Predicting Termination of Mobile Subscribers Using Classification Algorithms

By: Kiflemariam Mehari

Advisor: Beneyam Berehanu (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 Science in Telecommunication Engineering

Addis Ababa University

Addis Ababa, Ethiopia

December, 2019

DECLARATION

I, the undersigned, declare that the thesis consists of my own work in accordance with internationally accepted practices; I have fully recognized and referred all the materials used in this thesis.

Kiflemariam Mehari	_____
Name	Signature

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

This is to certify that the thesis prepared by **Kiflemariam Mehari**, entitled Predicting Termination of Mobile Subscribers Using Classification Algorithms and submitted in partial fulfillment of the requirements for the degree of Master of Science in Telecommunication Engineering complies with the regulations of the university and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Internal Examiner _____ Signature_____Date _____

External Examiner _____Signature_____Date _____

Advisor Beneyam Berehanu(PhD)_Signature_____Date _____

Co-Advisor _____ Signature_____Date _____

_____

Dean, School of Electrical and Computer Engineering

ABSTRACT

Service termination is one key challenge of operators that can significantly reduce their revenue. It is still a challenge even in Ethiopian monopoly market even if users do not have alternative service provider. For instance, the market has seen 1.6 percent overall prepaid mobile subscriber termination rate within a quarter, let alone specific service termination. To address this challenge, operators need to understand causes of termination and take timely proactive actions to mitigate number of terminations. For this purpose, they need to accurately and timely predict subscribers with potential service termination based on collected service related data. Performance of various classification algorithms have been investigated to predict subscribers' behavior. However, performance of such algorithms are not studied in the context of Ethiopia where ethio telecom has an enormous data that helps to define its subscribers' behavior.

Objective of this thesis work is to study performance of classification learning algorithms for predicting termination in the context of Ethiopian prepaid subscribers. Studied algorithms are *J48*, *Random Forest* and *Naïve Bayes* and their performance are compared in terms of prediction accuracy, performance, errors and interpretability of their model. Algorithms effectiveness and efficiency are evaluated considering two validation methods (*Percentage Split* and *Cross Validation*) and three data sets. For the performance evaluation, we used WEKA 3.8 tool algorithm implementation.

Obtained results show that Random Forest scores the highest prediction accuracy while Naïve Bayes scores the least. *Random Forest* and *Naïve Bayes* scores their best at 93.4% and 85.9% respectively, besides J48 scores 93.3%. J48 is as accurate & robust as *Random Forest*. Moreover, it provides the most interpretable and clear model.

Key words: *J48, Random Forests, Naïve Bayes, service termination, subscribers' defection, churn, confusion matrix*

ACKNOWLEDGMENTS

First and above all, I thank The Almighty God for giving me the strength, patience and knowledge to complete this research.

I would like to express my special gratitude to my advisor, Beneyam Berehanu (PhD), for his guidance, priceless advices, insightful discussion throughout this research. Dr. Beneyam, thank you for your valuable and constructive comments. I would also like to thank my evaluators Surafel Lemma (PhD) and Yihenew Wondie (PhD) for their vital and important feedbacks in the thesis progress seminars.

I would like to give specific thanks to ethio telecom especially for employees of the Corporate Application Section for their support and patience in the data collection.

Last but not the least, I pass my heart felt gratitude to my family who always show me their solidarity when I was in challenge.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AI	Artificial Intelligence
ARPU	Average Revenue Per User
CAF	Customer Application Form
CBS	Customer Billing System
CDR	Call Detail Record
CRM	Customer Relationship Management
DT	Decision Tree
ECA	Ethiopian Communication Authority
ECAF	Electronic Customer Acquisition Form
FN	False Negative
FP	False Positive
GPRS	General Packet Radio Service
GSMA	Global System For Mobile Communication Association
HDFS	Hadoop Distributed File System
IG	Information Gain
IMEI	International Mobile Equipment Identity
IPCC	IP Contact Center
IS	Information System
ML	Machine Learning
MMS	Multimedia Messaging Service
MVAS	Mobile Value Added Services
NB	Naïve Bayes
NLP	Natural Language Processing
OCR	Optical Character Recognition

ODM	Oracle Data Miner
PRS	Positioning Reference Signal
REP	Reduced Error Pruning
RF	Random Forest
SIM	Subscriber Identity Module
SMS	Short Message Service
TN	True Negative
TP	True Positive
WEKA	Waikato Environment For Knowledge Analysis

1 INTRODUCTION

Service termination is one key challenge of operators that can significantly reduce their revenue. Being able to predict the termination of subscribers in advance provides a company to have a valuable insight to maintain and improve its customer base [1], [2]. Early identification of subscribers that leave telecom services and retain them with attractive offers and discounts is one of the main business strategy of the telecommunications service providers. Since prepaid subscribers are not required to notify the service provider when they quit using service(s), they could leave the company unexpectedly. Fortunately, operators have enormous data to identify their subscribers who are inclined to leave a service. Telecom companies are trying to use their huge amount of day to day transactional data for purposes like understanding customer usage/ spending patterns and predicting their future behavior [3].

In most industries, retaining existing customers saves a cost anywhere between 10 to 25 times more than acquiring a new one. Thus, increasing customer retention rates by just 5% can increase profits by 25% to 95% [4]. Actually, when a telecom subscriber leaves from a service, operators lose not only the revenue from a subscriber but also the cost of material, labor and time resources incurred to acquire that subscriber.

Even though ethio telecom is a monopoly operator in Ethiopia, it scored prepaid subscribers termination rate of 1.6 percent from November 2018 to January 2019 as it is shown in Table 1-1. This rate is approximately similar to the churn rate on the first semi-annual report of Telefonica, an operator in Europe and Latin America countries, in 2019. It is reported that the average churn rate per month in the telecom sector is 2.2%. ethio telecom's subscriber termination rate is nearly becoming equivalent to the churn rate of operators that are working in a highly competitive environment [5], [6].

Table 1-1 ethio telecom Prepaid Subscribers Termination Rate

Prepaid Subscribers from Nov 2018 to Jan 2019		
A	Number of subscribers at the beginning of November 2018	37,280,894
B	Number of subscribers at the end of January 2019	38,167,654
C	Subscribers increment (B-A)	886,760
D	New subscribers acquired	1,474,572
E	Terminated Subscribers in the duration	587,812
F	Termination rate based on customer base (E/A)	1.6%
G	Termination rate based on new sales (E/D)	39.9%

As per the direction of the Ethiopian government, the Ethiopian Communication Authority (ECA) is in preparation to invite two telecom operators [7]. According to the director of ECA, the international bid will be announced until January 8, 2020. This will lead Ethiopia to have at least three telecom operators after April 2020 [7]–[10]. Therefore, ethio telecom is nearly going to face the competitive world in the telecom sector. ethio telecom has planned to attract and cooperate with experienced operators prior to competition. Thereby, 49 percent of its share will be available to the market to strengthen its capacity and confront the coming competition [8]–[10].

Many products and service offerings make the competition tighter, rather applying them wisely helps to win a competition [11]. To face the competition, a service provider first must know about its customers how they decide and what variable(s) influence their decision. One of the important data to study subscribers' behavior is Call Detail Record (CDR). A CDR contains at a minimum of: caller number, called number, call starting time, call end time, call duration, call type e.g. voice call, SMS, GPRS/MMS, call fee, terminal location and more than 30 different types of contents [12].

Data driven classification of subscribers is used to prepare a proper strategic plan, to provide proactive customer service, to design a better marketing strategy, to prepare customized retention plan, and so on. Machine learning can be easily applied to support the detection, classification and prediction of termination of subscribers [16]–[18], [23]–[26]. By analyzing an enormous data at the hand of a service provider, faulty subscribers can be identified early. Several prediction tools that support predictive modeling and classification process can be applied to find patterns from the actual data in a system. Therefore, the operator can apply a proactive retention activity or other strategies to absorb the loss from termination of subscribers.

1.1 STATEMENT OF THE PROBLEM

According to ethio telecom's Dashboard report, 670,128 prepaid mobile subscribers were terminated within four months (November 2018-February 2019). This significant number of subscribers' termination should be monitored in order to identify the factors behind their termination. Postpaid subscribers are often asked their reasons when they apply for termination of services but not for prepaid subscribers so far. In fact, prepaid mobile subscribers do not have an obligation or a binding contract to notify service provider prior to termination of a service. So it becomes very difficult to earn money from potentially disconnected subscribers. Once a subscriber is disconnected from telecom services, it has a challenge for the telecom operator to: campaign, offer goods and services, provide proactive customer service and context-based marketing. Additionally, mobile subscribers inclined to termination are deemed as active subscribers that overstates Average Revenue Per User (ARPU) expectations of the company. Thereby, a plan based on the expected number of active subscribers (customer base) is biased since potentially terminated subscribers are considered as active. Predicting the status of customers is, therefore, a priority for the planning of: a real customer base, customer service, promotion, and profit strategy. Moreover, entering to the emerging

telecommunications' competitive market without understanding the behavior of subscribers results in a customer churn that cannot be easily controlled.

Research questions

- Which features or attributes influence prediction of subscribers' termination?
- Which prediction technique can effectively and efficiently identify possible termination of subscribers?
- What are the tradeoffs to build understandable model?

1.2 OBJECTIVE

1.2.1 General Objective

The general objective of this study is to predict termination of prepaid mobile subscribers with effective & interpretable machine learning algorithm(s).

1.2.2 Specific Objectives

The specific objectives of this thesis are:

- To identify possible termination (deactivation) of prepaid mobile subscribers before quitting from services.
- To show how prepaid mobile subscribers quit a service based on combination of selected features (attributes).
- To select an effective & interpretable ML algorithm(s) for prediction of subscribers' termination.
- To model ethio telecom subscribers' termination.

1.3 SCOPE AND LIMITATIONS

1.3.1 Scope

The scope of this study is to predict possible termination of individual prepaid mobile subscribers in ethio telecom. As the behavior of subscribers is well known, their decision analysis will be more reliable. This research is about subscribers' behavior towards their status change to termination. It is mainly based on the information extracted from their CDR and profile data.

The data is combined from subscribers' categories: deactivated, suspended, barred and active status collected from Customer Billing System (CBS) and Customer Relation Management (CRM) systems. Even though unlimited data sets can be prepared in different arrangements, only six different arrangements are made to test the selected algorithms. Three classification algorithms (*J48*, *Random Forest* and *Naïve Bayes*) and two validation methods (*percentage split (70/30)* and *10-fold cross*) are applied for prediction on termination of subscribers.

1.3.2 Limitations

It takes a lot of time and resource to extract, pre-process, analyze and evaluate the data set of all subscribers. Thus, this analysis is limited to three months of subscribers' data, three classification algorithms and two validation methods. Lack of central data warehouse limits the study to apply a big data analytics in order to get a better prediction on termination of prepaid mobile subscribers.

1.4 CONTRIBUTIONS OF THE RESEARCH

This work adds value by creating a model that can effectively and efficiently forecast using machine learning classification algorithms to identify possible termination of subscribers based on their CDR usage and selected profiles. ethio telecom can thus

schedule proactive customer management, prepare an appropriate marketing strategy and retention programs to meet subscribers' need. In addition, ethio telecom can also be benefited from putting a strategic plan based on fairly categorized customer base such as ARPU. Moreover, evaluating the actions of existing subscribers tends to reduce churners in a competitive telecom market as early planning, which will undoubtedly occur in the near future. Finally, other researchers interested in predicting termination of subscribers from telecom services can base this research to improve or contribute their own.

1.5 LITERATURE REVIEW

Churn in literature defined as defection, customer loss, attrition, or turnover. Churn is simply used as the act of a subscriber switching from one service provider to another [1], [13], [14], [16], [17], [21]–[25]. In the context of this study, termination of a subscriber is defined as only quitting a service from the mobile operator. There is no other competitor to leave for in the case of monopoly market. The main feature that distinguishes churn prediction analysis from service termination prediction analysis is *On-net calls*. "*On-net calls*: describe monthly outgoing and incoming calls within the network. It includes duration of calls, number of calls, and the number of SMS sent or received." *Off-net calls* differ since the calls made are with another network (service provider) [3]. They are mainly used in the existing of other network(s) or competitors. Otherwise, all features and methods used in churn prediction can also be applied for prediction of subscribers' termination.

Unlike postpaid subscribers, the main challenge in predicting the termination of prepaid subscribers is that there is no binding contract requiring a prepaid subscriber to notify the operator before leaving a service [26]. However, telecom operators can also be benefited from the available enormous customers' data to predict defection by applying learning algorithms. In fact, the competition market across the world has forced service

providers to employ the best data mining algorithms that produce the most accurate prediction in order to remain on the market.

Predicting subscribers defection in advance provides an organization with valuable insight to retain and improve its customer base [1], [2]. The optimal size of the retention program takes into account the balance that the firm must make between increasing the target size to reduce defective losses and lowering the target size to reduce the cost of retention activities. Predicting defective subscribers is a critical and preliminary task ahead of retention activities. Retention practices are heterogeneous across the customer base as customers have different tenure, profile and service usage [16].

Learning algorithms contribute a lot to different applications, some of them are text or document classification, natural language processing to avoid stenography (shorthand writing), speech recognition, Optical character recognition (OCR) to read images that consist of typed, handwritten or printed content of text that are readable for humans but not for machines; computational biology applications, fraud detection, network intrusion detection, games, medical diagnosis, recommendation systems, search engines, information extraction systems, computer vision tasks such as image recognition, face detection and many more [27]. To predict customer defection, several machine-learning algorithms can be used. Since machine learning is not a one-time task, but instead a continuum, testing with different machine learning techniques and tuning parameters is the basis to avoid complexity and achieving the desired result. Parameter tuning on all machine learning techniques helps to create a clear model, to achieve the best performance and accuracy in order to predict based on a given data set [28], [29].

Based on subscribers' telephone numbers taken from a wireless telecom company in Taiwan a data set of about 160,000 subscribers, including 14,000 churners had been randomly selected by [21]. Customers' behavior extracted from customer demographics,

billing information, contract/service status, CDR, and service change log were able to deliver accurate churn prediction models using decision tree and neural network techniques. Before learning algorithms, customers were clustered into five segmentations (using K-means clustering) based on their loyalty, contribution, and usage. They got possible variables from other researches and telecom experts' interviews; then analyzed these variables. As a result, the performance of the neural network (backpropagation network, BPN) was better than the decision tree model without segmentation on both hit ratio and capture rate. But the performance of the decision tree model *without segmentation (without clustering)* was better than applying *segmentation (clustering)* on the dataset.

Researchers in [14] identified the most accurate ML algorithm to predict churners from their chosen ML algorithms. A CDR data of 98,514 subscribers for 183 days had been used for the research. A tree-based method was used to improve selection of features. As a result, SVM scored the highest accuracy of 89.4, whereas other algorithms scored: RF 88.4, KNN 88.2, AdaBoost 89.2 and Logistic Regression 89.3 in percent. Thus, the accuracy of an algorithm has implied the ability to identify churners.

Most studies in telecom industries have a focus to predict churners. However, Adeolu O. Dairo et al. [26] proposed a model to predict dormant customers. The researchers used CDR attributes and demographic variables of dormant customers to build the dormancy prediction model using a decision tree. Attributes were tested by using Oracle Data Miner (ODM) decision tree technique. Even though it had scored 79 percent at maximum, which is the least accurate compared with the churn prediction models' accuracy, the approach was unique and magnificent. It contributed to prepare a timely retention plan before customers go to the dormancy (suspension) stage. It is an easily adopted and applied model for the campaign management team to prevent dormancy and churners as well.

A. K. Ahmad et al. [18] analyzed a high volume and variety of data by ordinary analytical tools to present churn prediction model using big data analytics. To overcome a problem in getting customers' information (variety of data) from multiple systems and databases, different type of files structured, semi-structured (XML-JSON) or unstructured (CSV-Text) were extracted from SyriaTel telecom company. The research also dealt with a high volume of data related to customer's services and contact information, location information, complaints database, network logs data, CDR and mobile IMEI information are included in the study. The study used about ten million customers' information of 9 months (about 70 Terabytes on HDFS) to train, test, and evaluate in the Spark environment. The data also comes very fast and needs a suitable big data platform to collect, aggregate and extract features for each customer. XGBOOST algorithm were found as the best classifier compared to other algorithms *GBM tree algorithm*, *Random Forest* and *Decision Tree*.

Predicting behavior of customers (subscribers) is mainly based on records in the systems of a service provider. ethio telecom keeps record of subscribers in disintegrated systems such as ECAF, IPCC, CRM, MVAS, log file data, network systems (such as PRS, U2000) and other systems. It seems inapplicable to research the behavior of customers by collecting, integrating and processing all subscriber records. It requires a high storage device (central data warehouse) and an efficient processing machine. Thus, a few features that mostly related to predicting termination of subscribers have to be wisely selected. Based on works of literature, including mentioned above, CDR data hold most features that researchers commonly and necessarily used.

1.6 METHODOLOGY

Supervised learning approach, one of the machine learning techniques, is selected in order to build and test an effective and efficient model to predict subscribers' termination.

Classification algorithms that are mostly supervised learning used to predict subscribers' termination (deactivation). Three the most common classification algorithms, *J48 decision tree*, *RandomForest*, and *Naïve Bayes*, are applied to develop subscribers' termination model. The training and validation are made based on two validation methods as shown in Figure 1.1.

The steps followed to perform this study are explained as follows:

1. Literature Review

Works of literature were exhaustively reviewed from research articles, journals, books, and company white papers.

2. Data Collection

A formal supporting letter was taken from the Addis Ababa Institute of Technology (AAiT) then delivered to the Chief Human Resource Officer of ethio telecom. Then, relevant data was taken from system databases in IS division. Consequently, prepaid mobile subscribers profile and CDR data for voice, data, and SMS were collected.

3. Data Preprocessing

Aggregating data extracted in .csv format from systems, clearing missed values, removing unnecessary & highly correlated attributes and labeling instances were the main tasks at this stage. Based on proportionality of the labelling and the way the data was arranged; three data sets have been prepared. All data sets were finally converted to an appropriate format before ML algorithms have been applied.

4. Model Experiments

The experiment was made using the open-source "WEKA" machine learning tool. Classification algorithms were tested in six different scenarios. Scenarios were created based on three different data sets and two different validation methods (testing options): *percentage splits (70/30)* and *cross fold (k = 10)*. Algorithms' effectiveness and efficiency could be checked across data sets and validation methods.

5. Result Analysis and evaluation

Results were compared and contrasted so as result variation in each scenario were presented and discussed according to data set compilation and methods used. Finally, the results were evaluated based on their accuracy, performance, error measurement and interpretability metrics.

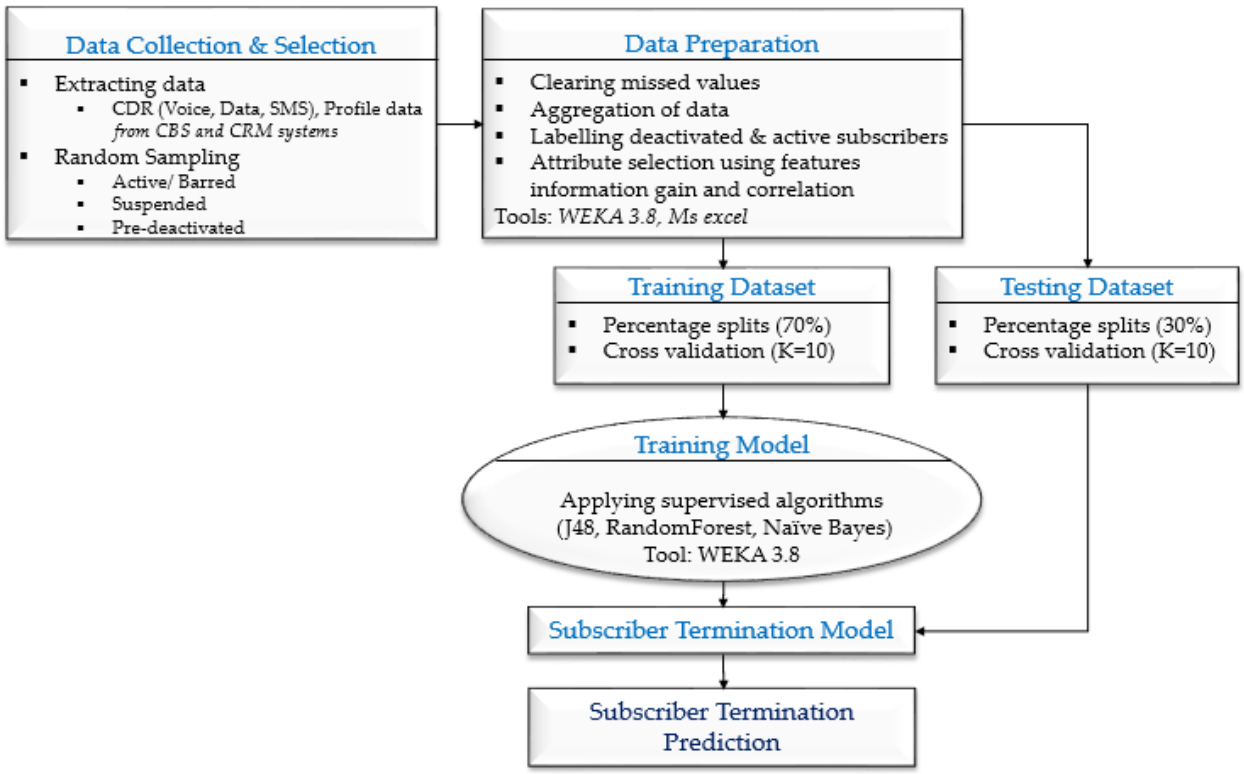

Figure 1.1 Process to Predict Prepaid Mobile Subscribers' Termination

1.7 THESIS ORGANIZATION

The remaining part of this paper is organized as follows: In Section 2, subscribers churn and service termination are briefly described. Section 3 service termination in ethio telecom trend are reviewed. The machine learning concept and selected classification algorithms for this study are thoroughly explained in Section 4. Section 5 explains how data is collected and prepared in different data sets; and how the prediction of ethio telecom subscribers' termination is done in different experiment approaches by putting evaluation metrics to compare algorithms each other. In Section 6 results are shown and discussed in brief and finally, the conclusion of the paper has been presented.

2 CHURN AND SERVICE TERMINATION

2.1 OVERVIEW

Customer churn sometimes known as customer attrition, customer defection or customer turnover. Most works of literature agreed up on churn in the terms of telecom industry that it is the loss of existing subscribers to another service provider [19]. Sometimes the definition of churn is put simply as the loss of customers [3]. Unlike churn, service termination is defined in this study as quitting of a service without leaving to competitors. In fact, in monopoly market customers (subscribers) has no option to migrate for rather they only quit. Besides, in this study prediction of subscribers' service termination and churn have common benefits. Additionally, the way subscribers terminate or churn are highly similar. Any user of operator services is considered as a customer whereas a customer with subscription is referred as a subscriber [30].

Subscribers could terminate from the service either voluntarily or involuntarily. In both ways prepaid mobile subscribers are not expected to notify their service providers before quitting. Thus operators become uncertain about their customer base which is the most necessary component to put a strategic plan. However, service providers have an enormous data to learn about their subscribers behavior such as: customer demography, billing data, Calls/SMS/Internet, international calls, on-net calls, complaints, network data and so on [18], [23], [31]. Predicting churners or mobile subscribers' termination has a lot of benefits that will be explained later.

Machine learning can be easily applied to support the detection, classification and prediction of termination of subscribers. By analyzing huge data at a service provider's hand, it is possible to identify defective subscribers early. Many prediction techniques can be used to find patterns from the actual data in a system that help predictive analysis

and classification process. The operator can therefore use a proactive retention operation or other approaches to be ready for the consequences after loss of subscribers [18], [24]–[25].

2.2 TYPES OF CHURN OR SERVICE TERMINATION

The customer life cycle provides another dimension to understand customers. This focuses specifically on the business relationship, based on the observation that the customer relationship evolves over time. Although each business is different, the customer relationship places customers into five major phases: prospects, responders, new customers, established customers and former customers [32].

According to [2], [33] a subscriber can leave the operator's service either voluntarily or involuntarily. Leaving voluntarily is also separated into two sub-categories that are deliberate and incidental as shown in Figure 2.1. Deliberate leaving could be occurred based on factors related to subscriber's sensitivity compared to subscriber's expectations such as high service price, low service quality, poor customer service, and other related inconvenience factors. Incidentally leaving service is not caused by subscriber's plan, but because of unexpected changes that could occur on subscriber life, such as a change in financial situation, changing home or work location (geographically relocated), time inconvenience, and others. Subscribers might leave a service they are using involuntarily if they are removed by the service provider itself due to deception, unpaid bills (loans), and non-usage of the service in a given period of time.

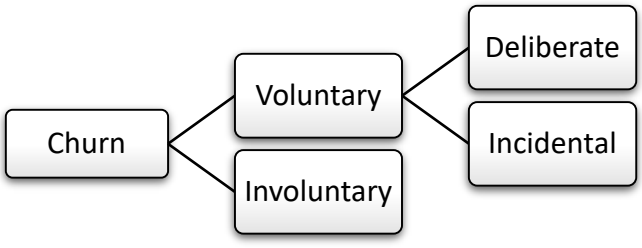

Figure 2.1 Types of Churn or Service Termination

2.3 BENEFITS OF PREDICTING CHURN OR SERVICE TERMINATION

Churn prediction and management have become of great concern to the mobile operators. Because of the importance of accuracy in churn prediction, a lot of researches has been directed into improving prediction model to identify churners [23]. Mobile operators want their subscribers to be maintained to meet their desires. They need to anticipate the possible churners by using their limited resources to maintain their customer base. Acquiring a new customer is 10 to 25 times more expensive than retaining and also reducing churn in just 5% can boost profitability by 75%. Improving retention has a greater impact on growth than acquisition. The probability of up selling or cross selling to an existing customer is 60-70%, but only 5-20% for a prospect [4].

Prediction of services termination in monopoly market has also similar benefits as churn prediction serves. It helps to send targeted offers to the concerned potential customers to improve their usage. Additionally, companies can gain an input from the prediction to:

- manage their customers proactively
- to improve their expectation on ARPU
- to prepare an appropriate marketing strategy and retention programs

3 MOBILE SERVICE TERMINATION IN ETHIO TELECOM

3.1 IDENTIFYING SERVICE TERMINATION IN ETHIO TELECOM

Before leaving telecom service(s), a postpaid subscriber should notify ethio telecom shops; otherwise, will be liable for any debt or failure under enforcement of the law. But most of the time, postpaid subscribers apply to quit service(s) so that their reason for leaving is registered on a CRM system. This helps the company to identify its subscribers interest and the most compliant area. Thus, a retention plan can be prepared for subscribers based on their interest so that maintaining the customer base becomes easy.

The challenge comes when prepaid customers quit; as they don't notify ethio telecom when they leave services. Actually, ethio telecom has a Customer Retention and Loyalty Section that has more than 60 employees. This section mainly involved in:

- Identifying the reasons why prepaid customers quit a service
- Confirming customers fault & other complaints are handled on time
- Identifying and awarding loyal customers
- Churn prediction

Data extracted and sent from IS division about terminated (deactivated) subscribers is used to identify their reason to leave a service. Employees used MS Excel to categorize and analyze subscribers' information. Identifying a reason of withdrawal of prepaid subscribers is done via a contact of another subscriber registered as an emergency contact for the terminated subscriber. Such action has a challenge for employees in investigating the reason through another customer. The contact number sometimes don't answer the call, becomes offended or may not have a willingness to answer the questionnaire.

3.2 SUBSCRIBERS TERMINATION RATE IN ETHIO TELECOM

Even though ethio telecom put an effort to identify and hold subscribers in the network, a significant number of mobile subscribers are terminated each month as it is shown in Table 3-1.

Table 3-1 Monthly Termination Rate in ethio telecom

Individual Mobile Subscribers Termination Rate in ethio telecom

Month	Category	Active	Sales in the Month	Termination Rate (Turnover Rate)	Number of Terminated Subscribers
Oct-18	Postpaid	54,161	346		
	Prepaid	37,280,894	518,810		
	Total	37,335,055	519,156		
Nov-18	Postpaid	54,902	1,438	1.29%	697
	Prepaid	37,469,103	492,681	0.82%	304,472
	Total	37,524,005	494,119	0.82%	305,169
Dec-18	Postpaid	55,392	856	0.67%	366
	Prepaid	37,805,636	484,484	0.39%	147,951
	Total	37,861,028	485,340	0.40%	148,317
Jan-19	Postpaid	56,418	1,489	0.84%	463
	Prepaid	38,167,654	497,407	0.36%	135,389
	Total	38,224,072	498,896	0.36%	135,852

Subscribers termination (turn over) rate of a month is determined by taking the difference in number of subscribers within a month (by excluding new subscribers) and dividing it to the number of subscribers at the beginning of the month. It is also possible to calculate the termination rate quarterly, semi-annually or annually accordingly. ethio telecom's termination rate from November 2018 to January 2019 is 1.6 percent based on the customer base. It becomes significant while compared to the new sales in the duration which is 39.9 percent.

3.3 SUBSCRIBERS LIFE CYCLE IN ETHIO TELECOM

In ethio telecom a subscriber should go through in either, few or all of the steps in the rules of a prepaid life-cycle subscription [34]. When SIM cards available to the market stays in a *pre-active* status until a subscriber fills out Customer Application Form (CAF) to take them. At this stage, the sim card has an unlimited validity period. When a customer is subscribed to get the service, the sim card status immediately changed to active. It becomes *active* after the first call is made. The validity period will last according to the air time hold/ purchased by the subscriber. The sim status can be changed to *One-way block* /*Barring*/ status if the probation period given to air time is ended. The sim stays for a given period of time at this stage; and will be transferred to *Two way blocked* /*Suspend*/. If a subscriber doesn't charge the balance within a month it will be terminated and recycled for a new sale. The detail subscriber life cycle in ethio telecom is shown in Appendix 1.

4 MACHINE LEARNING

4.1 OVERVIEW

Machine learning is emerging as one of the software industry's most important developments. While it has been around this advanced technology for decades, it is now becoming commercially viable. The world is moving into an age in which machine learning strategies become essential tools for creating value for businesses wanting to understand the hidden value of their data [28]. In [27] machine learning defined as computational methods using the experience to improve performance or to make accurate predictions. In 1959, Arthur Samuel defined machine learning as a "Field of study that gives computers the ability to learn without being explicitly programmed". In 1998, Tom M. Mitchell provided a widely quoted, more formal definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [35]. As per [36], machine learning usually refers to the changes in systems that perform tasks associated with Artificial Intelligence (AI). Such tasks involve recognition, diagnosis, planning, robot control, prediction, etc.

When to Use Machine Learning?

Machine learning has an interesting role to analyze tasks that are sophisticated to interpret. Predicting the label of a document, also known as document classification, is by no means the only learning task. Machine learning admits a very broad set of practical solutions to solve many analytical challenges.

Some tasks need the support of machine learning since they are very difficult to understand and analyze in usual methods for example speech or natural language recognition, text or document classification, computer vision application such as optical

recognition, game decisions, image identifier, and so on. Additionally, unless machine learning algorithms are used to learn from experience, it is sometimes impossible to extract meaningful results from input data.

Furthermore, machine learning is helpful for very large and complex data set that is highly interrelated with its features and beyond human capabilities to analyze such as astronomical data, turning medical archives into medical knowledge, weather prediction, analysis of genomic data, web search engines, and electronic commerce. The other reason to use is programmed tools are mostly inflexible to adapt their input data accordingly, but machine learning is adaptive for many tasks that can possibly change timely or from one user to another [27] [37].

Learning algorithms learn from the data itself and can put a possible prediction based on their experience. As they got big data sufficiently, they can predict more accurately. According to [21], data processing and storage approaches are facing many challenges in meeting the continuously increasing demands of big data in the real world. Challenges and solutions have four dimensions: data storage, analytics, online processing, and security. Learning tools like WEKA are limited in storing and processing big data [22]. For such a case Hadoop, Apache Spark, Python and other big data analytical tools have to be used.

4.2 MACHINE LEARNING CYCLE

In [28] creating a machine learning application or operationalizing a machine learning algorithm is an iterative process. It can't be simply trained a model once and leave it alone since data changes, preferences evolve, and new challenges will emerge. Therefore, the model should be kept fresh when it goes into production. While it is not possible to do the same level of training that was needed when the model is built, it can't be assumed that it will be self-sufficient.

The machine learning cycle is continuous, and choosing the correct machine learning algorithm is just one of the steps. The steps in the machine learning cycle are as follows:

- Identify the data: Identifying the relevant data sources is the first step in the cycle. In addition, as the machine learning algorithm is developed, it must be considered about expanding the target data to improve the system.
- Prepare data: Making sure the data is clean, secured, and governed. If a machine learning application is created based on inaccurate data, the application will fail.
- Select the machine learning algorithm: Several machine learning algorithms can be applied to the prepared data and business challenges.
- Train: The algorithm has to be trained to create the model. Depending on the type of data and algorithm, the training process may be supervised, unsupervised, or reinforcement learning.
- Evaluate: Evaluate your models to find the best performing algorithm.
- Deploy: Machine learning algorithms create models that can be deployed to both cloud and on-premises applications.
- Predict: After deployment, start making predictions based on new, incoming data.

Assess predictions: Assess the validity of predictions; the information gathered from analyzing the validity of predictions is then fed back into the machine learning cycle to improve accuracy.

4.3 APPLICATIONS OF MACHINE LEARNING ALGORITHMS

Machine learning algorithms are applied for most prediction problems, and the following problems are found in the real world but not the only [27], [38], [39]:

- Text or document classification: This includes problems such as assigning a topic to a text or a document, determining automatically if the content of a web page is inappropriate or too explicit; it also includes spam detection.

- Natural language processing (NLP): Most tasks in this field, including part-of-speech tagging, named-entity recognition, context-free parsing, or dependency parsing, are cast as learning problems. In these problems, predictions admit some structure. For example, in part-of-speech tagging, the prediction for a sentence is a sequence of part-of-speech tags labeling each word. In context-free parsing, the prediction is a tree. These are instances of richer learning problems known as structured prediction problems.
- Speech processing applications: This includes speech recognition, speech synthesis, speaker verification, speaker identification, as well as sub-problems such as language modeling and acoustic modeling.
- Computer vision applications: This includes object recognition, object identification, face detection, OCR, content-based image retrieval, or pose estimation.
- Computational biology applications: This includes protein function prediction, identification of key sites, or the analysis of gene and protein networks.
- Many other problems such as fraud detection (e.g. sim box fraud detection) for credit card, telecom or insurance companies, network intrusion, learning to play games such as chess, unassisted control of vehicles such as robots or cars, medical diagnosis, the design of recommendation systems, search engines, or information extraction systems, are tackled using machine learning techniques.

4.4 TYPES OF LEARNING

Machine learning scenarios differ from each other based on the types of training data available to the learner. Common machine learning types are listed as follows [27]:

Supervised learning: The learner receives a series of examples labeled as training data and makes predictions for all unknown levels that are most common for classification, regression, and ranking problems.

According to [24] it is possible to understand supervised learning in an even better way by looking at it through two types of problems.

Classification: Classification problems categorize all the variables that form the output. Examples of these categories formed through classification would include data such as marital status, sex, age, location, contract type and so on. The most common model used for this type of service status is the support vector machine. The support vector machines set forth to define the linear decision boundaries.

Regression: Problems that can be classified as regression problems include types where the output variables are set as a real number. The format for this problem often follows a linear format.

Unsupervised learning: the learner only receives unlabeled training data and makes predictions for all unknown levels. Since there is usually no labeled example available in this setting, it can be difficult to quantitatively measure the quality of the learner. Clustering and reducing dimensionality are examples of unsupervised learning problems.

Semi-supervised learning: The learner receives a learning sample data consisting of both labeled and unlabeled and makes predictions for all unknown levels. This learning is common in settings where unlabeled data is easily accessible but labels are expensive to obtain. Various types of problems arising in applications, including classification, regression, or ranking tasks, can be described as semi-supervised learning instances. The

idea is that the distribution of unlabeled data accessible to the learner can help to achieve a better performance than in the supervised setting.

Reinforcement learning: The training and testing phases are also intermixed in reinforcement learning. The learner frequently communicates with the environment to collect information, influencing the environment in some situations, and obtaining an immediate reward for each action. The learner's goal is to optimize its reward with the environment over a series of actions and iterations. Nevertheless, the environment does not provide long-term incentive input, and the learner faces the dilemma of discovery and exploitation, as it has to choose between discovering unknown acts to gather more information or exploiting the already collected information.

Online learning: it involves multiple rounds and intermixes training and testing phases. The learner gets an unlabeled training point at each round and forecasts then receives the true mark, and incurs a loss. The aim of the online setting is to minimize aggregate loss throughout all rounds. In comparison to the previous settings described above, online learning does not allow any distributional inference. In addition, in this case, instances and their labels may be chosen as adversarial.

4.5 SUPERVISED LEARNING

The learning of the classifier is "supervised" in that it is told to which class each training tuple belongs. It contrasts with unsupervised learning (or clustering), in which the class label of each training tuple is not known, and the number or set of classes to be learned may not be known in advance [40].

According to [35], a given problem of supervised learning is solved by performing the following steps:

- Determine the type of training examples: Before doing anything else, the user should decide what kind of data is to be used as a training set.
- Gather a training set: The training set needs to be representative of the real-world use of the function. Thus, a set of input objects is gathered and corresponding outputs are also gathered, either from human experts or from measurements.
- Determine the input feature representation of the learned function: The accuracy of the learned function depends strongly on how the input object is represented. Typically, the input object is transformed into a feature vector, which contains a number of features that are descriptive of the object.

4.5.1 J48 Decision Tree

Decision trees (DT) are well known for their simplicity and understandability. The DT is produced by algorithms that identify various ways of splitting a data set into branches (segments) [29]. Decision trees are highly effective tools in many areas such as data and text mining, information extraction, machine learning, and pattern recognition [41]. It is a classifier expressed as a recursive partition of the instance space.

The DT consists of nodes that form a Rooted Tree, meaning it is a Directed Tree with a node called root that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal node or test nodes. All other nodes are called leaves (also known as terminal nodes or decision nodes). In the decision tree, each internal node splits the instance space into two or more subspaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range.

Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target value having a certain value. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. Typically, the complexity of the tree is measured by the following metrics: the total number of nodes (tree size), total leaf, tree depth and the number of attributes used [42].

Entropy

Entropy is used to measure information content by calculating estimates the information gain from a feature. The feature subset that yields the maximum entropy reduction which is equivalent to maximum information gain will be selected. The reduction in entropy of the class given a feature is defined as the difference between the prior entropy and the posterior entropy after observing a feature. For homogenous feature values, Entropy will be zero, whereas for equally divided feature values Entropy will be one [22], [43]. The entropy of a discrete random variable is defined as shown in Equation (4.1) below:

$$E = -\sum_{y \varepsilon Y} P(y)\, log\ P(y) \tag{4.1}$$

The entropy of a random variable after observing a random variable is defined as shown in Equation (4.2).

$$E = -\sum_{x \varepsilon X} P(x) \sum_{y \varepsilon Y} P(y|x)\, log\ P(y|x) \tag{4.2}$$

Information Gain

The information gain is the difference between the entropy of classification label before and after the split and is achieved by replacing the function E in the previous expressions by the entropy function as shown in equation (4.3)

The information gain of Y given X.

$$InfoGain\ (Y,X) = Entropy(Y) - Entropy(Y|X)$$

$$InfoGain\ (Y,X) = Entropy(X) - Entropy(X|Y)$$

$$InfoGain\ (Y,X) = Entropy(Y) + Entropy(X) - Entropy(X|Y) \quad (4.3)$$

Pruning

Pruning is a technique in device reading that reduces the dimensions of decision trees by detaching parts of the tree that provide little control to categorize instances. A key motivation of pruning is ”trading accuracy for simplicity” [35], [44], [45].

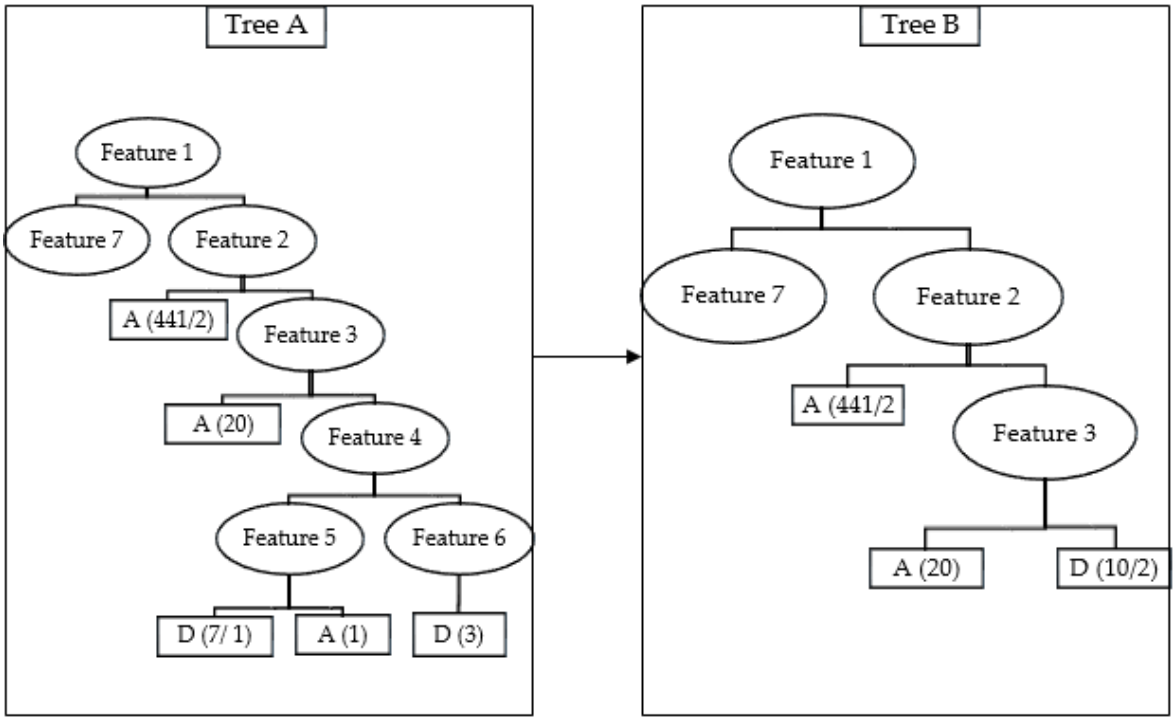

Figure 4.1 Pruning

The major advantage of generalization is creating simple visualization and clear evaluation for an easy decision, but it has a trade-off on the accuracy of the output [27]. The minimum number of instances indicates the minimum amount of data separation per branching. Instances per leaf in Figure 4.1. are limited to 10 so that all nodes in Tree A that hold less than 10 instances will be merged into their parent nodes as shown in Tree

B. Feature 5 and Feature 6 are merged to Feature 4 since they hold leaves less than 10 instances.

4.5.2 RandomForest

Random Forests are an ensemble learning method that is used for classification as well as regression; but however, it is mainly used for classification problems. As a forest is made up of trees, in *Random Forest* more trees mean more robust forest. Similarly, a *Random Forest* algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It uses an ensemble method which is better to reduce over-fitting by taking an average of the result [35], [46].

The designed strategy used in *Random Forest* is divide and conquer. It forms a number of DT and each DT is trained by selecting any random subset of attributes from the whole predictor attribute set. Each tree will grow up to a maximum extent based on the attribute present in the subset. Then, based on the average or weighted average method, the final DT will be constructed for the prediction of the test data set [20].

Random Forest algorithm follows some steps in the prediction process. First, start with the selection of random samples from a given data set. Next, this algorithm will construct a decision tree for every sample. After getting the prediction result from every decision tree, voting will be performed for every predicted result. At last, the most voted prediction result will be selected as the final prediction result.

Random Forest runs efficiently in a large range of data than a single decision tree does. It can handle thousands of input variables without variable deletion. It also handles the missing values inside the data set for training the model. Handle the unbalanced data set by using *Random Forest* is difficult. *Random Forest* algorithm is very flexible and possesses very high accuracy. It overcomes the problem of overfitting by averaging or combining the results of different decision trees. It maintains good accuracy even a large proportion

of the data is missing. Scaling of data does not require in *Random Forest* algorithm since it maintains good accuracy even after providing data without scaling.

Complexity is the main disadvantage of *Random Forest* algorithms. It is much harder and time-consuming than decision trees. It is less intuitive in case there is a large collection of decision trees and also requires more computational resources to implement.

4.5.3 NaïveBayes

[47] *Naive Bayes* algorithm is a classification technique based on applying Bayes' theorem with a strong assumption that all the predictors are independent of each other. In simple words, the assumption is that the presence of a feature in a class is independent of the presence of any other feature in the same class. For example, a customer may be considered as active if the service number hold by the customer has high voice usage, high sms usage, high data usage, high tenure, etc. Though all these features can be associated each other, they contribute independently to the probability that the customer is an active user.

In Bayesian classification [20] the main interest is to find the posterior probabilities i.e. the probability of a label given some observed features, $P\ (L|features)$. With the help of Bayes theorem, we can express this in quantitative form as shown below in Equation (4.4).

$$P(L|features) = \frac{P(L)P(features|L)}{P(features)} \tag{4.4}$$

Where, (L|features) is the posterior probability of class label L the attribute value "features"

$P(L)$ is the prior probability of class label "L".

$P(features|L)$ is the probability of the attribute value "features" when the given class label is "L".

$P(features)$ is the prior probability of predictor "features".

[47] *Naive Bayes* classification can be used for binary as well as multi-class classification problems. It is easy to implement and faster to converge than discriminative models like logistic regression. Since it's highly scalable in nature it requires less training data; it can also make probabilistic predictions and can handle continuous as well as discrete data. Bayesian classifier with decision tree and chosen neural network classifiers to be comparable in performance [40].

One of the disadvantages of *Naive Bayes* classification is its strong feature independence because in real life it is almost impossible to have a set of features that are completely independent of each other. Another issue with *Naive Bayes* classification is its 'zero-frequency' which means that if a categorical variable has a category but not being observed in the training data set, then the *Naive Bayes* model will assign a zero probability to it and it will be unable to make a prediction.

4.6 UNSUPERVISED LEARNING

Unsupervised learning is essentially a synonym for clustering. The learning process is unsupervised since the input examples are not class labeled [40]. Most of the unsupervised learning methods use a measure of similarity between patterns in order to group them into clusters. The simplest of these involves defining a distance between patterns. For patterns whose features are numeric, the distance measure can be ordinary Euclidean distance between two points in n-dimensional space [36].

The literature includes several clustering algorithms, but it is not the focus area of this paper to discuss them all in detail. Nonetheless, providing a fairly ordered image of clustering methods is useful. The most basic clustering methods can usually be grouped into the following categories based on their general characteristics shown in Table 4-1. A separate categorization of clustering methods is difficult to provide because these categories often overlap so that a system can have features from multiple categories.

Table 4-1 Characteristics of Clustering Methods

Method	General Characteristics
Partitioning methods	- Find mutually exclusive clusters of spherical shape - Distance-based - May use mean or medoid (etc.) to represent the center - Effective for small to medium size data sets
Hierarchical methods	- Clustering is a hierarchical decomposition (i.e., multiple levels) - Cannot correct erroneous merges or splits - May incorporate other techniques like micro clustering or consider object "linkages"
Density-based methods	- Can find arbitrarily shaped clusters - Clusters are dense regions of objects in space that are separated by low-density regions - Cluster density: Each point must have a minimum number of points within its "neighborhood" - May filter out outliers
Grid-based methods	- Use a multiresolution grid data structure - Fast processing time (typically independent of the number of data objects, yet dependent on the grid size

4.7 PARAMETER SETTING FOR PRUNING

Decision trees are simple to understand and interpret, and require less data preparation than many other types of classifiers. They are also generally fast, even on large data sets, and can create non-linear decision boundaries that fit data very well. In fact, overfitting is a prime risk with decision trees, so it is vital to check the model before deploying it to the test set. As a decision tree is grown, it naturally tends to over fit the data.

The pruning of decision trees is a crucial step towards maximizing the computational performance and the accuracy of the classification of such a model. Applying pruning methods to a tree typically leads to a reduction in tree size to avoid unnecessary complexity and to avoid over-fitting of the data set as new data is classified [45]. Pruning is a process by which the largest tree that is the most generalizable is selected, and all the branches below that level are pruned out.

J48 Parameters

J48 has the following major parameters to overcome overfitting problem [48], [49].

- *BinarySplits*: It determines how nominal data should be used with binary splits. This is a method whereby the tree is developed by considering one nominal value over all other nominal values rather than separately considering a break on each nominal value. This refers to a tree where from each node there are only two branches.
- *confidenceFactor*: This determines the aggressiveness of the pruning process. The higher this value, the more confident it will be that the data applied for learning will have a good representation of all possible events, but there will be less pruning. Smaller values induce to more pruning. It significantly affects the efficiency of the classifier.
- *minNumObj*: It determines what is allowed for the minimum number of observations on each tree leaf. This is another way of overfitting control.
- *numFolds*: Determines the amount of data used in reduced error pruning. One fold is used for pruning; rest is used for growing the tree.
- *SubtreeRaising*: This is a specific pruning method by which a whole set of branches are moved further down the tree to replace branches grown above it.
- *Unpruned*: This specifies if the tree should not be pruned.

Table 4-2 shows major parameters that would affect optimization while the *J48* classifier is used [48].

Table 4-2 Significant Parameters for Pruning in J48 Classifier

Parameter	Values range for adjustment	Default value
binarySplits	True/ False	False
confidenceFactor	[0.001, 0.5]	0.25
minNumObj	1, 2, 3, …., n	2
numFolds	1, 2, 3, …., n	3
subtreeRaising	True/ False	True
unpruned	True/ False	False

Random Forest Parameters

Random Forest parameters adjust the classifier to bring the desired accuracy, performance and visualization of the prediction. Some essential parameters including the description in WEKA 3.8 are shown with their default value in Table 4-3.

Table 4-3 Significant Parameters in Random Forest Classifier

RF Parameters	Description	Default Values
batchSize	The preferred number of instances to process if batch prediction is being performed.	100
computeAttributeImportance	Computes attribute importance via mean impurity decrease.	FALSE
doNotCheckCapabilities	If set, classifier capabilities are not checked before classifier is built (Use with caution to reduce runtime).	FALSE
maxDepth	The maximum depth of the tree, 0 for unlimited.	0
numIterations	The number of iterations to be performed.	100

Naïve Bayes Parameters

The parameters of Naïve Bayes do not rely on pruning but are mainly used to improve the performance and visualization of the classifier.

5 DATA ANALYSIS

5.1 DATA COLLECTION

All training and test data are collected from the ethio telecom data center. Three months of prepaid subscribers' CDR data and their profile information are extracted from CBS and CRM databases respectively. Fifty thousand and twenty thousand subscribers' data are requested from suspend/ pre-deactivate and active/ barred statuses respectively. Since active status subscribers have a significant number of service transactions, the required data have been minimized to balance the data extraction with suspend/ pre-deactivate subscribers.

5.2 DATA PREPROCESSING

5.2.1 Aggregation of Data

Features extracted from the two systems are saved in .CSV file format for further data preparation. The extracted data exceeds MS Excel sheet that has a capacity to hold around one million records only. Therefore, the condensed .csv files should be filed as much as extracted by using a splitter. All data have to be aggregated to a unique subscriber that is counted as an instance and saved in another file.

The three months of subscribers' data have been summarized based on unique service numbers in order to analyze subscribers CDR and profile. After a summary of subscribers' data averaged to a month for each unique subscriber tenure, average usage of voice, SMS, data and the average frequency of voice call and data usage that will be calculated and put in summary. Fields in the data are arranged appropriately based on their field type such as character, text, date, etc.

Before machine learning algorithms are used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned by removing inappropriate values and those with missed data [35].

5.2.2 Data Clearance

WEKA and MS Excel tool support in clearing missed values and removing unnecessary values from the data such as outliers, test service numbers, erroneously entered data like special symbols, missed values, null values, duplicated information, etc. Before feeding the collected subscribers data to the WEKA tool a data clearance is done using MS Excel and the remaining clearance is done using WEKA.

5.2.3 Attribute Selection

Features selection plays an important role in determining the performance of predictive models in terms of prediction rates. If a robust set of features can be extracted in this phase, the prediction rates can be significantly improved. However, obtaining such a good set of features is not an easy task. Most of the feature sets that have been proposed so far in the mobile telephony industry [21], [23] but still have room for improvement. Some steps are followed to filter out attributes that have a significant influence on the prediction to improve the prediction rates.

When attributes are selected, they should be evaluated relative to their power to give more information about the subscriber. In this study, outgoing CDR information and profiles are extracted from CBS and CRM systems respectively. Incoming CDR is not included since subscribers are limited by validation period. Validation period limits

subscribers' existence in the operator's system without using a service so long. Therefore, their behavior can be found in their outgoing usage.

Table 5-1 Features Extracted from CBS and CRM Systems

Features Selected				
Feature Set	**Feature Name**	**Calculation**	**Data Type**	**Behaviors to be learned**
Customer Profile	Activation Year	= *Year of activation date*	Numeric	Loyalty measurement
	Service Number (Serv_No)	-	Numeric	To identify the subscriber
	SIM status (Status)	-	Nominal	To rank or label subscribers' stage
	Service Type (Serv_Type)	-	Nominal	To identify the service a subscriber is registered
CDR	Voice Usage in Min (Voice_Usg)	$= \frac{Total\ voice\ duration}{Number\ of\ Calls}$	Numeric	Subscriber sensitivity to stay in a voice call
	Call Frequency (Voice_Freq)	$= \frac{Total\ number\ of\ calls}{Number\ of\ months}$	Numeric	To measure subscribers how often to make a voice call
	Voice Fee in Birr (Voice_Fee)	$= \frac{Total\ voice\ fee}{Number\ of\ months}$	Numeric	Subscribers sensitivity in spending money on voice call
	SMS Fee in Birr (SMS_Fee)	$= \frac{Total\ sms\ fee}{Number\ of\ months}$	Numeric	Subscribers sensitivity in spending money on SMS at all
	Frequency of Sending Messages (SMS_Freq)	$= \frac{Total\ number\ of\ sms}{Number\ of\ months}$	Numeric	Subscribers interest how often to send SMS at all
	Number of SMS sent to local numbers (SMS_Local)	$= \frac{Total\ no.of\ local\ sms}{Number\ of\ months}$	Numeric	Subscribers interest to send SMS for local mobile numbers
	Number of SMS sent to international numbers (SMS_Int)	$= \frac{Total\ no.of\ int.local\ sms}{Number\ of\ months}$	Numeric	Subscriber interest to send SMS for international mobile numbers
	Number of SMS sent to short code numbers (SMS_to_Code)	$= \frac{Total\ no.of\ int.code\ sms}{Number\ of\ months}$	Numeric	Subscriber interest to spend money on SMS for short code numbers
	Data usage in MB (Data_Usg)	$= \frac{Total\ data\ usage}{Number\ of\ months}$	Numeric	Subscriber interest to spend time on data usage
	Data Fee in Birr (Data_Fee)	$= \frac{Total\ data\ fee}{Number\ of\ months}$	Numeric	Subscribers sensitivity in spending money to use data
	Frequency of Data Usage (Data_Freq)	$= \frac{Total\ number\ of\ usage}{Number\ of\ months}$	Numeric	Subscriber interest to use data

Other personal profiles are not included since mostly they probably don't represent the person using the SIM card. Subscribers' profiles data cannot be fully trusted to study subscribers' behavior since people mostly share personal items including SIM cards [50]. The World Bank estimates about one billion people still lack official proof of identity [51]. Even though Ethiopia is one of the countries required to register SIM cards, the only country in Africa capturing and validating SIM users is Egypt, according to the GSMA 2018 study [52]. Features listed in Table 5-1 are extracted to learn subscribers' loyalty and usage behavior based on their CDR and profile.

It is important to pick the most related features to the target classification. In addition to improving classification performance, the prediction can be maximized by selecting attributes/ features based on certain evaluation techniques. Classifiers should also perform better than random assumptions [40]. There are various methodologies for assessing worthiness attributes that can be used depending on the type of information and attributes. Two of these methods have been used here, namely correlation and information gain ratio. Taking into account the data form of the attributes in this study, this choice is made; two nominal and the others left numerical. Details of how these techniques perform the evaluation are discussed here.

In memory of Karl Pearson, who originally developed it, the linear correlation coefficient is also referred to as Pearson's product-moment correlation coefficient. This statistic numerically represents the intensity of the straight-line or linear relationship between the two variables and the direction, whether positive or negative. Simply correlation is defined as the statistical association between two variables [53]. Therefore, the next step is to determine the worthiness of selected features in relation to the classification class.

Correlation: it evaluates an attribute's worthiness by measuring its linear relationship with the classification class. Pearson attribute correlation coefficient is calculated in Equation (5.1).

$$\frac{(\sum_{i=1}^{N} xy) - (\sum_{i=1}^{N} x)(\sum_{i=1}^{N} y)}{\sqrt{(\sum_{i=1}^{N} x^2 - (\sum_{i=1}^{N} x)^2)(\sum_{i=1}^{N} y^2 - (\sum_{i=1}^{N} y)^2)}} \quad (5.1)$$

Where, N: Number of instances in the given attribute
x: Instance values of the given attribute and
y: classification class label

Value of the coefficient (C) ranges between -1 and + 1. Values close to + 1 show a strong positive correlation, those close to -1 show strong negative correlation, and those closest to 0 show no relation. Since the correlation of Pearson establishes a linear relationship, the linear form is therefore considered to be analysed among the variables [54].

CorrelationAttributeEval function in WEKA evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class. Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average [55].

Information Gain Ratio (IG): Assess an attribute's worthiness by measuring an attribute's information gain ratio with the predefined classification class. IG measure is biased toward attributes with many values that help to avoid attributes with univariate bias in gaining information as it is explained in the previous section.

The data set used to evaluate attributes using worthiness evaluation techniques is shown below in Table 5-2.

Table 5-2 Data Set Used to Evaluate Attribute Worthiness

Status	Number of Subscribers
Active	19185
Deactivated (Terminated)	19185
Total	38370

Thirteen attributes, excluding the labelled (classification) class and service number, are ranked using the two attribute worthiness evaluation techniques as given in Table 5-3. Attributes that score a minimum value both in correlation and gain ratio compared and excluded from the data set.

Table 5-3 Attributes Worthiness

Attributes	Correlation	Rank by Correlation	Information Gain Ratio	Rank by IG
Activation Year	*0.37723*	*1*	*0.52765*	*1*
Voice_Fee	*0.32368*	*2*	*0.31474*	*3*
Voice_Usg	*0.28668*	*3*	*0.32051*	*2*
SMS_Fee	*0.21189*	*4*	*0.14542*	*7*
Data_Freq	*0.19523*	*5*	*0.15361*	*6*
Data_Fee	*0.17372*	*6*	*0.12889*	*8*
Data_Usg	*0.14095*	*7*	*0.12595*	*9*
SMS_Local	*0.08481*	*8*	*0.1991*	*4*
Serv_Type	*0.06046*	*9*	*0.0028*	*13*
Voice_Freq	*0.0363*	*10*	*0.15905*	*5*
SMS_to_Code	*0.02976*	*11*	*0.05382*	*11*
SMS_Int	*0.01605*	*12*	*0.0161*	*12*
SMS_Freq	*0.00466*	*13*	*0.0841*	*10*

Attributes with a better IG for prediction and correlation with the classification class are selected. The least predictor attributes in either of the measurements are removed. Therefore, seven attributes: activation year, voice fee, voice usage, SMS fee, data frequency, data fee, and data usage have been selected and will be combined with the classification class, Label. All of them have numerical data type whereas the classification class only has nominal data type with values 'A' and 'D' to classify active and deactivated subscribers respectively.

5.2.4 Data Labelling

All the data set is labeled based on the status of subscribers and it will be included in the data set as one feature. Making the data labeled helps to learn subscribers' behavior based on their labels; so that machine learning algorithms can learn and predict by comparing each subscribers' label with other features of every instance. Hence, prepaid subscribers with a status of *active* or *barred* are labeled as *Active Subscribers (A)*, whereas subscribers that are in *suspend*, *pre-deactivate* and *deactivate* status are labeled as *Deactivate (Terminate) Subscribers (D)* as shown below in Table 5-4.

Table 5-4 Data Labelling

Subscribers Status	Subscribers Category	Label
Active	Active	A
Barred		
Suspend	Deactivate (Terminate)	D
Pre-Deactivate (Termination)		
Deactivate (Recycling)		

5.3 EXPERIMENT APPROACH

Before applying machine learning algorithms, the data has to be arranged in different proportionality of *active* and *deactivated (terminated)* subscribers in the data set.

Datasets

There could be so many options to prepare possible data sets by changing the proportionality of active and deactivated subscribers' data size then after tuning the best learning algorithm by changing parameters that affect prediction effectiveness and efficiency. In this study, three types of data sets are prepared to check the learning impact on applied algorithms using 38370 instances from active and deactivated subscribers.

Table 5-5 Dataset Proportionality

Dataset	Proportionality
Dataset 1	1A:1D
Dataset 2	1A:2D
Dataset 3	2A:1D

The first data set is prepared in equal proportion from *active* and *deactivated* subscribers data set, while the second data set is prepared from *deactivated* subscribers who are twice *active* subscribers as well as *active* subscribers. The third data set is eventually prepared by *active* subscribers that doubled *deactivated* subscribers as shown here in Table 5-5.

Test Options

Two test options are used to test their impact on data selection for training and test purpose. These are Percentage Split and Cross-Validation:

Percentage Split: The solution proposed in this study divided the data into two groups: training data set and testing data set. The training data set consists of 70% of the total data set and aims to train the algorithms. The test group contains 30% of the data set and is used to validate the algorithms.

Cross-Validation: The second option is to use K-fold cross-validation with a value of k equals 10.

Experiment Scenarios

Algorithms go through the two testing options and the three data sets described above to find an effective learning algorithm. Upon comparing them in each case, there will be six different scenarios to choose the best algorithm as shown in Figure 5.1.

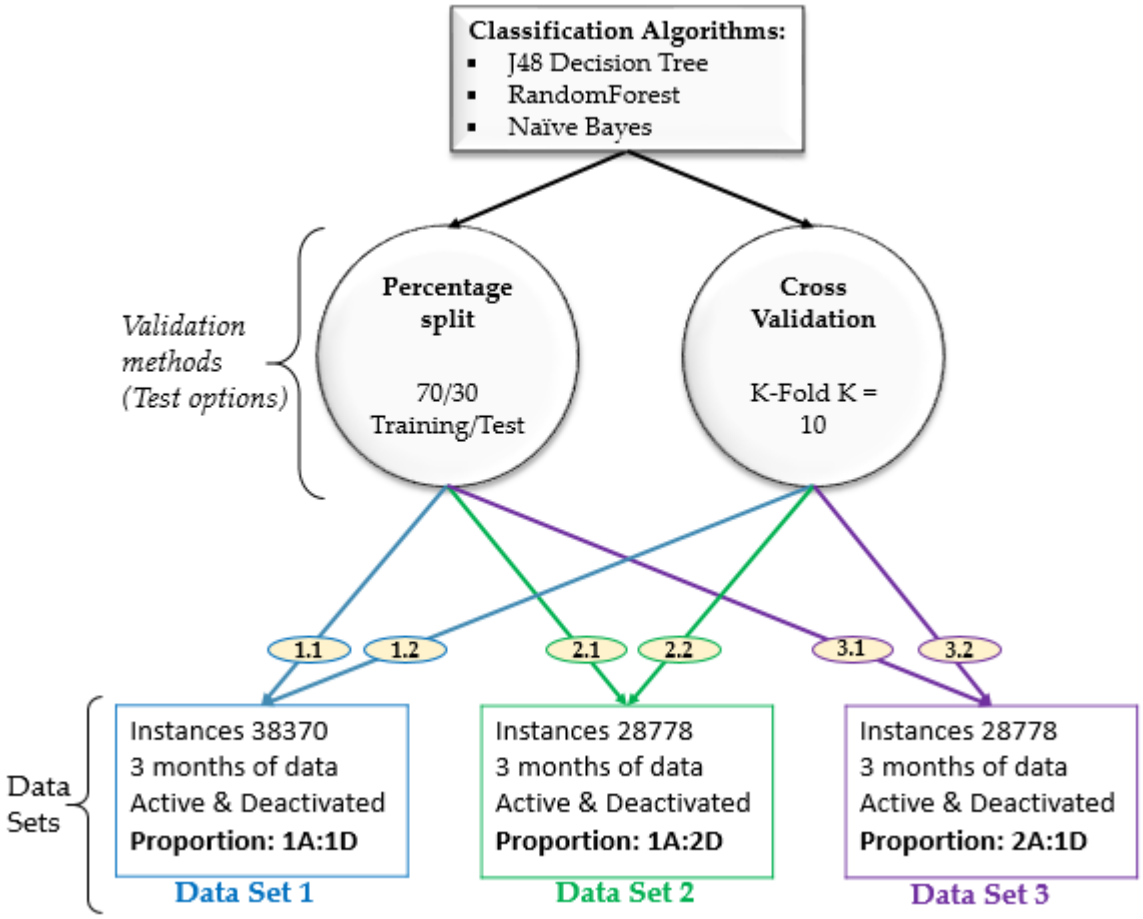

Figure 5.1 Experiments Scenario

5.4 Performance Evaluation Metrics

There are several standard metrics to evaluate classifiers' performance while predicting a termination of subscribers. Most commonly used metrics to measure and compare the effectiveness and efficiency of different classifiers on a given data set are listed here after based on [13], [20], [23], [35], [37], [40], [56]–[58].

5.4.1 Confusion Matrix

The literature proposes several standard performance metrics to compare the effectiveness of the various prediction classifiers. Evaluation metrics taken from confusion matrix can be used to measure the quality of any model built using both balanced and unbalanced data sets. The results of a classification (or clustering) algorithm

can be quickly visualized using a confusion matrix. There will be a two by two confusion matrix for a data set that has two labels for a decision making. Thus there will be four possible predictions as shown in Table 5-6:

Table 5-6 Confusion Matrix

ML Algorithm Result		**Predicted**		
		A	**B**	**Total**
Actual	**A**	**TP**	FN	m_1
	B	FP	**TN**	m_0
	Total	n_1	n_0	**n**

Where, A True Positive (TP): correctly predicts the positive value A.

A True Negative (TN): correctly predicts the negative value B.

A False Negative (FN): incorrectly predicts the positive value A.

A False Positive (FP): incorrectly predicts the negative value B.

Outcome Accuracy

Accuracy: is a measurement of correctly predicted instances from the total data and calculated as shown in Equation (5.2)

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (5.2)$$

Where, TP is the number of active subscribers,

TN is the number of deactivated subscribers,

FP is the number of deactivated subscribers predicted as active subscribers and

FN is the number of active subscribers predicted as deactivated subscribers.

Precision: It takes into account all positive predictions. Precision is the fraction of positively predicted points that are in fact positive as shown in Equation (5.3). The higher the precision is, the lower the number of false positive errors committed by the classifier.

$$Precision = \frac{TP}{TP+FP} \tag{5.3}$$

Recall: Fraction of positive instances correctly predicted by the classifier. Its value is equivalent to true positive rate. The higher the value of recall the fewer the number of instances misclassified as negative and calculated as shown in Equation (5.4).

$$Recall = \frac{TP}{TP + FN} \tag{5.4}$$

F1-Score: A combined measure or harmonic mean of precision and recall as calculated in Equation (5.5).

$$F1 = \frac{2*Precision * Recall}{Precision + Recall} = \frac{2TP}{(2TP + FN + FP)} \tag{5.5}$$

The main focus of this study is identification of subscribers' termination by improving precision and recall through reduction of FP and FN rates. Therefore, the effectiveness of classification algorithms will be measured using precision and recall.

Kappa: If the actual and predicted value randomly they would sometimes agree just by chance. Kappa gives a numerical rating of the degree to which this occurs. The calculation is based on the difference between how much it is actually predicted compared to how much expected as it is shown in Equation (5.8). The expectation and accuracy are calculated as shown in Equation (5.6) and Equation (5.7) respectively.

$$Expectation = Pe = [(n1/n) * (m1/n) + (n0/n) * (m0/n)] \tag{5.6}$$

Where, $n_1 = TP + FP$

$m_1 = TP + FN$

$n_o = FN + TN$

$m_0 = FP + TN$

$n = TP + FP + TN + FN$

$$Po = \frac{(TP + TN)}{n} \tag{5.7}$$

$$Kappa = \frac{(Po - Pe)}{(1 - Pe)} \tag{5.8}$$

5.4.2 Error Measurements

Error rate or misclassification rate of a classifier is calculated by subtracting rate of accuracy from one (100%) or can be found as shown in Equation (5.9).Other error measurements are calculated as shown in Equation (5.10), Equation (5.11), Equation (5.12) and Equation (5.13).

$$Error\ Rate = \frac{\text{FP+FN}}{\text{TP+TN+FP+FN}} \tag{5.9}$$

$$Mean\ Absolute\ Error = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{\theta}_i - \theta_i\right| \tag{5.10}$$

$$Root\ Mean\ Squared\ Error = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_i\right)^2} \tag{5.11}$$

$$Relative\ Absolute\ Error = \frac{\sum_{i=1}^{N}|\hat{\theta}_i - \theta_i|}{\sum_{i=1}^{N}|\bar{\theta} - \theta_i|} \quad (5.12)$$

$$Root\ Relative\ Squared\ Error = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}{\sum_{i=1}^{N}(\bar{\theta}_i - \theta_i)^2}} \quad (5.13)$$

Where, θ *: Actual value*

$\hat{\theta}_i$ *: Predicted value*

$\bar{\theta}$ *: Mean value*

5.4.3 Other Metrics

Classifiers can also be compared with the following additional factors in addition to accuracy-based measures: Speed, Robustness Scalability and Interpretability

Speed: it refers to the computational costs of creating and using the classifier.

Robustness: it is the classifier's ability to predict correctly given noisy data or incomplete information.

Scalability: it is the ability, despite large amounts of data, to construct the classifier efficiently. Usually, scalability is measured with a series of increasing data set size.

Interpretability: it refers to the level of understanding and knowledge the classifier or predictor offers. Interpretability is subjective, making it harder to evaluate. Decision trees and rules in classification can be easily interpreted, but their interpretability can diminish when they become more complex.

6 RESULTS AND DISCUSSION

Classification algorithms vary in their efficiency and effectiveness. Feature selection methods, composition of data set, data size, method of validation and related factors affect algorithm's accuracy and efficiency. In this analysis, three classification algorithms *J48*, *Random Forest* and *Naive Bayes* with two test options, *percentage split (70/30)* and *10 fold cross validation* are applied to three data sets prepared in different proportionality as previously shown in Table 5-5.

The results were analyzed here on the basis of the output in confusion matrix and time to build the model. Additionally, errors in prediction and interpretability of the model were considered in the comparison. A confusion matrix allows to get TP rate, FP rate, precision, recall, F-measure helps to see the detail. In addition, data set proportionality and validation methods are considered to see their effect regard to the accuracy and performance of classification algorithms.

Unlimited number of combination can be used to adjust classification algorithms parameter for prediction. But, their default values are used to restrict the infinite number of variations in classification algorithm parameter choices. Comparing them without further modification is rational. Therefore, analyses were carried out in WEKA (Version 3.8) using default parameters. Furthermore, after selecting the effective and efficient algorithm, its parameter can be modified until the possible desired accuracy, performance, and interpretability is obtained.

Accuracy

J48 and *Random Forest*, both scored their best at 93.3%, while *Naive Bayes* scored its best at 85.8%. All algorithms scored their most accurate prediction on *Dataset 3*. *J48* and *Naive*

Bayes have obtained their best outcome on percentage split method while *Random Forest* has been graded for its best on both data validation methods.

The composition of *Dataset 3*; a two-third of active subscribers and one-third of deactivated subscribers; is more suitable to learn much better about active subscribers compared to the other two data sets. *Random Forest* has got its best accuracy on the basis of true positive values that holds active subscribers. The precision and recall percent shown in Figure 6.1 and Figure 6.2 are as equivalent as the overall accuracy. Their F-measure in all scenarios is much closer to the overall accuracy, precision and recall as it is shown in Appendix 3, Appendix 4 and Appendix 5.

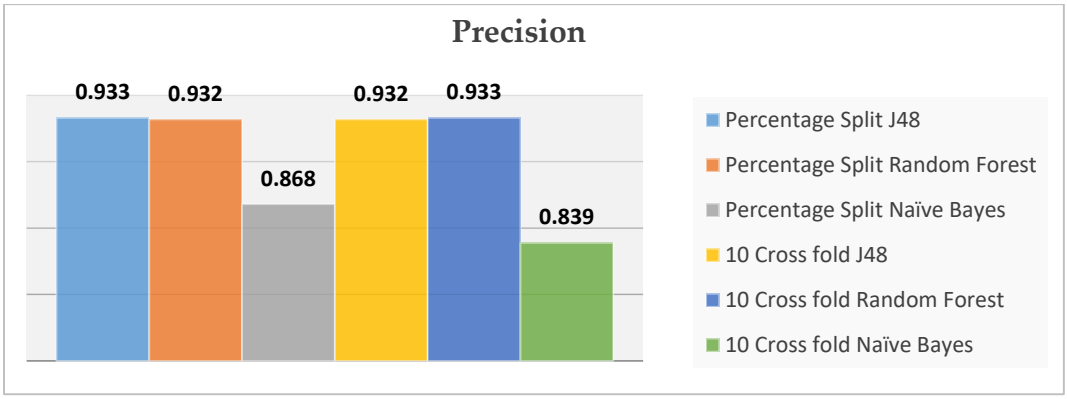

Figure 6.1 Precision in Dataset 3

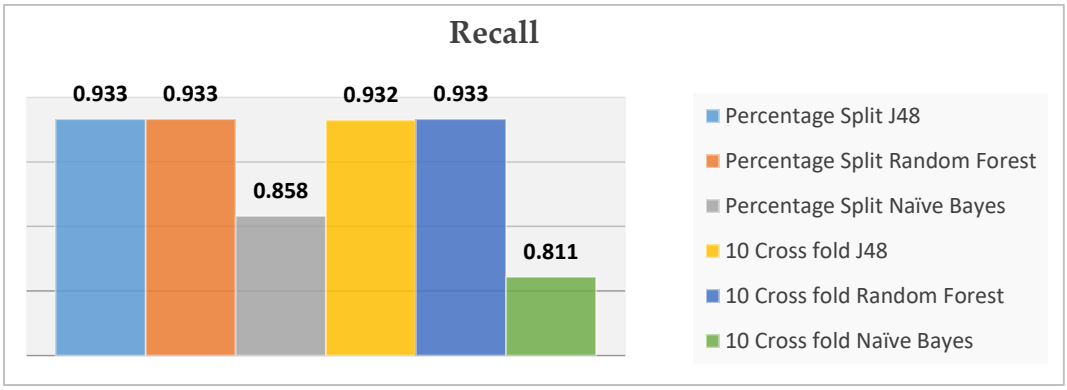

Figure 6.2 Recall in Dataset 3

The least accuracy is recorded by *Naive Bayes* at 74.7% on data set 1 while ten-cross fold validation method is used. *J48* has a better accuracy as approximately same as *Random Forest*. Accuracy result for all classification algorithms is shown in Appendix 2.

Random Forest creates a bootstrapped data set and a decision tree 100 times (the default value in WEKA) to make a vote and select the effective one. The algorithm is more robust so that it can work persistently in small or large data sets that may hold null or error values partially. It is also consistent while validation methods are changed. *J48* is also as approximately accurate as *Random Forest*. It uses the entropy and information gain for feature evaluation to make a better decision; that able to work in different data sets consistently.

Table 6-1 Accuracy Deviation Across Validation Methods

Classification Algorithms	Deviation	Dataset 1	Dataset 2	Dataset 3
J48	Between PS and CV	0.08%	0.53%	0.08%
Random Forest	Between PS and CV	0.17%	0.10%	0.02%
Naïve Bayes	Between PS and CV	2.81%	0.32%	4.75%

Where PS: Percentage Split
CV: Cross Validation

Naive Bayes is the least performer as compared to *J48* and *Random Forest* algorithms in all scenarios. It has a significant improvement in all validation methods while applied on *Dataset 3*. *Naive Bayes* classification algorithm is inconsistent across validation methods. Its accuracy varies significantly as the validation methods vary. A maximum accuracy deviation of 4.75% is observed between *percentage split* and *cross validation* methods as shown in Table 6-1. *J48* and *Random Forest* are much consistent in their accuracy as the validation method varies.

Table 6-2 Accuracy Deviation Across Datasets

Classification Algorithms	Deviation	PS	CV
J48	Between Dataset 1 and Dataset 2	0.5%	1.4%
	Between Dataset 2 and Dataset 3	0.04%	0.8%
	Between Dataset 1 and Dataset 3	0.8%	1.2%
Max Deviation – J48		**0.8%**	**1.4%**

Random Forest	Between Dataset 1 and Dataset 2	0.3%	0.8%
	Between Dataset 2 and Dataset 3	0.01%	0.8%
	Between Dataset 1 and Dataset 3	0.7%	1.2%
Max Deviation – RF		**0.7%**	**1.2%**
Naïve Bayes	Between Dataset 1 and Dataset 2	1.1%	9.5%
	Between Dataset 2 and Dataset 3	1.3%	5.0%
	Between Dataset 1 and Dataset 3	0.7%	2.3%
Max Deviation – NB		**1.3%**	**9.5%**

A significant variation on accuracy is occurred when Naïve Bayes classification algorithm is applied across the datasets using *cross validation* method. The maximum variation observed is 9.5% as it is shown in Table 6-2. The least deviation of accuracy is recorded in Random Forest as it is applied on *percentage split*; which is 0.01%. This minimum deviation is recorded between results in *Dataset 2* and *Dataset 3*.

Efficiency to build a model

Unlike *J48* & *Naive Bayes*, Random Forest takes a significant time to build a prediction model as shown in Table 6-3. The minimum and maximum time measurement *Random Forest* has lasted on are 5.79 and 9.67 seconds respectively. *Naive Bayes* is the most efficient classification algorithm in all scenarios of the analysis since it applies an independent measurement on features. The worst time in *Naive Bayes* is 0.08 seconds; and its fastest time elapsed to build its model is 0.05 seconds only. All classification algorithms took their worst time to build a model on *Dataset 1* since it is the largest data set compared to the other two data sets.

Table 6-3 Classification Algorithms' Efficiency in Prediction

VALIDATION TECHNIQUE	CLASSIFICATION ALGORITHM	TIME MEASUREMENT (SECONDS)		
		Dataset 1	Dataset 2	Dataset 3
PERCENTAGE SPLIT (70/30)	J48	0.4	0.38	0.4
	Random Forest	9.61	6.22	7.23
	Naïve Bayes	0.08	0.08	0.05
10 CROSS FOLD	J48	0.38	0.23	0.23
	Random Forest	9.67	5.79	7.25
	Naïve Bayes	0.08	0.05	0.05

Errors

Random Forest is robust in minimizing errors to be occurred. It scored minimum errors mostly in Dataset 3. *J48* has a slight difference of error values compared with *Random Forest*. *Naive Bayes* is the worst robust algorithm in minimizing errors in all scenarios of this study. Detail values of errors are shown in Appendix 6.

The first research question: "which prediction technique can effectively and efficiently identify possible termination of subscribers?" is answered in the analysis taken above to get the accuracy, efficiency and error of classification algorithms. As a result, Random Forest is the most effective and J48 is comparable to Random Forest. But Naïve Bayes is the most efficient compared to J48 and Random Forest. Although Naïve Bayes takes less time to build the model, it couldn't be taken as an advantage as we measure according to the aim of this study. Since such experiments are made once in a given period (week, month, year…), few seconds couldn't be a major issue to build a model.

Model Interpretability

J48 can build a prepaid mobile subscribers service termination model that can be clearly shown with the association of features to predict. According to this study, the only challenge observed in *J48* is overfitting problem. Overfitting minimizes interpretability of the model. In *J48* prediction, with default parameters, the model has 97 leaves with 193 total size of the tree. One of the parameters that makes the model more understandable

is pruning. Pruning can be done by limiting minimum number of instances per leaf. According to [59] there is no fixed method or clear guidelines for designing ensembles.

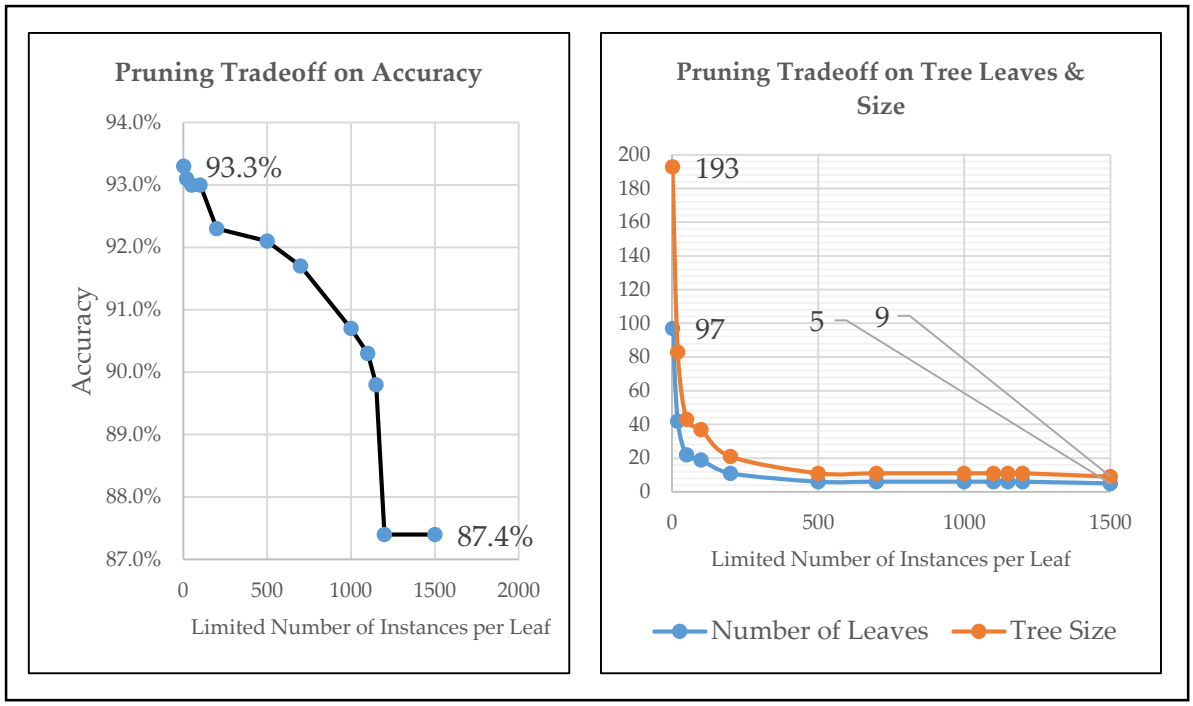

Figure 6.3 Pruning Tradeoff

The model is pruned by adjusting a parameter *minNumObj* with a value of 1500 to remove overfitting. Thus, a tree has been pruned to small leaves and size of the tree. In fact, pruning has a tradeoff on accuracy but it's better to have a clear model that helps a decision than overfitting [44]. After pruning is applied, *J48* model accuracy has been declined from 93.3% to 87.4%. The number of leaves and tree size with default parameters are 97 and 193; but after pruning their values are declined to 5 and 9 respectively as shown in Figure 6.3. Pruning in J48 has no impact in the information gain or rank of attributes importance. After pruning is applied, the tree becomes more visualized and easy to interpret for decision making.

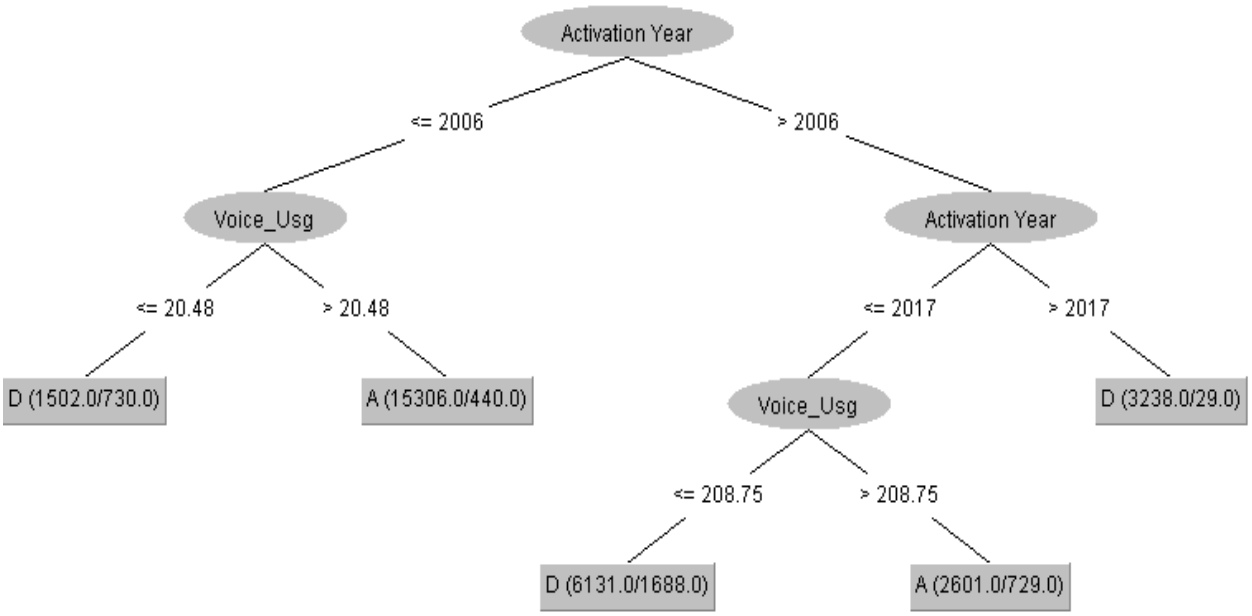

Figure 6.4 J48 Pruned Tree of Subscribers Termination Model

As it is depicted in Figure 6.4, *J48* pruned tree clearly shows 1502 subscribers (with FN of 730 subscribers) are terminated. The algorithm has learnt that subscribers with activation date prior to 2006 and with an average of monthly usage less than 20.48 minutes are mostly predicted as possible terminated subscribers. However, subscribers with a voice usage more than 20.48 minutes are mostly predicted as active subscribers. On the right side of the tree in Figure 6.4, when the subscribers' activation date is more recent, they are most likely to terminate the service unless they have an average 208.75 minutes voice usage per month.

Activation year has become the root node from the selected features applied in the prediction. It is directly proportional to the actual termination (deactivation) of subscribers. Subscribers with high tenure are loyal subscribers compared to the recently subscribed. Most recent subscribers have a high tendency to terminate.

The activation date mean of these subscribers in the data set is compared to all deactivated subscribers in ethio telecom. The sample data sets of terminated subscribers used in this study are too small compared to the total actual data in ethio telecom; their proportionality is about 1:683. As a result, actually terminated (deactivated) subscribers with activation year above 2006 are fairly represented in J48 model as shown in Figure 6.5.

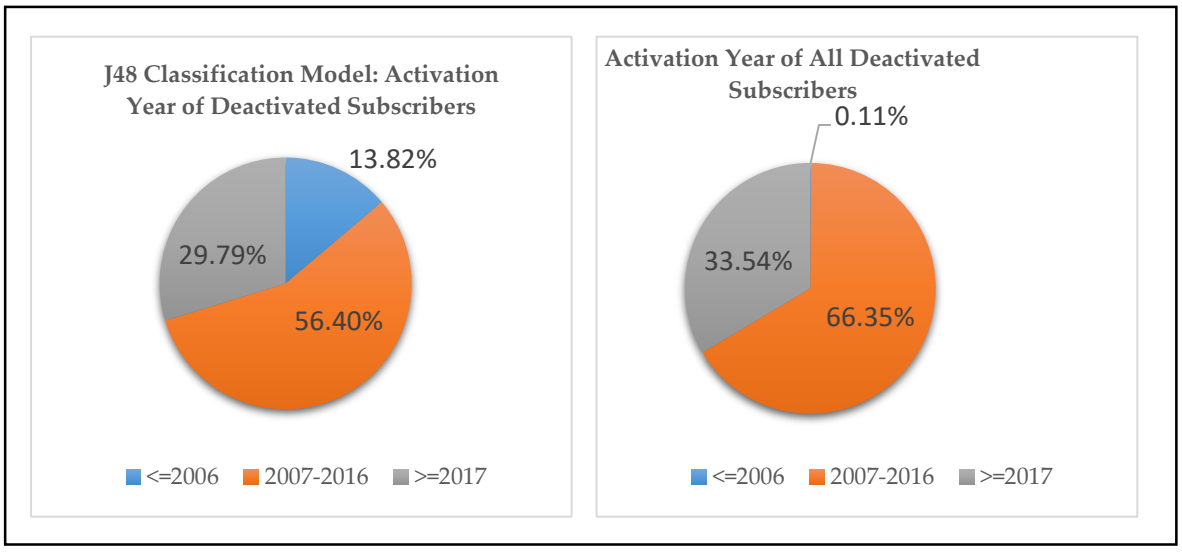

Figure 6.5 Activation Year Validation for J48 Model

The model built by *Random Forest* is not easy to make a decision. Unlike *Random Forest*, a better description is provided by the *Naive Bayes* model. *Random Forest* puts an importance of attributes for each feature as shown in Table 6-4.

Table 6-4 RF Prediction

Attribute (Feature)	Attribute Importance	Number of Nodes Used the Feature
Voice_Usg	36%	31526
Voice_Fee	31%	25683
Activation Year	31%	8647
SMS_Fee	26%	18520
Data_Fee	26%	15246
Data_Freq	22%	15135

Adjusting parameters in Random Forest has a significant impact on its accuracy and efficiency as well. Random Forest can be pruned by limiting its maxDepth parameter. In this study two parameters are adjusted. First, minimizing the tree from its default value (unlimited tree size) to a maximum depth of four. Second, by decreasing its iteration parameter from 100 to 10. As a result, its accuracy has been minimized from 93.4% to 91.1%, whereas time to build the model declines from 4.43 to 0.27 seconds.

Additionally, attributes importance value and number of nodes used by each features have been changed as shown in Table 6-5. Thus, pruning Random Forest tree has declined the accuracy, significantly minimized time to build the model and incurred inconsistency in ranking features that are helpful to make decisions.

Table 6-5 Pruned RF Prediction

Attribute (Feature)	Attribute Importance	Number of Nodes Used the Feature
Activation Year	21%	42
Data_Fee	12%	17
Voice_Usg	10%	31
Data_Freq	8%	16
Voice_Fee	6%	29
SMS_Fee	5%	15

Naive Bayes lacks to provide understandable model as compared to *J48*. But at least it puts the mean of each features with their standard deviations. Subscribers predicted as active (A) have an average year of activation around 2007 whereas deactivated (D) subscribers have around 2013. The standard deviation for both A and D are better closer to the mean as shown in Table 6-6. It describes active users have a better voice usage; they have also higher fee on voice, sms and data. Their data usage frequency is as twice as deactivated subscribers as well.

Table 6-6 Results in Naive Bayes

	A		*D*	
Measurement	Mean	Std.	Mean	Std.
Activation Year	2007.3225	2.9487	2012.8108	4.7136
Voice_Usg	281.2911	302.751	93.9162	228.0362
Voice_Fee	113.4437	126.9253	37.936	81.984
SMS_Fee	8.9491	17.6643	6.8023	18.5015
Data_Fee	52.9936	119.1724	30.0895	88.1817
Data_Freq	98.5613	194.3128	47.5389	112.9806

The activation year in Naïve Bayes model is much closer compared to the mean of whole deactivated (terminated) subscribers of ethio telecom as it is shown in Figure 6.6.

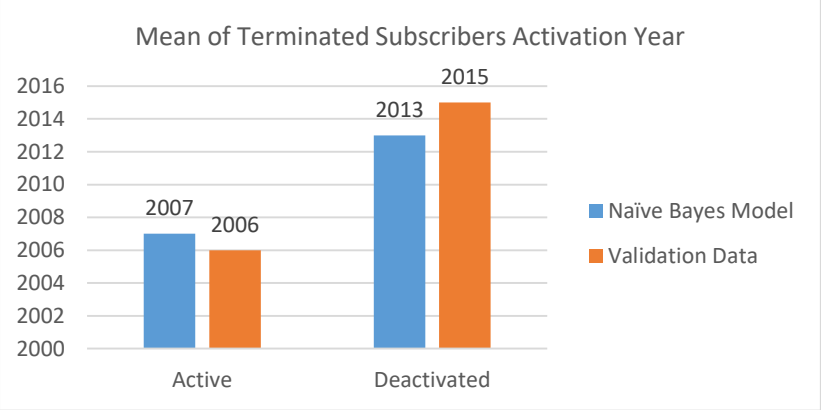

Figure 6.6 Activation Year Validation of Naïve Bayes

Regarding to answering the research questions raised at the beginning of this study, features or attributes influence prediction of subscribers' termination and the tradeoffs to build understandable model are identified. Activation year is the most influential feature in determining subscribers status as shown in Table 5-3, Table 6-5, Table 6-6 and Figure 6.4. As we minimize the size of a tree in order to get a clear and interpretable model, there is a tradeoff in reducing the accuracy as shown above in Figure 6.3. J48 is the most interpretable as compared to RF and Naïve Bayes.

7 CONCLUSION AND FUTURE WORK

7.1 CONCLUSION

CDR and profile data of prepaid mobile subscribers has been analyzed using three selected *J48*, *Random Forest* and *Naive Bayes* classification algorithms using WEKA tool. *Random Forest* is the most accurate with a minimum error values. It is consistent throughout all data sets and across all validation methods as well. The main drawback of this classification algorithm is taking a significant time to build its model. Additionally, the model cannot be easily understandable.

Unlike *Random Forest*, *Naive Bayes* is the least accurate classification algorithm as compared to *J48* and *Random Forest*. The main challenge to get a better result by *Naive Bayes* is its strong feature independence while all features in the selected data set are more dependent each other. Its model gives a better description compared with *Random Forest*; but its interpretability is poor as compared to *J48*. The only advantage in *Naive Bayes* is the efficiency it takes to build the model. Efficiency with poor accuracy couldn't bring the desired faulty subscribers prediction.

J48 is less predictive than *Random Forest*, with a slight difference in all scenarios. It takes moderate amount of time to build its model compared to *Random Forest* and *Naive Bayes*. *J48* has a model clarity to show the association of features to predict termination of subscribers. Random Forest is best in all metrics except its model clarity and time to build its model. In fact, elapsing time to build a model is not a big issue for this study. J48 has also a comparable effectiveness as Random Forest. Therefore, *J48* algorithm is recommended for its good classification accuracy, model clarity, robustness to error and consistency throughout the datasets and validation methods. *J48* model represents ethio telecom individual mobile subscribers' termination.

According to *J48* model, depicted in Figure 6.4, recent subscribers are more likely to terminate their service. But they incline to active if their average voice usage is more than 208.75 minutes per month. In addition, subscribers with high tenure (subscribed before 2007) and voice usage less than 20.48 minutes in average per month are also more likely to terminate. However, subscribers more than 20.48 minutes have high probability to be active user. Marketing management should plan to minimize termination of recently subscribed prepaid mobile users by considering their respective service usages.

Once prepaid subscribers are disconnected from telecom services, it is difficult to return them back. So, early detection of termination of subscribers has the advantage for ethio telecom to campaign, offer goods and services, proactive customer service, context-based marketing and improving ARPU preparation. Additionally, winning the nearly coming competition market without exhaustively knowing subscribers' usage behavior will be a big challenge. Therefore, preparing a plan to predict the termination of prepaid mobile subscribers is vital for ethio telecom to act as one of the world-class operators.

7.2 FUTURE WORK

Predicting subscriber termination is not a one-time task rather continuum. Thus, future works proposed to proceed starting from this research include:

- *Comparing classification models with customer experience:* This helps to identify the real problems of termination behind the figures in classification models.
- *Big data analytics:* Future studies can employ a big data analytics to organize a better dataset and features in order to get best prediction. Within and outside the network should be analyzed when competitors exist.

REFERENCES

[1] K. Dahiya and K. Talwar, "Customer Churn Prediction in Telecommunication Industries using Data Mining Techniques-A Review," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 5, no. 4, p. 2277, 2015.

[2] S. Kaur, "Literature Review of Data Mining Techniques in Customer Churn Prediction for Telecommunications Industry," *J. Appl. Technol. Innov.*, vol. 1, no. 2, pp. 28–40, 2017.

[3] G. Maji and S. Sen, "Data warehouse based analysis on CDR to retain and acquire customers by targeted marketing," *2016 5th Int. Conf. Reliab. Infocom Technol. Optim. ICRITO 2016 Trends Futur. Dir.*, no. September, pp. 221–227, 2016.

[4] F. F. Reichheld, "Prescription for cutting costs," *Bain Company. Harvard Bus. Sch. Publ.*, vol. September, no. September, 2001.

[5] T. Group, "Telefónica Group's Report January-June 2019," 2019.

[6] M. Azeem and M. Usman, "A fuzzy based churn prediction and retention model for prepaid customers in telecom industry," *Int. J. Comput. Intell. Syst.*, vol. 11, no. 1, pp. 66–78, 2018.

[7] "International bidding to invite two telecom operators in Ethiopia." [Online]. Available: "https://www.ethiopianreporter.com/index.php/article/17086/." [Accessed: 29-Oct-2019].

[8] "Ethiopia to open telecom for three operators." [Online]. Available: "https://newbusinessethiopia.com/investment/ethiopia-to-open-telecom-for-three-operators/." [Accessed: 2-Sep-2019].

[9] "Ethiopia opens market for multiple telecom operators" [Online].

"https://newbusinessethiopia.com/technology/ethiopia-opens-market-for-multiple-telecom-operators/."[Accessed: 5-Sep-2019].

[10] "Ethiopia plans to issue telco licenses by year end" [Online]. "https://www.reuters.com/article/us-ethiopia-telecoms-exclusive/exclusive-ethiopia-plans-to-issue-telco-licenses-by-year-end-id." [Accessed: 12-Sep-2019].

[11] J. Hidayati, L. Ginting, and H. Nasution, "Customer behaviour for telecommunication service provider," *J. Phys. Conf. Ser.*, vol. 1116, no. 2, 2018.

[12] "Applying Call and Event Detail Records to Customer Segmentation and CLV," *Int. J. Inf. Commun. Technol. Res.*, vol. 3, no. 8, 2013.

[13] V. Umayaparvathi and K. Iyakutti, "A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics," *Int. Res. J. Eng. Technol.*, pp. 2395–56, 2016.

[14] M. R. Khan, J. Manoj, A. Singh, and J. Blumenstock, "Behavioral Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty," *Proc. - 2015 IEEE Int. Congr. Big Data, BigData Congr. 2015*, pp. 677–680, 2015.

[15] S. Kumar and C. D., "A Survey on Customer Churn Prediction using Machine Learning Techniques," *Int. J. Comput. Appl.*, vol. 154, no. 10, pp. 13–16, 2016.

[16] L. Aurelie and S. Gupta, "Managing Churn to Maximize Profits Aurélie Lemmens Working Paper Managing Churn to Maximize Profits," *Harvard Bus. Sch.*, 2013.

[17] K. H. Liao and H. E. Chueh, "Applying Fuzzy Data Mining to Telecom Churn Management," *Commun. Comput. Inf. Sci.*, vol. 134, no. PART 1, pp. 259–264, 2011.

[18] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J. Big Data*, vol. 6, no. 1, 2019.

[19] M. Kaur and P. Mahajan, “Churn Prediction in Telecom Industry Using R,” *Int. J. Eng. Tech. Res.*, vol. 3, no. 5, pp. 46–53, 2015.

[20] A. Mishra and U. S. Reddy, “A comparative study of customer churn prediction in telecom industry using ensemble based classifiers,” *Proc. Int. Conf. Inven. Comput. Informatics, ICICI 2017*, no. January 2019, pp. 721–725, 2018.

[21] S. Y. Hung, D. C. Yen, and H. Y. Wang, “Applying data mining to telecom churn management,” *Expert Syst. Appl.*, vol. 31, no. 3, pp. 515–524, 2006.

[22] E. H. X. Nguyen, “Customer Churn Prediction for the Icelandic Mobile Telephony Market,” pp. 1–109, 2011.

[23] C. Kirui, L. Hong, W. Cheruiyot, and H. Kirui, “Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining,” vol. 10, no. 2, pp. 165–172, 2013.

[24] A. Q. Ahmed and D. Maheswari, “Churn prediction on huge telecom data using hybrid firefly based classification Churn prediction on huge telecom data,” *Egypt. Informatics J.*, vol. 18, no. 3, pp. 215–220, 2017.

[25] O. A. Amoo, B. O. Akinyemi, I. O. Awoyelu, and R. E. Adagunodo, “Journal of Emerging Trends in Computing and Information Sciences Modeling & Simulation of a Predictive Customer Churn Model for Telecommunication Industry,” vol. 6, no. 11, pp. 630–641, 2015.

[26] A. O. Dairo and T. Akinwumi, “Dormancy Prediction Model in a Prepaid Predominant Mobile Market : A Customer Value Management Approach,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 4, no. 1, pp. 33–39, 2014.

[27] M. Mohri, A. Rostamizadeh, and A. Talwalkar, "*Foundations of Machine Learning*," Second edi. London, England: The MIT Press, 2018.

[28] P. Langley and J. G. Carbonell, "*Approaches to machine learning*," vol. 35, no. 5. 1984.

[29] N. Prasasti and H. Ohwada, "Applicability of machine-learning techniques in predicting customer defection," *ISTMET 2014 - 1st Int. Symp. Technol. Manag. Emerg. Technol. Proc.*, no. May 2014, pp. 157–162, 2014.

[30] AN ORDERGROOVE WHITE PAPER, "Do subscription services modify consumer spend ?," pp. 2–5, 2019.

[31] A. T. Jahromi, "Predicting customer churn in telecommunications service providers," p. 88, 2009.

[32] G. L. Michael J.A. Berry, "*Data mining techniques : for marketing, sales, and customer relationship management*," 2nd ed. 2004.

[33] J. Hadden, A. Tiwari, R. Roy, and D. Ruta, "Churn prediction: Does technology matter," *World Acad. Sci. Eng. Technol.*, no. 16, pp. 973–979, 2008.

[34] ethio telecom, "ethio telecom Prepaid Mobile," 2019. [Online]. Available: https://www.ethiotelecom.et/prepaid-mobile/. [Accessed: 24-Jul-2019].

[35] P. Dönmez, "Introduction to Machine Learning, 2nd ed., by Ethem Alpaydın. Cambridge, MA: The MIT Press2010. ISBN: 978-0-262-01243-0. $54/£ 39.95 + 584 pages.," *Nat. Lang. Eng.*, vol. 19, no. 2, pp. 285–288, 2013.

[36] N. J. Nilsson, "Mlbook.Pdf (Predmet Application/Pdf)," 2005.

[37] S. Shalev-Shwartz and S. Ben-David, "*Understanding machine learning: From theory to algorithms*," vol. 9781107057. 2013.

[38] M. Kargari, "Introducing a heuristic method for combining fraud evidence based on outlier detection," pp. 1–13.

[39] A. H. Elmi and S. Ibrahim, "Detecting SIM Box Fraud Using Neural Network

Abdikarim," *IT Converg. Secur.*, vol. 215, no. January, 2012.

[40] S. Agarwal, "*Data mining: Data mining concepts and techniques*," 2014.

[41] G. Sharma, R. Bhargava, and M. Mathuria, "Decision Tree Analysis on J48 Algorithm," *Int. J. Adv. Res. inComputer Sci. Softw. Eng.*, vol. 3, no. 6, pp. 1114–1119, 2013.

[42] W. N. H. W. Mohamed, M. N. M. Salleh, and A. H. Omar, "A comparative study of Reduced Error Pruning method in decision tree algorithms," *Proc. - 2012 IEEE Int. Conf. Control Syst. Comput. Eng. ICCSCE 2012*, pp. 392–397, 2012.

[43] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," no. April, 1999.

[44] P. Kapoor, R. Rani, and R. JMIT, "Efficient Decision Tree Algorithm Using J48 and Reduced Error Pruning," *Int. J. Eng. Res. Gen. Sci*, vol. 3, no. 3, pp. 1613–1621, 2015.

[45] S. Drazin, "Decision Tree Analysis using Weka," *Mach. Learn. II, Univ. Miami*, pp. 1–3, 2010.

[46] "Classification Algorithms Random Forest." [Online]. Available: https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm. [Accessed: 02-Jul-2019].

[47] "Classification Algorithms Naive Bayes." [Online]. Available: https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_naive_bayes.htm. [Accessed: 02-Jul-2019].

[48] V. Koblar and B. Filipič, "Optimizing parameters of machine learning algorithms," 2012.

[49] R. Sharma, A. Ghosh, and P. K. Joshi, "Decision tree approach for classification of remotely sensed satellite data using open source support," *J. Earth Syst. Sci.*, vol. 122, no. 5, pp. 1237–1247, 2013.

[50] Global Affairs Canada, "Cultural Information with Conversations." [Online]. Available: https://www.international.gc.ca/cil-cai/country_insights-apercus_pays/ci-ic_et.aspx?lang=eng#wb-cont.

[51] The World Bank, "wb_id4d_dataset_2018_0." [Online]. Available: http://id4d.worldbank.org/global-dataset. [Accessed: 04-Aug-2019].

[52] Y. Theodorou, K. Okong'o, and E. Yongo, *Access to Mobile Services and Proof of Identity 2019: Assessing the impact on digital and financial inclusion*. 2019.

[53] D. Kiernan, "Correlation and Simple Linear Regression." [Online]. Available: https://textbooks.opensuny.org/natural-resources-biometrics/. [Accessed: 20-Aug-2019].

[54] S. Kumar and I. Chong, "Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states," *Int. J. Environ. Res. Public Health*, vol. 15, no. 12, 2018.

[55] M. Hall, "Class CorrelationAttributeEval," *University Of Waikato*, 2018. [Online]. Available: http://weka.sourceforge.net/doc.dev/weka/attributeSelection/CorrelationAttributeEval.html. [Accessed: 04-Sep-2019].

[56] M. J. M. G. Anthony J. Viera, "Understanding Interobserver Agreement: The Kappa Statistic," no. May, pp. 360–363, 2005.

[57] D. M. Ammar A. Ahmed, "Churn prediction on huge telecom data using hybrid firefly based classification," *Egypt. Informatics J.*, 2010.

[58] V. Valdimarsson, "Customer Churn Prediction in the Icelandic Telecommunication

Market," 2014.

[59] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, "Ensemble selection from libraries of models," *Proceedings, Twenty-First Int. Conf. Mach. Learn. ICML 2004*, no. 1996, pp. 137–144, 2004.

APPENDIX

Appendix 1 ethio telecom Subscriber Life Cycle

Subscriber Life Cycle	Description
Prepaid Life Cycle Rule Pre-active	▪ It is the status of the sim card before sales are made or before filled out Customer Application Form /CAF/. ▪ The unlimited validity period of the sim card on pre-active period.
Active Period	▪ It is the time which is equivalent to your prepaid account balance validity period. ▪ This period starts from the first call made till the expiry date of the validity period of your prepaid account. ▪ During this period, you can make & receive both voice call & SMS. ▪ The maximum accumulated validity period of an active period is 120 days.
One-way block /Barring/	▪ This is the period between prepaid account balance expiry date and the beginning of a two-way block. ▪ During this time, a customer can receive call/SMS but can't make voice call/SMS. ▪ You can recharge your account and become active. The existing balance will be kept. ▪ This period survives for 11 months (330 days).
Two way blocked /Suspend/	▪ This is the period between one-way block & prepaid account termination. ▪ During this period, you can't receive any voice call/SMS and can't make any voice call/send SMS. ▪ You can recharge your balance and become active, but the unused balance will be Zero. ▪ This period will only last for 30 days.
Termination	▪ After the expiration of a two-way block period, the SIM Card will be terminated instantly. ▪ You can't recharge your account. ▪ Your request for reconnection will not be accepted. ▪ Terminated numbers will be recycled instantly.
Recycling	▪ Terminated (Deactivated) numbers will be ready for reconnection.

Appendix 2 Accuracy of Classification Algorithms

Validation Technique	Classification Algorithm	Confusion Matrix		Predicted		Dataset 1	Dataset 2	Dataset 3
				A	D			
Percentage Split (70/30)	J48	Actual	A	5321	352	92.4%	91.9%	93.3%
			D	528	5310			
	Random Forest		A	5317	356	92.7%	92.4%	93.3%
			D	482	5356			
	Naïve Bayes		A	3650	2023	77.5%	76.4%	85.8%
			D	565	5273			
10 Cross fold	J48		A	17966	1219	92.4%	92.4%	93.2%
			D	1683	17502			
	Random Forest		A	18026	1159	92.5%	92.5%	93.3%
			D	1700	17485			
	Naïve Bayes		A	11226	7959	74.7%	76.1%	81.1%
			D	1744	17441			

Appendix 3 Classification Algorithms Accuracy on Dataset 1

Validation Technique	Classification Algorithm	TP Rate	FP Rate	Precision	Recall	F-measure
Percentage Split (70/30)	J48	0.924	0.076	0.924	0.924	0.924
	Random Forest	0.927	0.073	0.927	0.927	0.927
	Naïve Bayes	0.775	0.229	0.793	0.775	0.771
10 Cross fold	J48	0.924	0.076	0.925	0.924	0.924
	Random Forest	0.925	0.075	0.926	0.925	0.925
	Naïve Bayes	0.747	0.253	0.776	0.747	0.74

Appendix 4 Classification Algorithms Performance on Dataset 2

Validation Technique	**Classification Algorithm**	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-measure**
Percentage Split (70/30)	J48	0.919	0.075	0.923	0.919	0.920
	Random Forest	0.924	0.081	0.926	0.924	0.925
	Naïve Bayes	0.764	0.415	0.761	0.764	0.741
10 Cross fold	J48	0.924	0.083	0.925	0.924	0.924
	Random Forest	0.925	0.083	0.926	0.925	0.926
	Naïve Bayes	0.761	0.409	0.759	0.761	0.739

Appendix 5 Classification Algorithms Performance on Dataset 3

Validation Technique	Classification Algorithm	TP Rate	FP Rate	Precision	Recall	F-measure
Percentage Split (70/30)	J48	0.933	0.084	0.933	0.933	0.933
	Random Forest	0.933	0.093	0.932	0.933	0.933
	Naïve Bayes	0.858	0.139	0.868	0.858	0.861
10 Cross fold	J48	0.932	0.091	0.932	0.932	0.932
	Random Forest	0.933	0.093	0.933	0.933	0.933
	Naïve Bayes	0.811	0.156	0.839	0.811	0.816

Appendix 6 Error Rates in Algorithms' Classification

Validation Techniques	Error Types	Error Rates in Algorithms' Classification								
		Dataset 1			**Dataset 2**			**Dataset 3**		
		J48	**RF**	**Naïve Bayes**	**J48**	**RF**	**Naïve Bayes**	**J48**	**RF**	**Naïve Bayes**
Percentage Split	Mean absolute error	0.122	0.114	0.284	0.124	0.115	0.250	0.107	0.103	0.285
	Root mean squared error	0.252	0.241	0.402	0.256	0.245	0.400	0.234	0.229	0.396
	Relative absolute error	24.5%	22.9%	56.8%	27.9%	26.0%	56.4%	24.0%	23.3%	64.1%
	Root relative squared error	50.4%	48.2%	80.3%	54.6%	52.2%	85.3%	49.7%	48.7%	83.9%
10 Cross fold	Mean absolute error	0.122	0.115	0.291	0.120	0.115	0.255	0.107	0.102	0.288
	Root mean squared error	0.252	0.244	0.413	0.250	0.244	0.403	0.237	0.229	0.401
	Relative absolute error	24.4%	23.0%	58.2%	27.1%	25.9%	57.4%	24.0%	22.8%	64.8%
	Root relative squared error	50.5%	48.8%	82.5%	53.0%	51.8%	85.5%	50.4%	48.6%	85.0%