraud. The intention of a subscriber plays a central role in the definition of fraud [9]. Due to subscribers incontradiction use of the services based on their subscription agreements, revenue loss has been occurred to telecom operators [5].

According to a report by CFCA, revenue loss caused by subscription fraud alone is billions of dollars [10] additionally see Table 1.0.1.

Ethio telecom is the only telecommunication company operating in Ethiopia. As part of the telecom service provider, its revenue losses due to fraudsters by the years 2013, 2015 and 2017 is tabulated on Table 1.1.1.

Table 1.1.1: Ethio telecom Revenue Loss [6] [11] [12].

year	Ethio telecom revenue losses in $Million
2013	50
2015	33
2017	89

Ethio telecom provides different types of services to its customers. The two main service types are voice and data. Depending on the customer interest they can use prepaid as well as postpaid payment methods for their service requests. Prepaid means pay as you go and postpaid for credit to a month or above a month. For any service requests ethio telecom registers its customers on the Customer Relation Management (CRM) system for billing purpose. However, it will not have the chance to know these subscribers who are fraudulent or not at the time of service request or applications. Subscription fraud could be done with false identification numbers, stolen credentials and stolen accounts which helps to get customers full information for creating many new accounts for a fraudulent purpose. So, once the subscriber registered and get accounts to access the network they have a chance to start their fraudulent activity.

Ethio telecom by now deploys and use a traditional type of Fraud Management System (FMS) to detect fraudulent activities. Due to the inflexibility rules of this traditional detecting system, fraudsters get a chance to adapt the existing FMS detection rules and policies. Then after, they increase the company's revenue loses as well as damages of subscribers' trust relationship with the company and loses the company's brand. This traditional management system cannot handle the new behavioural change of fraudsters activity.

To overcome such problems there is a methodology or Machine Learning (ML) (data-driven) approach which could give solution. This approach extracts the hidden knowledge's or patterns from subscribers CDR data. This data-driven detection type learns from the history of the subscriber's data for their behavioral change. Such approach gives the capability of detecting fraudaters without predefined rules.

By considering the above problems, this research needs to analyze and answer the following research questions.

Research questions

1. What kind of data features can be used to identify Subscription fraud?
2. What kinds of ML algorithm can be used to detect subscription fraud?

1.2 OBJECTIVES

1.2.1 *General Objective*

The general objective of this thesis is to analyze which ML algorithm perform better for detecting subscription fraud.

1.2.2 *Specific Objectives*

This study has the following specific objectives:

- To select relevant attributes for building Subscription fraud detection model
- To choose appropriate machine learning tools, algorithm and techniques for Subscription fraud detection
- To detect Subscription frauds based on subscribers historical usage behaviors
- To compare the prediction performance of algorithms

1.3 SCOPE AND LIMITATION

There exist more than 200 fraud types in the telecom industry [13]. However, this specific research focused only on subscription fraud type. The row data used is limited to CDR data, which has been callected from ethio telecom. Due to storage limitation, two-month prepaid mobile subscriber data were used.

1.4 SIGNIFICANCE OF THE STUDY

The primary contribution of this thesis is to propose a subscription fraud detection model based upon ML technique. This thesis delivers practical and theoretical contributions which include:

- Discovers uncovered hidden fraudulent behaviour
- Gives awareness and enables telecommunication operator to identify and detect subscription fraudster
- Deliver practical suggestions that may help anti-fraud managers and employees to get a better understanding of subscription fraud.
- Provide insight about the company traditional fraud management system and its limitation.
- Proposes suitable machine learning algorithms for subscription fraud detection.
- Could be a preference for further researches regarding to subscription fraud detection.

1.5 RELATED WORK

In order to have detail understanding of this research topic's about fraud detection and prevention mechanisms specifically subscription fraud detection, related literatures like journals, articles, magazines besides with the Internet were reviewed.

In the past few years, subscription fraud to be the trendiest and the fastest-growing type of fraud [14]. In a similar spirit, Abidogun [15] characterize subscription fraud as being the most significant and prevalent worldwide telecommunications fraud type.

Stevez *et al.* [9] describes the identification of fraudsters on the time of applying to a service request or subscription basis for fixed telecommunication. Two strategies have been proposed for detecting subscription fraud: examining account applications and tracking customer behavior. They use a classification module and a prediction module. The classification module (fuzzy rule) classifies subscribers according to their previous historical behavior into four different categories: subscription fraudulent, otherwise fraudulent, insolvent and normal whereas the prediction module (ANN) allows them to identify potential fraudulent customers at the time of subscription.

Regarding to their experimental test Stevez *et al.* [9] use a database containing 10,000 real subscribers in a major telecom company in chile and 2.2% subscription fraud was detected and a multi-layer perception neural network was implemented for prediction purpose and it identifies 56.2% from true fraudsters and screening only 3.5% of all the subscribers in the test set. Their study was tested and significantly preventing subscription fraud in telecommunications by analyzing the application information at the time of customer application.

Kabari *et al.* [16] present a design and implements of a subscription fraud detection system using Artificial Neural Networks and Neurosolutions for Excel was used to implement the Artificial Neural Network. The study was grounded on customers Internet data usage. The system was trained and tested and 85.7% success rate achieved. The designed system found to be user friendly and effective.

On the other hand, Kabari *et al.* [2] identify the different subscription services provided by the telecommunications sector, identifys the different ways telecommunications fraud is perpetrated and proposes the use of Naïve Bayesian Network technology to detect subscription fraud in the telecommunications sector. The system takes care of the challenges encountered by the rule-based system of detecting fraud. The paper grounded on customers Internet data usage.

Farvaresh and Sepehri [5], describe that one of the common types of fraud is subscription fraud in which usage type is in contradiction with subscription type. The study aimed at identifying customers' subscription fraud by employing data mining techniques and adopting knowledge discovery process based on the leased line telephone services. A hybrid approach consisting of preprocessing, clustering, and classification phases was applied and appropriate tools were employed commensurate to each phase. Specifically, for clustering phase Self Orgamized Map and k-Means were used and in classification phase decision tree (C.45), ANN and SVM as single classifier and bagging, boosting, stacking, majority and consensus voting assemblies were examined. The results showed that SVM among single classifiers and boosted trees among all classifiers have the best performance in terms various metrics. The result is significant both theoretically and practically.

Saravanan *et al.* [14], describe about the identification of true high usage customers from illegitimate customers based on their calling patterns fraud detection. The paper implements a probability-based method for fraud detection in telecommunication sector used Naïve-Bayesian classification to calculate the probability and an adapted version of KL-divergence to identify the fraudulent customers on the basis of subscription. Each user's data corresponds to one record in the database. This paper overcomes the problem of identifying fraudulent customers in telecommunication sector by classifying the true fraudulent customers alone. In other words the paper result indicates that the normal high usage customers has a similar behavioral patterns whereas fraudsters behavioral changes indicating using the service with some time with high usage and disappearing from the system for some time.

All related papers presented have their own roles of detecting subscription fraud depending on the nature of the telecom service provider's customer usage behavior. So that we can see different types of methods and techniques of subscription fraud detection. What the researcher clearly sees is that types of algorithm, feature selection, dataset size limitations are the most significant impact of classification accuracy. However, these papers do not consider aggregated information of the subscriber usage behavior. In this study, by considering subscription fraud usage behaviors the researcher uses feature number of incoming calls, Total number of calls, Distinct calls, ratio of Distinct calls to total calls, Ratio of international call to Total calls.

1.6 METHODOLOGY

In order to achieve the objectives of this thesis and answering the research questions the methods that the researcher follows

1. Review telecom fraud litratures focuse on subscription fraud detection to understand and formulate the problem domain. Additionally, domain expertis were consulted.

2. Collect legitimate and fraudulent subcribers CDR data, attribute selection, preprocessing (i.e data cleaning, integration and aggregation) and creates the dataset for both subscribers to train and testing the model.

3. Since the data were labeled, Supervised ML techniques were chosen. WEKA ML tool were prefered for this specific research. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization[17].

4. The developed model was tested and evaluated using performance measurement parameters (Confusion matrix, Receiver Operating Characteris-

tic (ROC), F-Measure and Accuracy). Conclusively, after comparing models a better efficient model were proposed and recommendation were provided.

1.7 THESIS ORGANIZATION

This thesis paper contains six chapters, objective, research methodology and problem formulations are discussed in chapter 1. Chapter 2 discuses about telecommunication services and its fraud types including subscription fraud theory. Third chapter discussed about machine learning techniques and algorithms which are applied in this thesis. Chapter 4 concerned to where and how the necessary data collected, prepare for experimentation in addition to explaining performance measurement parameters. The fifth chapter relay on classifier classification performance result evaluation and comparison. After all these things are applied the researcher's conclusion and ways of further researches regarding to subscription fraud detection pointed on chapter 6

2

TELECOMMUNICATION SERVICES AND FRAUDS

This chapter discusses telecommunication services and fraud types. The first section describes about the telecom mobile voice services whereas the second section describes telecommunication fraud types related to subscription behaviours.

2.1 TELECOMMUNICATIONS MOBILE SERVICES

Globally telecommunication companies provide different services to their subscribers due to the market computation and winning their customers heart. In a new mobile industry era user subscribe different mobile telecom services via their Subscriber Identity Module (SIM) card numbers for their personal usage. There are two main mobile services provided by TSPs, prepaid and postpaid services.

2.1.1 *Prepaid Mobile Services*

As the name implies in 'Pre-paid' all transaction in this service is pay-as-you-go that means unless you have account balance money you can not dial, send SMS and access the Internet data. In this days fraudsters use this type of services because there is no requirement of guarantee during the service request time to the TSP. Due to lack of proper identification of the subscribers, a single subscriber can have many SIM numbers with fabricated Credentials. The most fraud type which uses this service is Subscription and SIM-Box fraud. With this and other issues, this is the reason subscription fraud challenges today's TSPs globally.

2.1.2 *Postpaid Mobile Services*

It is the most conservative service offered by telecom operators. 'Post-paid' implies that credit facilities are given for services used for some period usually between 1 to 6 months. Due to the credit period time duration and lack of proper identification of the subscribers TSP provides this service with some guarantee's in case the subscriber defaults. Other services provided by ethio telecom are susceptible to fraud includes: Roaming Services, Value Added Services and Premuim Rate Service (PRS).

2.2 TELECOMMUNIACTION FRAUD

The telecommunication industry has expanded dramatically in the last few years with the development of affordable mobile phone technology [18]. With the increasing number of mobile phone subscribers, global mobile phone fraud is also set to rise. Telecommunication fraud is defined as unauthorized use, tampering or manipulation of a mobile phone or service[19]. The problem with telecommunication fraud is the huge loss of revenue and it can affect the credibility and performance of telecommunication companies. Telecommunication fraud which attracts particularly to fraudsters as calling from the mobile terminal is not bound to a physical location and it is easy to get a subscription. This provides a means for illegal high-profit business for fraudsters requiring minimal investment and relatively low risk of getting caught due to possibility of subscribing with fabricated identity.

2.3 COMMON TYPES OF TELECOMMUNICATION FRAUD

In this day there is a number of fraud types in telecommunication sector which causes huge amount of dollars lost every year. Different researchers categorize fraud types into different manners. Based on Shawe-Taylor *et al.* [20], fraud cate-

gorized to six these are subscription fraud, PABX fraud, handset theft, premium rate fraud, free phone call fraud and roaming fraud. Hilas and Mastorocostas, [21] categorized fraud in to four technical fraud, contractual fraud, hacking fraud, and procedural fraud. The third Scholars Kang and Yang [3] categories fraud types in to two subscription and superimposed frauds.

In recent year report (2017), the top five fraud were listed as Subscription Fraud (Identity), PBX Hacking, IP PBX Hacking, Subscription Fraud (Application) and Internal Fraud/Employee theft as shown in Figure 2.3.1.

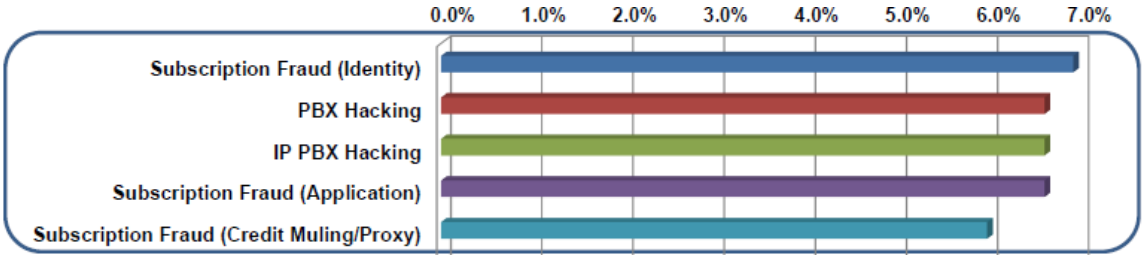


Figure 2.3.1: CFCA 2017 report [10]

By the coming sub sections, the researcher discuss some fraud types which are subscription fraud properties Superimposed fraud, SIM swapping, SIM cloning, SIM-Box and Roaming fraud.

2.3.1 *Subscription Fraud*

Subscription fraud is characterised by a fraudster using own, stolen or fabricated identity to get services with no intention to pay. Subscription fraud recognized as the most damaging of all non-technical fraud types. It is usually the pre cursor to other types of fraud such as Premium Rate Fraud, International Revenue Share fraud, SIM-Box fraud and Roaming fraud which are lethal in their own rights [22]. The real impact of this type of fraud is difficult to measure because it does not stop with revenue loss alone. The effects can be catastrophic in terms of escalating complaints, poor customer experience and dissatisfaction among support staff [23].

Subscription fraud is a contractual fraud [16]. Subscription fraud is the most common since with a stolen or manufactured identity, there is no need for a fraudster to undertake a digital network's encryption or authentication systems. Their preferred methods are using low techniques which is using the network under the threshold level of FMS. This has less chance of detection.

Regarding to Koi-Acroti *et al.* [8], in this day subscription fraud to be the trendiest and the fastest-growing type of fraud. In similar spirit, characterizes subscription fraud as being probably the most significant and prevalent worldwide telecommunications fraud type. In subscription fraud, a fraudster obtains a subscription (possibly with false identification) and starts a fraudulent activity with no intention to pay the bill.

2.3.1.1 *subscription fraud call properties*

The following list of characterstics describes the behaviour of subscription frauds:

- Fraudsters generate large amount of outgoing calls
- The majority destination of fraudulent subscribers' calls is International calls
- After repeatedly calling fraudulent subscriber disappears from the network
- There could be a small number of incoming calls
- Fraudulent subscribers also sending international SMS as well as using high amount of Internet data
- Use High duration per calls specially International calls

2.3.1.2 *Subscription Fraud scenario*

Subscription fraud is characterized by a fraudster using own, stolen or fabricated identity to get services with no intention to pay. Subscription fraud remains one of the top ranked types of fraud and is widespread across all operations. Subscription Fraudsters employing schemes such as identity theft and use of fabricated, stolen details identities information [14].

Subscription fraud involves the fraudulent individual obtaining the customer information required for signing up for telecommunication service with authorization. The usage of the service creates a payment obligation for the real or normal customer [2].

Fraudsters obtain an account without intention to pay the bill. In such cases, abnormal usage occurs throughout the active period of the account. The account is usually used for call selling or intensive self-usage. Cases of bad debt, where customers who do not necessarily have fraudulent intentions never pay a single bill, also fall into this category. These cases, while not always considered as "fraud", are also interesting and should be identified.

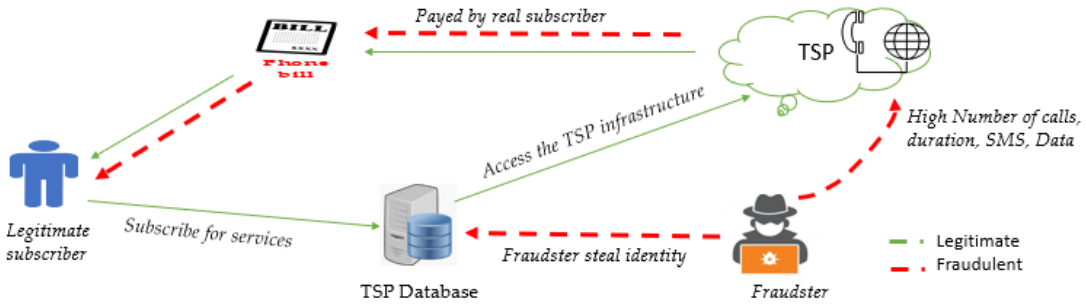


Figure 2.3.2: Subscription Fraud Scenario [2] [14]

In this scenario case as shown in Figure 2.3.2 the normal user subscribe a service from TSP and signed with the service provider and accessing the TSP infrastructure and pay depends on their usage. However, the fraudulent steals normal users identity, cloning the SIM or stolen phones then after accessing the TSP's infrastructure as much as they want without the intention to pay but billing will be handled by normal customers.

2.3.2 *Superimposed Fraud*

Superimposed fraud is the most common fraud scenario in private networks. This is the case of an employee, the fraudster, who uses another employee's authorization code to access outgoing trunks and costly services [24]. Unlike subscription

fraud, a legitimate account will be used in superimposed fraud. Mobile phone cloning and obtaining calling card authorization are among the several ways to carry out this fraud type. Fraudsters make use of the legitimate account for an illegitimate use by different means. In such cases, abnormal usage is observed, and it is somewhat challenging to detect. Such kind of frauds come into knowledge when the authorized users complaint about the excessive billing. This fraud can be overcome by building a robust system to check the authenticity of users.

2.3.3 *SIM swapping*

A SIM swapping attack works by convincing call center representatives working for a mobile phone provider to port a phone number to a new device. If they do that, they will innocently transfer control of the victim's phone number to the attacker. A SIM swap can be considerably easier when there is a collaborative insider to leverage. With someone working for the mobile carrier, an attacker doesn't even need to carry out a social engineering ruse to gather the necessary information about the victim. It has become increasingly popular for cybercriminals to recruit mobile phone provider employees through social media accounts to scale their SIM swapping attacks. By posing as company hiring for open positions through these accounts, attackers have an opportunity to engage insiders through the promise of monetary gain [25].

2.3.4 *SIM Cloning*

SIM cloning has the same goal as SIM swapping, but cloning does not require calling the mobile carrier. Rather, it is more about technical sophistication. The cloning attack uses smart card copying software to carry out the actual duplication of the SIM card, thereby enabling access to the victim's international mobile subscriber identity (IMSI) and master encryption key. Since the information is seared onto the SIM card, physical access to it is a requirement. That means taking the SIM card

out of the mobile device and placing it into a card reader that can be attached to a computer where the duplication software is installed.

After the initial stealthy SIM replication takes place, the attacker inserts that SIM into a device they control. Next, the victim has to be contacted. The trick may begin with a seemingly innocuous text message to the victim asking them to restart their phone within a given period of time. Then, once the phone is powered off, the attacker starts their own phone before the victim restarts and, in doing so, initiates a successful clone followed by an account takeover. Once the victim restarts their phone, the attack is complete, and the attacker will have successfully taken over the victim's SIM and phone number. Then the legal phone user then gets billed for the cloned phone's calls. Cloning mobile phones is achieved by cloning the SIM card contained within, not necessarily any of the phone's internal data [25] .

2.3.5 *SIM-BOX*

A SIM box fraud is a setup in which fraudsters install SIM boxes with multiple low-cost prepaid SIM cards most of the time all these SIM cards will be subscribed with forged credentials. The fraudster then can terminate international calls through local phone numbers in the respective country to make it appear as if the call is a local call. This allows the box operator to bypass international rates to fraudulently below the prices charged by Mobile Network Operators (MNO) and evade the tax charged by the government. This act denies telecommunications and government from benefiting from international phone calls. Besides the loss of revenue, SIM Box operators cause degradation of call quality which prevents them from meeting service level agreements. The fraudster will pay the network for a national call but will charge the Wholesale operator for every minute he terminated; the Network Operator loses the Interconnection fee [26].

2.3.6 *Roaming*

According to Maciá-Fernández [27] Subscription fraud is one of the fraudster's preferred methods for digital roaming fraud. Due to the delay in the home provider receiving roamer call data can be anywhere from one to several days, Stealing mobile phones belonging to roamers, usually in vacation destinations. It is the ability to use telecom services like voice or data services outside the home network with no intention to pay for it. In these cases, fraudsters use the longer time-frames required for the home network. Roaming fraud can starts as an internal or subscription fraud in the home network when obtained SIM cards are sent to a foreign network.

3

MACHINE LEARNING

Machine learning is an application of Artificial Inteligent (AI) that provides systems with the ability to automatically learn and improve from experience without being explicitly programmed [28]. ML provides smart alternatives to analyzing vast volumes of data. By developing fast and efficient algorithms and data-driven models for real-time processing of data. It can produce a better accuracy results and data analysis.

In this day ML categorized in to four namely unsupervised, semi-supervised , supervised and reinforcement. The upcoming sections describe about these ML categories. The first, second and third sections describe unsupervised, reinforcement and semi-supervised techniques conducted respectively. But, due to the labelled data used in this study, supervised ML techniques with the proposed algorithms will be described in the last section.

3.1 UNSUPERVISED LEARNING

Unsupervised learning involves the analysis of unlabeled data under assumptions about structural properties of the data. unsupervised learning is the partitioning or segmentation of the data in to groups or clusters. In unsupervised or undirected models there is no output field, just inputs. The pattern recognition is undirected; it is not guided by a specific target attribute. The goal of such technique is to uncover data patterns in the set of input fields [29].

3.2 SEMI-SUPERVISED LEARNING

Semi-supervised algorithms require a combination of labelled and unlabeled data. Semi-supervised learning is a class of machine learning tasks and techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning and supervised learning.

3.3 REINFORCEMENT LEARNING

Reinforcement learning algorithms use agents how to behave the environment in other words instead of training examples that indicate the correct output for a given input, the training data in reinforcement learning are assumed to provide only an indication as to whether an action is with or not.

Reinforcement learning is a behavioural learning model. The algorithm receives feedback from the data analysis, guiding the user to the best outcome. Reinforcement learning differs from other types of supervised learning because the system isn't trained with the sample data set. Rather, the system learns through trial and error. Therefore, a sequence of successful decisions will result in the process being reinforced.

3.4 SUPERVISED LEARNING

Supervised machine learning algorithms need labelled datasets for learning and classifying the dataset. In supervised algorithms, the goal is to predict an event or estimate the values of a continuous numeric attribute. In supervised algorithms, there are input fields or attributes and an output or target field. Input fields are also called predictors because they are used by the model to identify a prediction function for the output field. The predictor needs to be labelled or previously identified data which needs to train the algorithm. After learning is accomplished

the outcome tells to the researcher that how well the algorithm predicts the new input class or instances.

Supervised learning by itself can be categorized into Regression and Classification. Under Supervised machine learning, there are algorithms categorizes SVM, Decision Trees, Neural Network, Naïve Bayes and Nearest Neighbor algorithm.

3.4.1 *Regression*

Regression analysis is a subfield of supervised machine learning. It aims to model the relationship between a certain number of features and a continuous target variable. In regression problems, the output comes up with a quantitative answer, with predicting the prices of a house or the number of seconds that someone will spend watching a video.

3.4.2 *Classification Techniques*

Classification is the process of building a model of classes from a set of records that contain class labels and used to classify the item according to the features with respect to the predefined set of classes datasets [30]. In today's scientific world ML algorithms are known to effectively classify complex datasets of two and multi-class datasets.

Classifier: An algorithm that maps the input data to a specific category

The objective of this study is to compare and select the best classifier algorithms based on their classification performance measures. Following this objective, the researcher selects three supervised ML algorithms J48 decision tree, ANN and SVM.

The next subsections will explore these classification algorithms in detail.

3.4.2.1 *J48 - Decision trees*

Decision Trees are a non-parametric supervised learning method used for classification and regression [31]. A decision tree follows a top-down approach to split data recursively into smaller mutually exclusive subsets that include a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. The learning process involves finding the best variables for the partition at each split. Decision tree works by creating a tree to evaluate an instance of data, start at the root of the tree and moving down to the leaves (roots) until a prediction can be made.

Based on patit *et al.* [32], J48 classifier is a simple C4.5 decision tree for classification. J48 examine the normalized information gain that actually the outcomes splitting the data by choosing an attribute randomly from the given dataset. In order to make the decision, the attribute utmost standardized information gain is used (equation 3.2). The splitting methods stop if a subset belongs to the same class in all the instances. J48 constructs a decision node using the expected values of the class.

Sequences of steps that J48 algorithm performs to accomplish its classification:

Step 1: The leaf is labelled with the same class if the instances belong to the same class.

Step 2: For every attribute, the potential information will be calculated and the gain in information will be taken from the test on the attribute with equation 3.1 and 3.2 .

Step 3: Finally the best attribute will be selected for root.

To determine root attributes and decision tree size, Entropy and information gain are statistical measures which are used to construct tree of the algorithm. Attributes which has the highest information gain value will be the root or decition nodes of the tree but nodes which have an entropy of zero are considered to be

leaf node while nodes with entropy greater than zero will further split until the entropy is zero.

So that classifying with J48 Algorithm could benefite:

- Can handle lost or missing attribute values of the data
- Overcome Over-fitting
- Reduced error pruning
- differing attribute costs Subtree raising with different confidence
- Pre-pruning or Post-pruning
- precision can be increased by pruning
- Both the discrete and continuous attributes are handled by this algorithm. A threshold value is decided by C4.5 for handling continuous attributes. This value divides the data list into those who have their attribute value below the threshold and those having more than or equal to it.

Counting Gain

This process uses the "Entropy" which is a measure of the data disorder. The Entropy of the attributes is calculated by:

$$Entropy(j|y) = \frac{|yi|}{|y|} \log \frac{|yi|}{|y|} \qquad H(X) = - \sum_{i=1}^{c} \frac{|yi|}{|y|} \log \frac{|yi|}{|y|} \tag{3.1}$$

where:

y attributes
H(X) Entropy of the dataset

Entropy controls how a Decision Tree decides to split the data. It actually effects how a Decision Tree draws its boundaries.

$$Gain(y, j) = Entropy(y - Entropy(j|y)) \tag{3.2}$$

Based on equation 3.1 and 3.2 the size of the tree, root and leaf will be determined.

Pruning

Pruning is a technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances [33]. Pruning is performed for decreasing classification errors which are being produced by specialization in the training set. Additionally, it reduces the complexity of the final classifier which results improving predictive accuracy and reduction of overfitting.

3.4.2.2 *Artificial Neural Network - ANN*

Classification is one of the most active research and application areas of neural networks. An ANN is a computational model that is inspired by the way biological neural networks in the human brain process information. Artificial neural networks are composed of nodes called neurons or processing elements, which are connected together to form a network of nodes. Neural networks are physically cellular systems which can acquire, store and utilize experimental knowledge.

ANN architecture generally consists of input layer, output layer and hidden layer(s). Each contact between these nodes has a set of values called weights which contribute to the determination of the values resulting from each processing element based on the input values of that element. ANNs can learn from their Architecture of the network through an iterative process by adjustments of its synaptic weight and bias level. It can be used to model a complex relationship between inputs and outputs which can find patterns in data.

There are two major classes of ANN Feed Forward Neural Network (FFNN) and Recurcisve Neural Network (RNN). FFNN is the first and simplest type of artificial neural network. It contains multiple neurons arranged in layers. In a FFNN, the information moves in only one direction which is forward from the input nodes, through the hidden nodes (if any) and to the output nodes. There are no cycles or loops in the network. Single layer and Multi Layer Perception (MLP) are the two

examples of FFNN. RNN on the other hand are dynamical networks with recurring path of synaptic connections serve as controlling time-dependent problems.

ANN used for supervised and unsupervised techniques based on its learning paradigm. In the case of supervised learning it uses associative learning which requered a training inputes data but in the case of unsupervised neural network, it only requires input patterns from which it develops its own representation of the input stimuli.

Perceptron

The most simple neural network unit is called "Perceptron"[33]. Perceptron has just two layers, input layers and output layers. Often called a single-layer network on account of having 1 layer of links, between input and output. Input nodes are connected fully to a node or multiple nodes in the next layer. A node in the next layer takes a weighted sum of all its inputs.

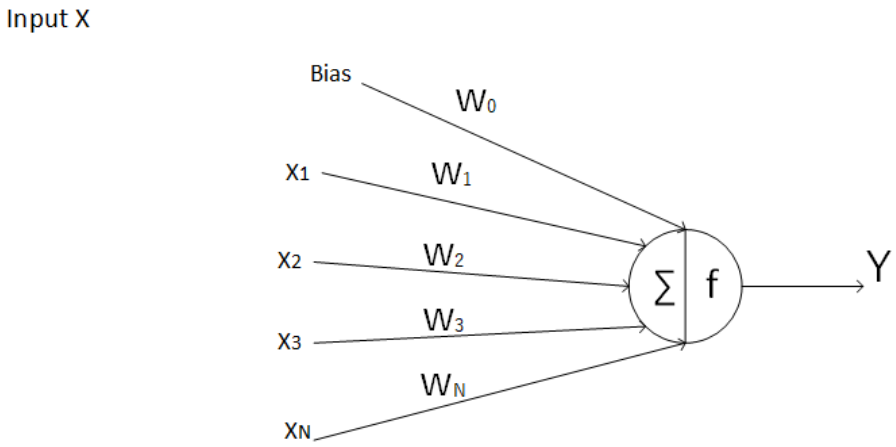


Figure 3.4.1: Perceptron

where,

x_n, feature inputs to the network are represented by the mathematical symbol

W_n, Each of these inputs are multiplied by a connection weightY, generatetheoutput.

The outputs of all neurons in the hidden layer are calculated by the summation function Y (see Equation (3.3)).

$$Y = f(\sum_{n=1}^{N}(w_1x_1 + w_2x_2 + ... + w_Nx_N + w_0) \tag{3.3}$$

where:
Y generate the output.

f function (sigmoid)

x inpute (attributes)

w weight of bias

These products are simply summed and fed through the transfer function with equation (3.4)

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.4}$$

where:
f(x) generate the output

Multi-Layer Perceptron (MLP)

MLP contains one or more hidden layers (apart from one input and one output layer) as shown in figure 3.4.2. While a single layer perceptron can only learn linear functions but a multi-layer perceptron can both learn linear and non – linear functions. A good point of MLPs is their applicability to any field of pattern recognition tasks of supervised learning [34]. MLP can solve problems which are not linearly separable. MLP is often applied to supervised learning problems. It is a FFNN that generates a set of outputs from a set of inputs. MLP is a neural network connecting multiple layers in a directed graph goes in one direction only. MLPs are widely used for pattern recognition, classification, prediction, and approximation, optimization, control, time series modelling and data mining.

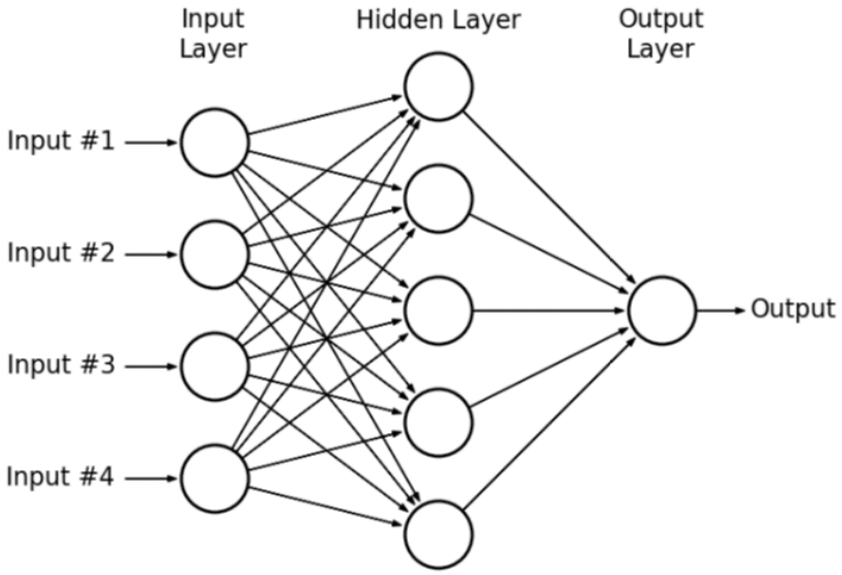


Figure 3.4.2: Multi-Layer Perceptron network diagram

MLPs use backpropagation algorithm to train connection weights between layers. The backpropagation algorithm relies on gradient descent and computing time can be long[34]. For the Backpropagation case, it is used for adjusting the weights by comparing the output of the feed-forward neural network against the desired output. Training a neural network is the process of setting the best weight on the edges connecting all the units in the network.

Classifying with ANN features has desirable characteristics like high accuracy, noise tolerance, independence from prior assumption, ease of maintenance, overcoming the drawbacks of other statistical methods, ability to be implemented in parallel hardware, minimized human intervention (highly automated) and suitability to be implemented in non-conservative domain are major ones. However ANN has limitations like poor transparency, trial and error design, data hungriness (requires large amount of data), over fitting, lack of explicit set of rules to select a suitable neural network, dependency on the quality and amount of data, lack of classical statistical properties (confidence interval and hypothesis testing) and the techniques are still evolving (not robust).

3.4.2.3 *Support Vector Machine - SVM*

Support vector machine is a supervised classifier which has been proved highly effective in solving a wide range of pattern recognition. SVMs are a new technique suitable for binary and multi-class classification tasks in addition to a new promising non-linear, non-parametric classification technique [35]. SVM is state of the art classification and regression algorithm and optimization procedure maximize predictive accuracy while automatically avoiding over-fitting the training data. However, they suffer from the important shortcomings of their high time and memory training complexities[36]. The SVM classifier uses only a small subset of the total training set for classification, thus reducing the computational complexities. The reduction is achieved by the use of kernel trick and overfitting of data is avoided by classifying with a maximum margin [37].

Initially, the hyperplane randomly classifies the classes by making a line or hyperplanes. Input vectors that just touch the boundary of the margin H1 and H2 – as shown in figure 3.4.3 are called support vectors. In this approach, the training stage is used to delimit the region (boundary) where data is classified as normal. In the testing stage, the instances are compared to that region, if they fall in the delimited region they are classified as normal, if not as fraudulent. The main objective is to minimize the number of points within the margin as much as possible.

$$z(x) = w_1.x_{j1} + w_2.x_{j2} + w_3.x_{j3} + w_4.x_{j4} + ... + w_n.x_{jn} + b \tag{3.5}$$

In a compact form:

$z(x) = x_j^T w_j + b$

where:

z(x) Linearly discriminant function

x feature vector selected for classification

w space weight of the hyperplane

b space bias controlling hyperplane positioning

Maximizing the margin

Maximum perpendicular distance between the nearest data point and hyperplane - Margin SVM algorithms find the function (hyperplane) that returns the largest minimum distance to the data points. This distance is called a margin, and the data closest to the margins are then termed support vectors. In figure 3.4.3 the points laid 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, the remaining data are not important in developing the model.

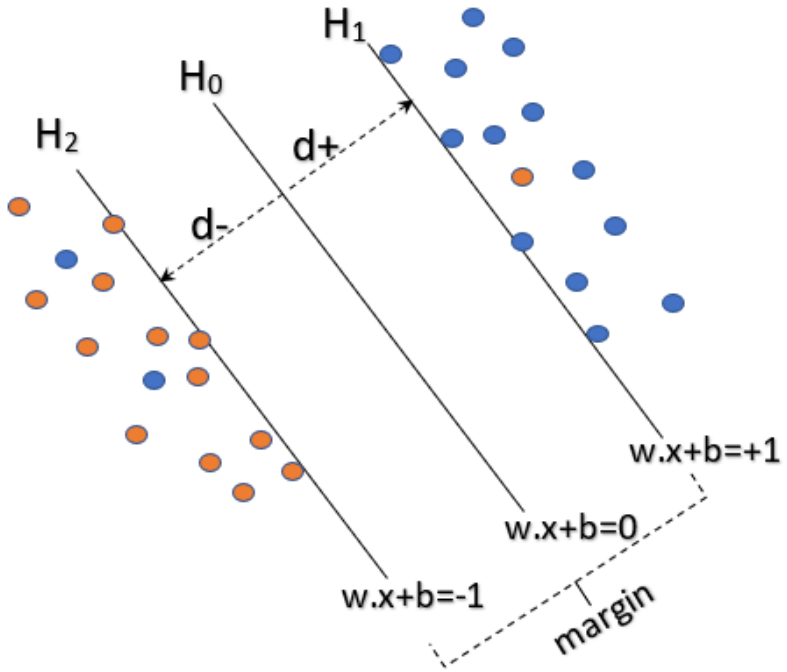


Figure 3.4.3: SVM Classification According to Berwick [38]

$$H_1 = wx_i + b \geqslant +1 \quad \text{when } H_i > +1 \tag{3.6}$$

$$H_2 = wx_i + b \leqslant -1 \quad \text{when } H_i = -1 \tag{3.7}$$

The points on the planes H1 and H2 are the tips of the Support Vectors The plane Ho is the median in between, where w $x_i + b = 0$

d+ = the shortest distance to the closest positive point

d- = the shortest distance to the closest negative point

The distance between the median hyperplane H_0 *and hyperplane* H_1 *is then* :

$$d+ = \frac{|wx + b|}{||w||} = \frac{1}{||w||} \tag{3.8}$$

and also on the other side the distance between the median and hyperplane H_2*is* :

$$d- = \frac{|wx + b|}{||w||} = \frac{1}{||w||} \tag{3.9}$$

The margin (gutter) of a separating hyperplane is

$$d+ \text{ + } d- \tag{3.10}$$

In other words from equation 3.8 and 3.9 we can compute the total margin distance between H1 and H2 is thus:

$$margin = (d+ \text{ - } d-).\frac{w}{||w||} = \frac{(1-b)+(1+b)}{||w||} = \frac{2}{||w||} \tag{3.11}$$

Hence, Maximizes the margin while minimising some measure of loss on the training data. The optimization problem for the calculation of w and b can thus be expressed by:

$$min_w \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} E_i \tag{3.12}$$

where:

E_i *is errors*

C "capacity" is a tuning parameter

In the first part of equation (3.12) C is a tuning parameter, which weights in-sample classification errors and thus controls the generalisation ability of an SVM.

The higher is C, the higher is the weight given to in-sample misclassifications, the lower is the generalization of the machine. Low generalisation means that the machine may work well on the training set but would perform miserably on a new sample.

4

DATA PREPARATION

To achieve the objective of this study, the researcher constructs a methodology or experimental processes which includes data collection (Raw CDR Data), preprocessing and evaluation of the algorithm as shown in figure 4.0.1.

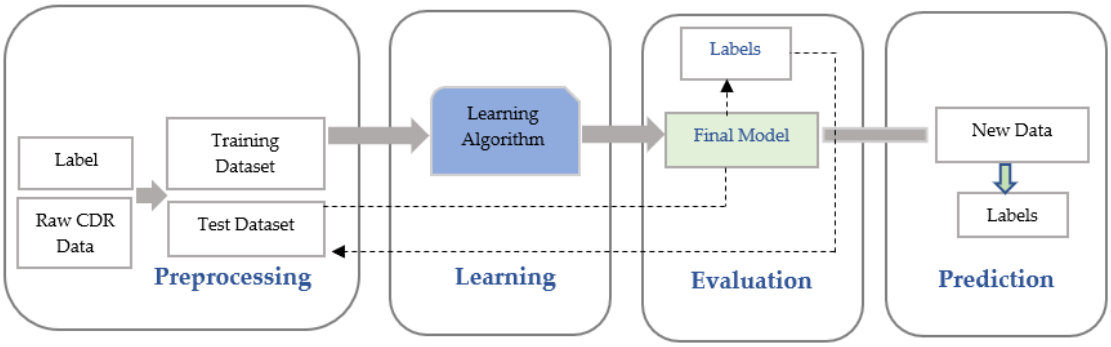


Figure 4.0.1: System Model

Under the upcoming sections details of how the data were collected, understanding the collected data, selecting a relevant attribute, preprocessing such as data clearing, data integration and aggregation tasks will be conducted.

4.1 DATA COLLECTION

The main objective of this study is subscription fraud detection grounded on subscribers usage call patterns, CDR. A two month period from May 25 to July 25, 2019 were collected.

For collecting and processing purpose windows server 2012 R2 standards V.6.3.9600 with 8GB RAM and quad-core Processor, 2.5 TB storage mounted on windows server and Oracle 11th generation version database was deployed on the same

server. The CDR data collected and stored every day to the server in text format as shown in Figure 4.1.1 and it has 33 attributes. Each day an average of 29 million subscribers accessing the network and about 175 million calls recorded which requests 30Gb space to store for a single day activity.

```
20891769667|1|93xxxxx68|1|93xxxxx68|25193xxxx432||636013088261406||20190624141408|20190624141508|60|5003|251|1|1|636011700432401||20190624141512|101530120057433089
20891769668|1|91xxxxx31|1|91xxxxx31|25197xxxx732||636013094710676||20190624141447|20190624141508|40|3335|251|1|1|636010110232554||20190624141512|100190130067844703
20891769669|1|91xxxxx52|1|91xxxxx52|25194xxxx351||636019925350994||20190624141256|20190624141509|140|11673|251|1|1|636012116114884||20190624141512|1004701500542230
20891769670|1|91xxxxx18|1|91xxxxx18|25191xxxx388||636019912566900||20190624141459|20190624141508|20|1668|251|1|1|636011401010035||20190624141512|1004404000034160529
20891769671|1|94xxxxx79|1|94xxxxx79|25191xxxx573||636019939411438||20190624141200|20190624141509|200|16675|251|1|1|636011100413012||20190624141512|1635204001577844
20891769672|1|92xxxxx05|1|92xxxxx05|25193xxxx125||636019926394557||20190624141405|20190624141508|66|0|251|1|1|636010110430129||20190624141512|100960140075982457|20
20891769673|1|99xxxxx74|1|99xxxxx74|25198xxxx035||636019927797682||20190624141353|20190624141509|80|6670|251|1|1|636011500411363||20190624141512|100120400173617465
20891769674|1|96xxxxx60|1|96xxxxx60|25196xxxx779||636013062448843||20190624141413|20190624141509|60|5003|251|1|1|636011600510747||20190624141512|146200400012771788
20891769675|1|96xxxxx30|1|96xxxxx30|25191xxxx639||636013066747605||20190624141350|20190624141509|80|6670|251|1|1|636011700912052||20190624141512|152710400042287435
20891769678|1|99xxxxx62|1|99xxxxx62|25193xxxx526||636019928411588||20190624141449|20190624141510|40|3335|251|1|1|636010170230491||20190624141513|177990400190823486
20891769680|1|91xxxxx73|1|91xxxxx73|25192xxxx156||636019926784417||20190624141418|20190624141510|60|0|251|1|1|636011101914063||20190624141513|100270120059101088|20
20891770121|1|96xxxxx60|1|96xxxxx60|25192xxxx864||636013062811063||20190624141401|20190624141510|80|6670|251|1|1|636010130130115||20190624141513|146560400018063740
20891770123|1|93xxxxx28|1|93xxxxx28|25193xxxx880||636013025725953||20190624141440|20190624141510|40|3335|251|1|1|636011400814003||20190624141513|101260110052984522
20891770124|1|92xxxxx95|1|92xxxxx95|25191xxxx084||636019926394845||20190624141424|20190624141510|60|5003|251|1|1|636011102310987||20190624141513|100710140005177823
20891770125|1|92xxxxx96|1|92xxxxx96|25193xxxx927||636013015548358||20190624141450|20190624141510|30|0|251|1|1|636011200212907||20190624141513|100740140005775132|20
```

Figure 4.1.1: Dumped CDR

Due to space limitation seven tables were created, one table which is a source table for inserting or uploading the original dumped CDR data as it is and six table (three for fraudulent and three for legitimate) created based on the type of the service SMS table, data table and voice table with selected attributes. These daily collected text format data uploaded to the source table with batch files (see in appendix A.2) and from the source table the selected attributes based on the services inserted to the three tables with another batch files (see in appendix A.3). Source table has similar attributes as collected text format CDR data. After inserting these data to their related tables the original dumped text format CDR data were deleted (daily) and the source table truncated.

4.2 UNDERSTANDING CDR DATA

This thesis work grounded on ethio telecom prepaid mobile CDR data. So, the CDR data is collected to the database. Before going through preprocessing and further tasks the researcher works together with domain experts to understand and evaluate the data with their respective attribute values to the problem domain, it is a key task. Additionally, verifying usefulness of the data, completeness, redundancy, missing values, and reasonableness of attribute values with respect to the ML goals.

4.3 DATA SELECTION

All attributes from the collected CDR data cannot be used for this study. Based on Kamel [39], before starting to select attributes we need to identify the parameters that could represent and lead to hitting the objectives of study. The expect output from the classification algorithms fully depend on the quality of the selected input data.

A two months prepaid mobile CDR data from MAY 25 to July 25 2019 were collected. After CDR data collected and understanding the value of data, the researcher selects relevant attributes which could identify the behaviors' of subscription fraud (see Section 2.3.1.1). Irrelevant data to this research objective removed from the collected CDR data in order to learn relevant information to the machine learning algorithm. In this specific study, subscription fraud behavior were used as an input to this data selection task. In this data selection task domain expert advise has been a key role. Reducing unrelated attributes will improve the training time, algorithms performance and reduce complexity of the algorithm task.

4.3.1 *Attribute Selection*

In many supervised learning problems attribute selection is important for a variety of reasons: generalization performance, running time requirements, constraints and interpretational issues imposed by the problem itself [40]. Feature selection is important to reducing the number of attributes which helps not only speed up the learning process but also prevents the learning algorithms from getting misled into generating an inferior model due to the presence of irrelevant attributes.

For this specific study, nine out of 33 attributes as shown in Table 4.3.1 were selected. The selected attributes identifies the behaviour of subscribers usage. The remaining attributes which are irrelevant to the objective of this study eliminated.

Some attributes are redundant or having the same value such as charge_fee and Call_Fee, calling_number and service_number, third_party and called_number.

Table 4.3.1: Original Selected Attributes

No	Attribute	Reason
1	CALLING_NUM	For identifying the subscriber who initiating the call
2	CALLED_NUM	To identify the destination number
3	Call_Type	To identify the call destination (Local or International)
4	START_TIME	To describe the calling start time
5	DURATION	duration of a call spent
6	CALL_FEE	How much money is spent per call
7	DATA_USAGE	To describe how much data the subscriber uses
8	SMS	describe how much SMS the subscriber sent/reciev
9	SERVICE_TYPE	To identify the service type

This study is mainly depend on prepaid mobile subscribers usage pattern and the features are differentiated based on combination of the following three category inaddition to subscription fraud behaviour:

- *Call Indicator:* Indicator of calls is either a local call or international call or destination of calls.

  CALLED_NUM, Service_type, call_type

- *Time Dimension:* Usage of call is captured timely.

  CALLING_NUM, START_TIME, DURATION

- *Usage category:* number of calls made (call frequency), call charge and duration of calls made.

  CALLING_NUM, CALL_FEE, Data_usage, SMS, DURATION, Service_Type

4.3.2 *Sampling*

The goal of sampling is to learn via sampling from the statistical data in order to estimate the characteristics of the subscription fraud. ML algorithms adaptively improve their performance as the number of samples available for learning increases [41]. But, Using the entire two months collected CDR data is impossible.

A small number of fraudulent were identified. So, we need to make some proportionality between normal and fraudulent numbers. Depending on the problem domain area the number of proportionalities between the classified parties could be different. But as discussed with domain experts and scholars research related to this thesis Farvaresh and Sepehr [5] and Estévez *et al.* [9] uses the number of different sample sizes based on their problem domains. We decided to use 25% fraudulent and 75% legitimate users to create a dataset.

A total of about 29 million subscriber data were stored in the database and out of these subscribers 15,000 legitimate subscribers has been selected using a simple random sampling technique. In this technique, each instance in the database has an equal chance of being selected as a subject. The entire process of sampling is done in a single step with each subject selected independently of the other instances of the database. In addition to the legitimate subscriber 5,000 high usage (subscription) fraudulent subscriber were identified with the help of domain experts has been used for further experiment and analysis purpose. A total of 20,000 subscribers used as a sample in this study. Table 4.3.2 shows the sampled data with their number of records.

Table 4.3.2: Sampled data records

	Fraud	Normal	Total
Subscriber	5,000	15,000	20,000
Record	2,989,540	6,475,503	9,463,043

4.4 DATA PREPROCESSING

Real-world data is often incomplete, inconsistent, noisy and lacking in certain behaviours or trends and is likely to contain many errors. Well performed preprocessing steps are important in order to increase the classification performance and adequately analyse the result. Preprocessing tasks had been done on the database like clears null values, removing noisy data (some wild characters), merging tables, attribute aggregation and integrations of tables performed (see Appendices Section A.4).

Under this subsection preprocessing stage data cleaning, data integration and data aggregation will be conducted. For this study there are number of oracle database scripts used in addition to weka preprocessing features.

4.4.1 *Data Cleaning*

Data cleaning is a method for fixing missing values, outliers, and possible inconsistent data. Missing data is common [42]. Data cleansing is a process in which we go through all the data within the database. The presence of missing, irrelevant and noisy values in a dataset can affect the performance and accuracy of a classifier constructed using that dataset as a training sample.

The collected CDR data have a null value, incomplete values, missing value and duplicate records. As the objective of this study is grounded on prepaid mobile, the researcher selects only those calling numbers from the collected CDR data. But records different from prepaid is discarded. The collected CDR has 33 columns, however, duplicated and null valued attributes removed. Records having missing values, outliers and incomplete were discarded. In this study, we found less than 1% missed value (77,332) from a total records of data (9,463,043) however based on Acuna and C. Rodriguez [42] rates of less than 1 % missing data are generally considered trivial. However, it was removed.

On the other hand, an outlier is unusual variables of value that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. Its appears could be at the maximum or minimum of the variable that distorts the distribution of the data [43]. It is therefore important to identify them prior to modelling and analysis. Outlier detection algorithms evaluate instances based on distance, density, projections, or distributions. But the most common data distribution measures is an Interquartile Range (IQR) for identifying outliers.

IQR describes the middle 50% of values when ordered the data from lowest to highest. To find the interquartile range (IQR), first, find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3) respectivelly.

$$IQR = Q_3 - Q_1 \tag{4.1}$$

The interquartile range is a value that is the difference between the upper quartile value (Q_3) and the lower quartile value (Q_1).

Equation 4.1 result tells how spread out the "middle" values is, it can also be used to tell when some of the other values are "too far" from the central value. These "too far away" points are called "outliers" because they "lie outside" the range .

In this study, researcher applies IQR techniques to identify and remove the outliers for reasonable and acceptable classification results on weka tool platform. A total of 64,202 records were detected and removed (below -3IQR or above +3IQR)

4.4.2 *Data Integration*

Data integration is the first step toward transforming data into meaningful and valuable information of the subscribers. In this study, data were integrated from multiple sources to have a single view of the overall sources. The researcher already have six different tables (voice, Data and SMS) for fraudulent and legitimate user. Attributes from these tables need to be integrated to form a subscriber level. The collected CDR data were inserted to these three tables daily basis for two

month. These integration processes were done for both fraudulent and legitimate subscribers.

4.4.3 *Data Aggregation*

Data aggregation is a process in which information is gathered and expressed in a summary form. In this research aggregation were done based on subscriber's usage patterns of call duration, call frequency, call fee, usage data and Total Incoming call. This accumulated data behavior of a subscriber gives a better fraud detection capability. Data is aggregated in a daily span of time based on the selected attributes from the collected CDR data (see Table 4.3.1).

Aggregated output of SMS, voice call and Internet data are integrated to form single instance per subscriber level. Moreover, a class label field that identifies the subscriber type added for training purpose. These aggregated attributes are described on Table 4.4.1.

Table 4.4.1: Aggregated and derived features description

Attribute	Description
TOT_CALLS	Subscribers total number of calls
DIST_CALLS	Number of Unique called numbers
OUT_CALL_DURATION	Total out going calls duration
RATIO_DISTCALL_TOTAL	Ration of Dist_Call & Total_Call
TOT_CALL_FEE	Sum of Total call fee
INT_OUT_CALL	Total International calls
RATIO_INT_TOT	Ration of INT_TOT
INC_CALL	Number of Incoming Call
TOTAL_SMS	Number of SMS sent
DATA_USAGE	Total data usage
FRAUD_STATUS	Identifying the subscriber type

After aggregating attributes, the last task is preparing the data in file formate that is suitable to ML tool. The tool accepts CSV and arrf file formats.

4.4.4 *Validation Techniques*

Crosss Validation options is a resampling procedure used to evaluate machine learning models on a limited data sample. Split the dataset into k-partitions or folds (default 10-folds). Train a model with all of the partitions except one that is held out as the test set, then repeat this process k times and creating k-different models (Table 4.4.2) and give each fold a chance of being held out as the test set. In 10-fold cross-validation, a single instance used for both testing and training. The accuracy estimate is the overall number of correct classifications from the 10 iterations divided by the total number of samples in the initial dataset. In its training and testing phase a single instance was used for testing and training.

Table 4.4.2: 10-Fold Cross-Validation Process

No	Train classifier on folds	Test against fold
1	2, 3, 4, 5, 6, 7, 8, 9, 10	1
2	1, 3, 4, 5, 6, 7, 8, 9, 10	2
3	1,2, 4, 5, 6, 7, 8, 9, 10	3
4	1, 2, 3, 5, 6, 7, 8, 9, 10	4
5	1, 2, 3, 4, 6, 7, 8, 9, 10	5
6	1, 2, 3, 4, 5, 7, 8, 9, 10	6
7	1, 2, 3, 4, 5, 6, 8, 9, 10	7
8	1, 2, 3, 4, 5, 6, 7, 9, 10	8
9	1, 2, 3, 4, 5, 6, 7, 8, 10	9
10	1, 2, 3, 4, 5, 6, 7, 8, 9	10

This is the most used testing method. Table 4.4.2 describes how train classifier on each fold and test against folds works. When the second iteration applied which

trained on folds 1 to 10 except fold 2 and tested with the 2nd fold and iterate it 10 times. The accuracy estimate is the ratio of sum of correctly classification from the whole iteration to the total number of instances in dataset

Use training set: The classifier is evaluated on how well it predicts the class of the instances it was trained on.

Supplied test set: The classifier is evaluated on how well it predicts the class of a set of instances loaded from a file. In this supplied test experimental process, the researcher creates two datasets for training and test from the total 349,164 instances with resampling method. The first dataset has 261,873 instances for training and the second dataset which has 87,291 instances for testing as shown in Figure 4.4.1. In case of supplied test a single instance never used for training and testing.

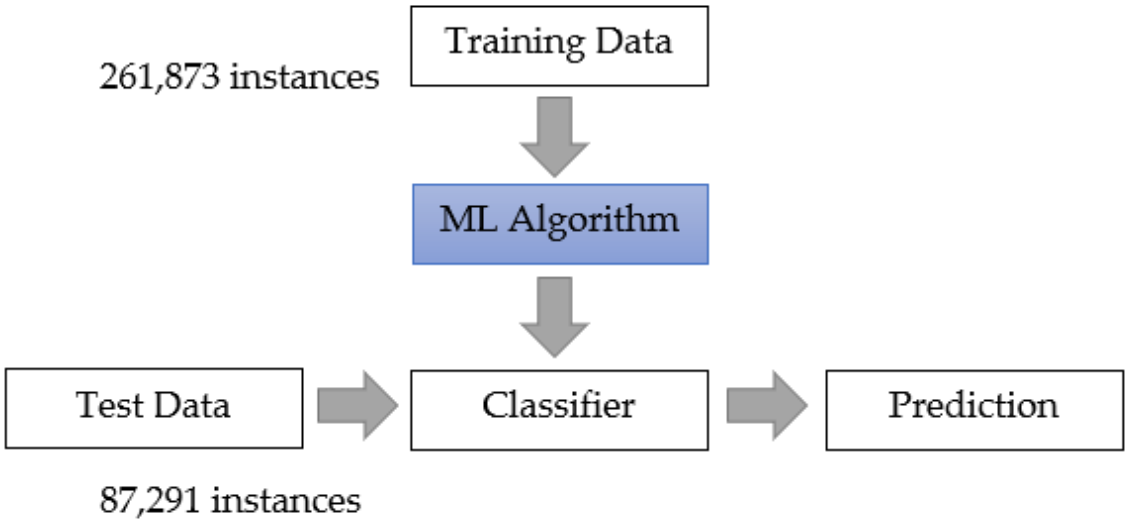


Figure 4.4.1: Supplied Test Technique

4.4.5 *Algorithm Training*

A total of six experiments performed with both 10-Fold cross-validation and supplied test options. For the cross-validation case the three algorithms have been using same dataset size of 349,164 instances whereas for the second test options

supplied test the three algorithms use 261,873 for train and 87,291 for testing. For training purpose, a label attribute was added to the dataset before the experiment started. In this study algorithms were experimented with their default parameters. In many cases, using the default setting parameters scores adequate results, however, compare results /models and achieve the research objectives other options are considered. Results of each experiments evaluated with algorithm's performance measurement parameters which discussed under Section 4.5.

4.5 PERFORMANCE MEASUREMENT PARAMETERS

The next tasks after training and testing the algorithms, evaluating the algorithms outcomes based on their performance measures parameters. The upcoming subsections describe about parameters for comparing and analyzing the algorithms.

4.5.1 *Confusion Matrix*

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. As we have two classes for this study a confusion matrix will be a 2X2 matrix which contains True Positive, False Positive, True Negative, and False Negative values for the class labels. Confusion matrix for two class labels ('Actual' class and 'Predicted' Class) is shown in Table 4.5.1.

Table 4.5.1: Confusion Matrix

		Predicted Class	
		Class Positive	Class Negative
Actual Class	Class Positive	True positive	False Negative
	Class Negative	False Positive	True Negative

True positive and true negatives are the observations that are correctly predicted. A good classifier minimize false positives and false negatives values.

True Positives - TP - These are the correctly predicted positive values which means that the value of actual class is Fraudulent and the value of predicted class is also Fraudulent.

True Negatives - TN - These are the correctly predicted negative values which means that the value of actual class is Legitimate and value of predicted class is also Legitimate.

False positives and false negatives, these values occur when actual class contradicts with the predicted class.

False Positive - FP – When actual class is Negative and predicted class is Positive.

False Negative - FN – When actual class is Positive but predicted class in Negative.

Once we understand these four parameters then we can calculate Accuracy, Precision, Recall and F-Measure.

4.5.2 *Accuracy*

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations (rate of total correct classification).

Mathmatical Equation 4.2

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.2}$$

4.5.3 *F-Measure*

F-measure is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test. It can be mathmatically expressed as in Equation 4.3

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.3}$$

Precision can be intuitively understood as the classifier's ability to only predict really positive samples as positive. For example, a classifier that classifies just everything as positive would have a precision of 0.5 in a balanced test set (50% positive, 50% negative). Classifies only the true positives as positive would have a precision of 1.0. So basically, the less false positives a classifier gives, the higher is its precision.

Mathmatical expression of precision shown in Equation 4.4

$$Precision = \frac{TP}{TP + FP} \tag{4.4}$$

Recall can be interpreted as the amount of positive test samples that were actually classified as positive. A classifier that just outputs positive for every sample, regardless if it is really positive, would get a recall of 1.0 but a lower precision. The less false negatives a clasifier gives, the higher is its recall.

Mathmatically Recall expressed as : Equation 4.5

$$Recall = \frac{TP}{TP + FN} \tag{4.5}$$

4.5.4 *Root Mean Squared Error (RMSE)*

Root Mean Squared Error (RMSE) is a frequently used measure of the differences between sample values predicted by a model and the values observed. RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It's the square root of the average of squared differences between prediction and actual observation. The Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction.

RMSE expressed mathematical as in Equation 4.6.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(d_i - f_i)^2} \tag{4.6}$$

4.5.5 *Receiver Operating Characteristic Curve - ROC*

ROC curve is a graph of FPR Vs TPR. The area measures the ability of the classifier to correctly classify the test data. It shows performance of models across all possible thresholds. The wider the coverage area of the model a better classifier it is.

5

RESULT AND DISCUSSION

In this chapter a comparative performance analysis of three different machine learning algorithms, done on a subscription fraud detection is presented. For each algorithm, two experiments were performed with the two validation techniques supplied test and cross-validation. The dataset size and validation techniques are described in Section 4.4.4. Each algorithm model trained with ML tools using the default configuration parameters. The results of the two validation techniques on each algorithm were recorded with respect to their processing time and performance measurement parameters (shown in Table 5.1.1). Conclusively, the best model for subscription fraud detection proposed.

5.1 RESULTS AND COMPARISON

Performance metrics that were using for comparisons are accuracy, precision, recall, F-Measure, RMSE and ROC curves. Table 5.1.1 shows classifiers classification performance results of the algorithms besides the validation techniques. The highest classification accuracy noted from cross-validation and supplied test options are J48 algorithms with 99.3% and 99.2% respectively. Both validation technique results are comparable in the case of J48 algorithm. On the contrary, SVM scores the smallest accuracy of 94.71% from the total experimentation results using cross-validation option and relatively bigger with 1.29% in its supplied test option than cross-validation. ANN is relatively the second-highest classifier in cross-validation and supplied test 97.51% and 96.57% respectively.

Table 5.1.1: summarized performance metrics of all the classifiers

Features	Cross Validation			Supplied Test		
	Algorithm			Algorithm		
	SVM	J48	ANN	SVM	J48	ANN
Build time	196	9	341	49	8	267
Evaluate	1,239	42	2,528	339	21	550
Precision	0.949	0.993	0.975	0.961	0.992	0.967
Recall	0.947	0.993	0.975	0.96	0.992	0.966
F-Measure	0.948	0.993	0.975	0.959	0.992	0.965
RMSE	0.2298	0.074	0.1347	0.1998	0.0750	0.1730
ROC	0.94	0.981	0.933	0.892	0.979	0.908
Accuracy	94.71	99.3	97.51	96	99.22	96.57

Considering the comparison of classification in terms of time (model building and evaluation), J48 using supplied test took a better building time of 8s whereas ANN's 341s with cross-validation technique is the elongated building time over other algorithms. In terms of model evaluation time, J48 using supplied test has a better evaluation time of 21s followed by J48 algorithm 42s using cross-validation techniques. But ANN with cross-validation took longer evaluation times (2,528s) than other algorithms.

The confusion matrix allows the visualization of the performance of an algorithm on a set of test data for which the true values are known. As described in the summary of confusion matrix Table 5.1.2, J48 cross-validation technique achieves relatively smaller incorrectly classified 0.7% or 2,483 from the total 349,164 instances whereas the two algorithms SVM and ANN incorrectly classifies 3,486 and 8,693 from their highest-scoring using cross-validation performance results respectively.

Table 5.1.2: Summary of confusion matrix

		Correctly Classified		Incorre Classified	
Test Mode	Algorithm	TP	TN	FP	FN
Supplied Test	J48	20,930	65,688	673	0
	SVM	18,402	65,403	3,201	285
	ANN	18,618	65,685	3	2,985
Cross Validation	J48	85,366	261,315	1,045	1,438
	SVM	80,101	250,617	6,310	12,136
	ANN	82,400	258,071	4,011	4,682

Classification performance of algorithms can be identified by their error. This error level is described by RMSE value as shown in Figure 5.1.1. The lower RMSE tells us the level of the Proficiency of the algorithm's detecting frauds. J48 using cross-validation scores 0.074 RMSE which is lower relative to all other experimental RMSE value. SVM using CV records the highest RMSE (0.2298) than other algorithms. ANN using CV scores (0.1347) relatively smaller than its' ST (0.173) value but larger than J48 using ST value 0.075.

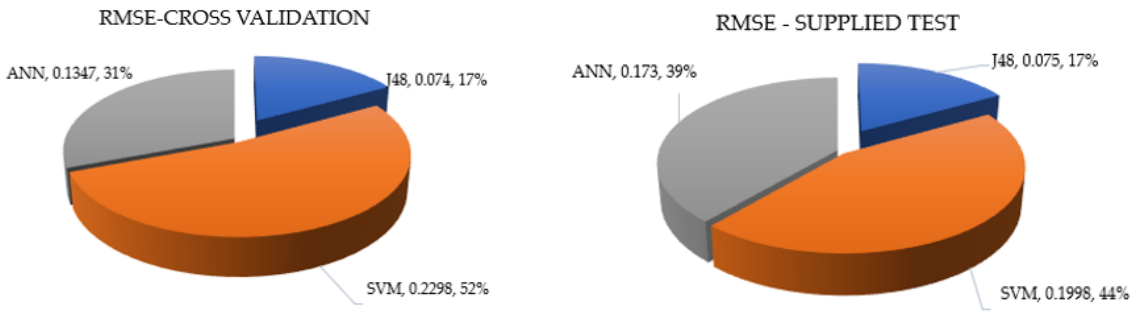


Figure 5.1.1: Comparison of models with RMSE value

When algorithms performance measurements in terms of TP and FP detection rate, the classifier needs to be evaluated in terms of precision, recall and F-measures using supplied test options shown in Figure 5.1.2. From the graph, we can observe

that the J48 algorithm scores the highest with similar values of 0.992 for the three performance measures compared with SVM and ANN algorithms. However, ANN 0.967 is the second-highest precision value than SVM 0.961 value. SVM F-measures 0.959 is the least measuring value of the supplied test option in other words SVM is relatively the lowest classifier than the other two algorithms related to the three performance metrics. The higher the precision value means the lower false positive detected. In fraud detection, lesser false positive detection is preferable because it minimizes the risk of blocking legitimate subscribers.

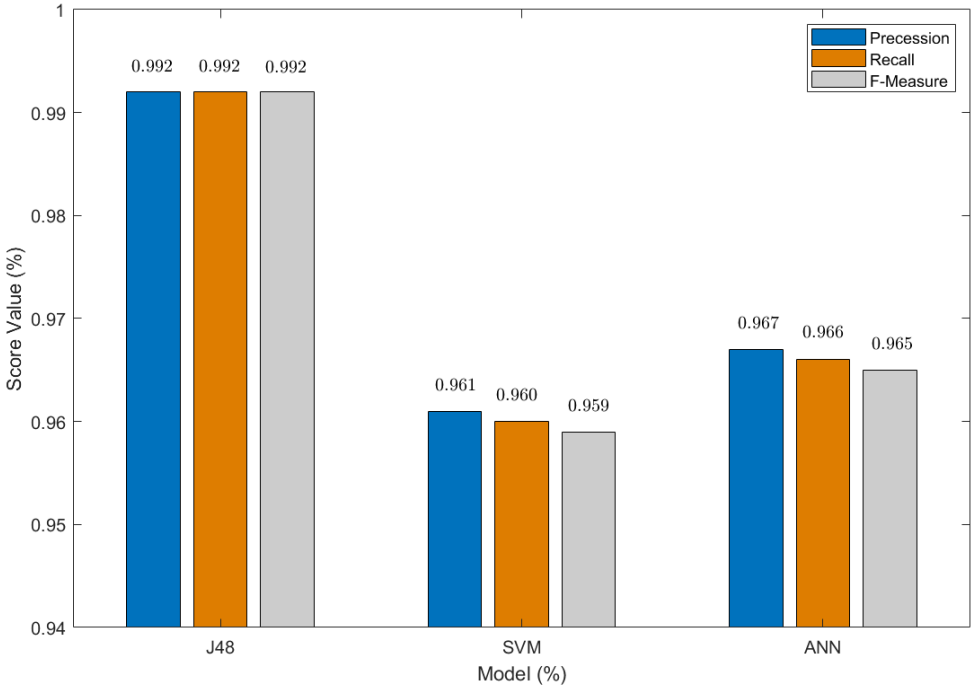


Figure 5.1.2: Comparison of models on Precision, Recall and F-measure using ST

On a similar fashion, algorithms need to compare using cross-validation options' with the same performance measurment metrics precision, recall and F-measure values as shown in Figure 5.1.3. J48 and ANN scores similar results for the three performance measurment metrics 0.993 and 0.975 respectively. J48 is scored the highest for the three metrics values. But, SVM scores the smallest value of the three measured metrics precision (0.949), recall (0.947) and F-measure (0.948). The highest the precision and recall value indicates that the classifier is best.

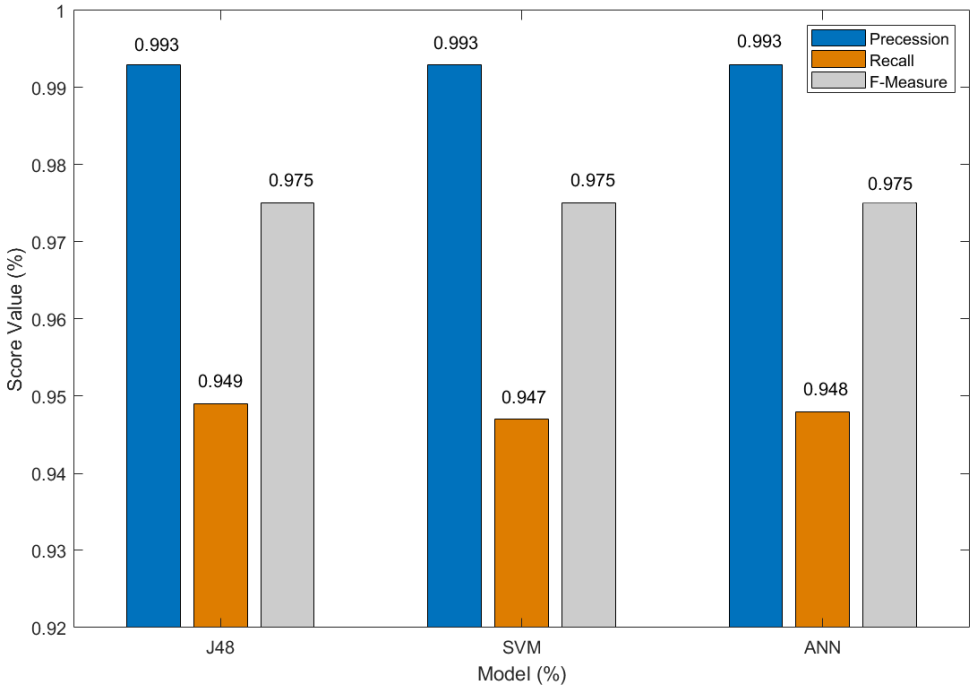
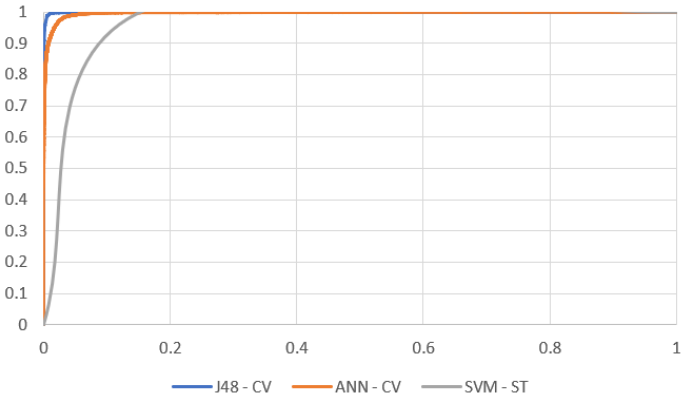


Figure 5.1.3: Comparison of models on Precision, Recall and F-measure using CV

ROC curve is another choice to compare the performances of the proposed algorithms. It is a standard technique which is used for summarizing classifier performance through a range of tradeoffs between false positive and true positive error rates. As it is described in performance measure Table 5.1.1 the lower ROC value recorded on each algorithm on SVM, J48 and ANN with suppled test validation 0.892, 0.979 and 0.908 respectively. But, each algorithm's uppermost ROC value is plotted as shown in Figure 5.1.4 for further comparison to select the superlative classifier based on Area Under the Curve (AUC). Among the smallest ROC results SVM is the least.

ROC Figure 5.1.4 shows that J48 followed by ANN is the best classifier due to their wider coverage area of the curve. J48 curve laid to the vertical true positive rate axis which approximated to [0,1). This indicates that the algorithm was accurate in other words it measures the proportion of actual positives that are correctly identified as positive values. In contrary, SVM is the least classifier compared to the two selected (highest) algorithms based on its coverage area.

Figure 5.1.4: Comparison of ROC curve to the highest classifier algorithms



The main objective of this study was analysing which algorithms perform better for detecting subscription fraud based on their classification performance. As the result depicted in the above sections comparisons in both validation options the highest accuracy results of each algorithm tabulated in Table 5.1.3. So, the final findings of the study help to answer the research questions which has been initiated at the beginning of the study.

Research question

1. What kind of data features can be used in order to identify Subscription fraud?

   As the objective of this research, subscription fraud detection grounded on the subscribers usage patterns, usefull data features were identified.

   - Calling number - identify the subscriber who initiate the call
   - Called number which identify the destination number
   - Call type which identify the call destination
   - Start time which describe the calling start time
   - Duration which describe duration of call spent
   - Call fee which describe how much money is spent per certain period
   - Data usage which describe how much data the subscriber uses
   - SMS which describe how much SMS the subscriber sent/receive

- Service type which identify the service type

2. What kinds of ML algorithm can be used in order to detect subscription fraud?

   The highest accuracy results of each classifier tabulated on Table 5.1.3. So that, it is concluded that the J48 decision tree using cross validation performs better than the other two algorithms with accuracy, precision, recall, ROC and RMSE for newly arriving instances of subscription fraud.

Table 5.1.3: Summary of highest accuracy

Validation	Algorithm	Accuracy
Supplied test	SVM	96
10-fold	J48	**99.3**
10-fold	ANN	97.51

6

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

In the twenty-first century the development of technology is advancing very rapidly. This telecommunication industries era stimulates certain characteristics of the fraudsters. Due to fraudsters adaptability or advancing themselves with new technologies and behavioral changes telecommunication companies suffer from fraudsters activity. Among many fraud types, one of the common and predominant types of fraud is subscription fraud in which usage type is in contradiction with subscription type. It is a contractual fraud type and is the starting point of other fraud types. Despite many controls in place, subscription fraud is still widespread and affects every telecommunation operators.

The focusing point of this study was train and analysing which algorithms perform better for detecting subscription fraud based on their classification performance. To achieve the goal of this research a CDR data were collected from ethio telecom and preprocessing tasks were applied to clear unnecessary missing data values and to remove outliered data. Attribute selection task is a key for fraud detection. To select attributes subscription fraud behaviours were identified by the help of domain expert advice in addition to related paper review. After eliminating irrelevant attributes nine out of 33 were selected. To have full information of the subscriber, attributes were aggregated in subscriber level to discriminate legitimate subscribers from fraudulent based on the behaviour of subscription fraud.

A total of six experiments were done using the two validation techniques tenfold cross-validation and separate test data on the three ML algorithms namely J48, ANN and SVM. In separate test case instances in the training dataset never

included in the test dataset. However, in CV cases the validation technique has knowlages of each instances as it uses a single instance for training and testing.

The performance of all algorithms using both validation technique were evaluated based on various metrics including accuracy, precision, Recall, F-measures, RMSE, ROC and time (building and evaluation). These metrics results provides quantitative understanding of their suitability to subscription fraud detection.

As a result, J48 algorithm (99.3%) is a superlative classifier that perfectly fits the prediction solution of subscription fraud detection based on the performance measure evaluation matrices. This result happens because of its capable of learning disjunctive expressions in addition to it reduced error pruning. Pruning decreases the complexity in the final classifier, and therefore improves predictive accuracy from the decrease of over fitting. The two algorithms highest scores of ANN (CV) and SVM ST with 97.51% and 96.0% respectively. For the experimentation techniques, SVM shows a little improvements in supplied test options by 1.29%. But ANN on the other hand improves on the cross-validation techniques by 0.94%. In case of J48 both validation technique results are comparable. The performance measurement results show that cross-validation technique preferred as compared to the results recorded from the supplied test.

This research will play an important role in controlling and preventing the current fraudulent threat in ethio telecom specifically subscription fraud detection. Moreover, this research gives a telecom operator to identify the fraudulent subscribers from legitimate and benefits the telecommunications company to decrease revenue losses caused by fraudsters, increase their profitability, increase trust relationships with their customers and build the company brand.

6.2 FUTURE WORK

Proposed future works from the researcher are:

This study was carried out on prepaid mobile with two months sample data. Increasing the dataset size may improve the performance and accuracy of the technique.

with similar methodology and techniques research can be performed with additional unseen attributes.

with these similar techniques and algorithms research can be conducted for other fraud types

This study was carried out on prepaid mobile but the techniques proposed here could be extended to subscription fraud in fixed and post-paid mobile communications.

BIBLIOGRAPHY

[1] S. Qayyum, S. Mansoor, A. Khalid, Z. Halim, A. R. Baig, *et al.*, "Fraudulent call detection for mobile networks," in *2010 International Conference on Information and Emerging Technologies*, IEEE, 2010, pp. 1–5.

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

[3] S. Wu, N. Kang, and L. Yang, "Fraudulent behavior forecast in telecom industry based on data mining technology," *Communications of the IIMA*, vol. 7, no. 4, p. 1, 2007.

[4] R. A. Becker, C. Volinsky, and A. R. Wilks, "Fraud detection in telecommunications: History and lessons learned," *Technometrics*, vol. 52, no. 1, pp. 20–33, 2010.

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

[6] K. Gonzales. (Apr. 2018). Telecom fraud: 29billionandcounting–whyitmattersmoretho [Online]. Available: `https://www.telecomengine.com/article/telecom-fraud-29-billion-and-counting-why-it-matters-more-than-ever-in-the-digital-era/`.

[7] oseland. (2013). Communications fraud control association. 2013 global fraud loss survey. N. (CFCA), Ed., [Online]. Available: `http://www.cfca.org/press.php`.

[8] G. Y. Koi-Akrofi, J. Koi-Akrofi, D. A. Odai, and E. O. Twum, "Global telecommunications fraud trend analysis," *International Journal of Innovation and Applied Studies*, vol. 25, no. 3, pp. 940–947, 2019.

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

[10] C. F. C. Association *et al.*, "2017 global fraud loss survey," *Press Release, June*, 2017.

[11] fanabc. (Oct. 1, 2019). Telecomfraud. fanabc, Ed., [Online]. Available: `https://www.fanabc.com/english/2018/10/ethio-telecom-mulls-over-preventing-telecom-fraud`.

[12] A.-A. Ababa. (Mar. 6, 2017). Ethiopia-telecom fraud. A.-A. Ababa, Ed., [Online]. Available: `http://apanews.net/en/news/ethiopia-loses-over-52m-to-telecom-fraud-official`.

[13] A. Wiens, T. Wiens, and M. Massoth, "A new unsupervised user profiling approach for detecting toll fraud in voip networks," in *The Tenth Advanced International Conference on Telecommunications (AICT 2014) IARIA*, 2014, pp. 63–69.

[14] P Saravanan, V Subramaniyaswamy, N Sivaramakrishnan, M Prakash, and T Arunkumar, "Data mining approach for subscription-fraud detection in telecommunication sector," *Contemporary Engineering Sciences*, vol. 7, no. 11, pp. 515–522, 2014.

[15] O. A. Abidogun, "Data mining, fraud detection and mobile telecommunications: Call pattern analysis with unsupervised neural networks," PhD thesis, University of the Western Cape, 2005.

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

[17] S. S. Aksenova, "Machine learning with weka weka explorer tutorial for weka version 3.4. 3," *sabanciuniv. edu*, 2004.

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

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

[20] S.-T. et al., "Business applications of neural networks. the state-of-the-art of real world applications," *Novel techniques for profiling and fraud in mobile telecommunications.*, Dec. 18, 2000.

[21] C. S. Hilas and P. A. Mastorocostas, "An application of supervised and unsupervised learning approaches to telecommunications fraud detection," *Knowledge-Based Systems*, vol. 21, no. 7, pp. 721–726, 2008.

[22] G. M., "Telecoms fraud," *Computer Fraud & Security*, Jul. 15, 2010.

[23] WeDo. (Dec. 18, 2019). How to handle subscription fraud. WeDo, Ed., [Online]. Available: https://web.wedotechnologies.com/hubfs/10_RAID.Cloud/Subscription%20Fraud/Datasheets/RAID-CLOUD-Subscription-Fraud-datasheet.pdf.

[24] H. et.al, "An application of supervised and unsupervised learning approaches to telecommunications fraud detection," *Knowledge-Based Systems*, vol. 21, no. 7, pp. 721–726, 2008.

[25] D. Bales. (Oct. 10, 2019). Clone or swap? sim card vulnerabilities to reckon with. L. Kessem, Ed., [Online]. Available: https://securityintelligence.com/posts/clone-or-swap-sim-card-vulnerabilities-to-reckon-with/.

[26] kahsu hagos, *Sim-box fraud detection using data mining techniques: The case of ethio telecom*, D. Ephrem, Ed., Nov. 3, 2018.

[27] G. Maciá-Fernández, "Roaming fraud: Assault and defense strategies,"

[28] R. Domingues, *Machine learning for unsupervised fraud detection*, 2015.

[29] K. K. Tsiptsis and A. Chorianopoulos, *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons, 2011.

[30] G. Kaur and A. Chhabra, "Improved j48 classification algorithm for the prediction of diabetes," *International Journal of Computer Applications*, vol. 98, no. 22, 2014.

[31] S. K. Jayasingh, J. K. Mantri, and P Gahan, "Comparison between j48 decision tree, svm and mlp in weather forecasting,"

[32] T. R. Patil, S. Sherekar, *et al.*, "Performance analysis of naive bayes and j48 classification algorithm for data classification," *International journal of computer science and applications*, vol. 6, no. 2, pp. 256–261, 2013.

[33] A. F. Sheta and A. Alamleh, "A professional comparison of c4. 5, mlp, svm for network intrusion detection based feature analysis," in *The International Congress for global Science and Technology*, vol. 47, 2015, p. 15.

[34] H. Sug, "The effect of training set size for the performance of neural networks of classification," *WSEAS Transactions on Computers*, vol. 9, no. 11, pp. 1297–1306, 2010.

[35] L. Auria and R. A. Moro, "Support vector machines (svm) as a technique for solvency analysis," 2008.

[36] J. Nalepa and M. Kawulok, "Selecting training sets for support vector machines: A review," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 857–900, 2019.

[37] S. Subudhi and S. Panigrahi, "Use of fuzzy clustering and support vector machine for detecting fraud in mobile telecommunication networks.," *IJSN*, vol. 11, no. 1/2, pp. 3–11, 2016.

[38] R. Berwick, "An idiot's guide to support vector machines (svms)," *Retrieved on October*, vol. 21, p. 2011, 2003.

[39] M. Kamel, *Data preparation for data mining*. 2009.

[40] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for svms," in *Advances in neural information processing systems*, 2001, pp. 668–674.

[41] H. Taherdoost, "Sampling methods in research methodology; how to choose a sampling technique for research," 2016.

[42] E. Acuna and C. Rodriguez, "The treatment of missing values and its effect on classifier accuracy," in *Classification, clustering, and data mining applications*, Springer, 2004, pp. 639–647.

[43] I. Ben-Gal, "Outlier detection," in *Data mining and knowledge discovery handbook*, Springer, 2005, pp. 131–146.

A

APPENDIX

A.1 CDR TABLE

Table A.1.1: CDR TABLE

No	Attributes	Description
1	CDR_ID	CDR Sequence Number
2	RE_ID	Service Identifier
3	BILLING_NBR	Billing Number
4	CDR_TYPE	Call type Id(The types of call 0: local call , 1: toll call within a charging area, 2: toll call between charging areas ,3: international toll call)
5	CALLING_NUMBER	Calling Number(call initiate number)
6	CALLED_NUMBER	call destination number
7	CALLING_IMEI	International mobile equipment identity
8	CALLING_IMSI	IMSI of the calling party
9	THE_THIRD_PARTY_NUMBER	Third Party Number
10	CALL_START_TIME	the time when call start
11	CALL_END_TIME	the time when call end
12	CALL_DURATION	Call duration
13	CALL_FEE	the actual money deducted
14	CALLED_COUNTRY	called number country code of

Continued on next page

Table A.1.1 – continued from previous page

No	Attributes	Description
15	CALLING_CARRIER	Calling carrier
16	CALLED_CARRIER	Called carrier
17	CALLING_DISTRICT	Cell ID of the calling party
18	CALLED_DISTRICT	Cell ID of the called party
19	STATUS_DATE	Billing date
20	CALLING_SUB_ID	Calling subscriber ID
21	BILLING_CYCLE_ID	Billing cycle ID
22	CHARGE_1	Charge amount of you spend
23	CHARGE_2	Charge amount of you get disscount
24	RATE_ID1	Rate ID
25	ACCOUNT_ITEM_ID1	Account item ID
26	UPLOAD_TRAFFIC	Upload tra fficc
27	DOWNLOAD_TRAFFIC	Download traffic
28	BILLING_OFFERING_ID	Billing offering ID
29	ERROR_CDT_TYPE	Error CDR Indicator
30	CALLFORWARDINDICATOR	Call Forward Indicator
31	HOTLINEINDICATOR	Hot Line Indicator (voice mail)
32	CALLING_TRUNK_ID	Calling Trunk ID
33	CALLED_TRUNK_ID	Called Trunk ID

A.2 FILE UPLOADER SCRIPT

This script uploads dat files to the database

==========

@echo off @setlocal enabledelayedexpansion set source=E:Root

:Endegena for /f "delims=" 'dir /b /a-d /tw /od ") do (RENAME "if /f :Aleke DEL PAUSE

:MV ::if set g=!Fileholer:0,-3! MOVE "

:clldr cd "sqlldr userid=tis/tis control=E:
FTP_Root

:loadingerror :: this part of the code writes error log with the name of the file which caused error echo error on !Fileholer! >

:delete DEL

:rnm :: This function moves .bad files to sepcific directory ::echo the value of !Fileholer! MOVE "RENAME "RENAME "

PAUSE

==========

A.3 FILE LOADER

```
@echo off
@setlocal enabledelayedexpansion
set source=E:\derebe\\active
set dest=E:\derebe\\run_active
set badfiles=E:\derebe\\run_active\Badfiles
:Endegena
for /f "delims=" %%a in (
    'dir /b /a-d /tw /od "%source%\*.csv"'
) do (RENAME "%source%\%%a" "CBS_Billing.csv" && set Fileholer=%%a && goto MV )
if /f %source%\*.csv NEQ ""
:Aleke
DEL %dest%\CBS_Billing.csv
PAUSE
:MV
::if %%a == [] goto Aleke
set g=!Fileholer:0,-3!
MOVE "%source%\CBS_Billing.csv" "%dest%\" && goto clldr
:clldr
cd "%dest%\"
sqlldr userid=fit/fit control=D:\Research\Fraud\Exported_Data\run_active\original_sqlldr.ctl
log=D:\Research\Fraud\Exported_Data\run_active\log.log && if errorlevel == 1 goto loadingerror
if errorlevel == 0 goto rnm
echo exiting on loading error !errorlevel! && PAUSE
:loadingerror
:: this part of the code writes error log with the name of the file which caused error
echo error on !Fileholer! > %dest%\!Fileholer!_error.txt && goto Aleke
:delete
DEL %dest%\CBS_Billing.csv && goto Endegena
:rnm
:: This function moves .bad files to sepcific directory %badfiles% and rename it
to the loaded dat file name additionaly renames log files to the loaded data file name
::echo the value of !Fileholer!
MOVE "%dest%\CBS_Billing.bad" "%badfiles%"
RENAME "%badfiles%\CBS_Billing.bad" "!Fileholer!.bad"
RENAME "%dest%\log.log" "!Fileholer!.log"  && goto delete
PAUSE
```

Figure A.3.1: File loader

A.4 ORACLE SCRIPTS

Merging two tables

```
-------------SMS and Voice Agrr----------------------------
MERGE INTO F_TOTAL_AGRR d
  USING F_sms_agrr b
ON (d.msisdn = b.msisdn and d.called_date = b.called_date)
WHEN MATCHED THEN
UPDATE SET d.total_sms = b.tot_sms;

-------------------------------------------------------------
```

Figure A.4.1: Merging two tables

Aggregating voice data

```
INSERT INTO AGGR_L_VOICE s (s.MSISDN,s.Tot_calls,s.Distnct_Calls,
s.Tot_Call_Duration,s.Ratio_DistCall_TotaL,s.Tot_CALL_FEE)
select msisdn,count(msisdn),count(distinct(called_number)),round(sum(Call_duration/60),2),
round(count(distinct(called_number))/count(called_number),2),round(sum(call_fee/10000),2)
from dere.subsc_L_voice where call_start_time BETWEEN
TO_DATE('25-JUN-19 00:00:00', 'DD-MON-YY HH24:MI:SS')
AND TO_DATE('25-JUL-19 04:00:00', 'DD-MON-YY HH24:MI:SS')
group by msisdn;

update aggr_L_voice set flage=0 where flage is null;
```

Figure A.4.2: Attribute Aggregation

Table Joining

select * from aggr_l_voice lv , agrr_l_data ld

where lv.flage = ld.flage

```
select lv.*,ld.conn_duration,ld.USAGE_DATA_MEG,
       ld.CONN_DURATION,TOT_DATA_FEE
          from aggr_l_voice lv
       LEFT join agrr_l_data ld
             on lv.flage = ld.flage
       left join aggr_l_voice ls
             on ls.flage = ld.flage
```

Figure A.4.3: Table Joining

Removing unneccessary called_numbers is premuim numbers

remove * from source_voice_table where called_number like '994%', '905%', '991%';