ST. MARY'S UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF COMPUTER SCIENCE

Application of Data Mining Technique for Predicting Airtime Credit Risk: The Case of Ethio Telecom

BY

OLIYAD TAREKEGN

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ST. MARY'S UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

June 23, 2019

Addis Ababa, Ethiopia

ST. MARY'S UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF COMPUTER SCIENCE

Application of Data Mining Technique for Predicting Airtime Credit Risk: The Case of Ethio Telecom

BY

OLIYAD TAREKEGN

Approval by Board of Examiners

_____ _____

Chairman, School of Graduate Committee

Dr. Getahun Semeon _____ _____

Advisor

Dr. Asrat Beyene _____ _____

Internal Examiner

Dr. Temtim Assefa _____ _____

External Examiner

DECLARATION

I, the undersigned, declare that this MSc thesis is my original work, has not been presented for fulfillment of a degree in this or any other university, and all sources and materials used for the thesis have been acknowledged.

Oliyad Tarekegn _____

Name Signature

Place: Addis Ababa, Ethiopia

Date of submission: June 14, 2019

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

Dr. Getahun Semeon _____

Advisors' Name Signature

DEDICATION

This thesis is dedicated to my MOM of whom I am enormously proud.

ACKNOWLEDGMENT

A word of thanks goes to God, Almighty, who has been with me throughout this period. During the time of doubt and wanting to give up, He raised me up so that I could stand on mountains, walk on stormy seas and I got stronger because I knew I was safe on His shoulders.

I am greatly indebted to my advisor Dr. Getahun Semeon whose insights, knowledge, patience, guidance, and humanity made me believe in this thesis. Your words of encouragement and 'straight talk breaks no friendship' approach are the reasons why I was able to see the finish line.

I would like to express my gratitude towards my family, particularly Habtamu Tarekegn and Letera Tarekegn for their encouragement, following up the progress of my work and support which helped me in completion of this thesis.

Finally, I would like to thank my friends, classmates and Ethio Telecom domain experts especially Dejen Hayelom, Tamirat Tesfaye and Abdu Nursbo for their kind help and support during the study.

TABLE OF CONTENTS

LIST OF ACRONYMS

Acronym	Description
3G	Third Generation
4G LTE	Fourth Generation Long Term Evolution
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ARFF	Attribute Relation File Format
ARPU	Average Revenue Per User
BSS	Business Support System
CAAZ	Central Addis Ababa Zone
CBS	Convergent Billing System
CDR	Call Data Record
CRISP	Cross Industry Standard Process
CRM	Customer Relationship Management
CSV	Comma Separated Value
DM	Data Mining
EAAZ	East Addis Ababa Zone
ER	East Region
ESB	Enterprise Service Bus
ETOP UP	Electronic top up
GMDB	Global Memory Database
GSM	Global Standard for Mobile Communication
ISP	Internet Service Provider
IVR	Interactive Voice Recorder
MLP	Multilayer Perceptron
MNO	Mobile Network Operator
MW	Middleware
NAAZ	North Addis Ababa Zone
NR	North Region
NWR	North West Region

ROC	Region of Convergence
SAAZ	South Addis Ababa Zone
SIM	Subscriber Identity Module
SMS	Short Message Service
SMSC	Short Message Service Center
SWAAZ	South West Addis Ababa Zone
SR	South Region
SWR	South West Region
KDD	Knowledge Discovery in a Data
VAS	Value Added Service
VC	Voucher Card
WAAZ	West Addis Ababa Zone
WEKA	Waikato Environment for Knowledge Analysis
WR	Western Region

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Airtime credit is a service that enable prepaid mobile subscribers to use telecom services any time even after running out of balance and pay for it later. This created convenience among users, and it became an additional source of revenue for operators. But this service has its own risk due to many subscribers failing to repay their credit and ending up as defaulters. The fact that telecom prepaid service users are not required to present any guarantee to get airtime credit makes the risk even worse.

This study explored the role of data mining in predicting airtime credit risk. An open source data mining tool called WEKA was used to conduct the experiment. Various classification algorithms were applied in order to find the best performing model. These algorithms were J48 decision tree, Naïve Bayes, Multilayer Perceptron and Logistic Regression. Ethio Telecom prepaid subscriber's usage data which consisted 86, 024 instances and eleven attributes were used for building and testing the algorithms. For all experiments performed, WEKA's tool 10-fold cross validation and percentage split test options were used. Confusion matrix was also used to evaluate the performance of the models using different measures such as accuracy, precision, recall, f-measure and ROC area.

The model built with J48 decision tree outperformed the other classifiers by an accuracy of 98.5632%, and Precision, Recall and F-measure of 0.986 and its ROC area threshold 0.996 with 10-fold cross validation test option. The model built with Logic regression has an accuracy of 97.1717%. Whereas Multilayer Perceptron and Naïve Bayes classifiers recoded an accuracy of 96.7622% and 94.6355% respectively. From the selected classifier there are some important rules and parameters generated which can help in airtime credit decision making process. Data usage is the main attribute which showed the potential prediction power. Which is, for a subscriber having high data usage with other usages set to low can predict a subscriber ending up as defaulter. Also, attributes such as voice usage and topping up channel has shown high airtime credit risk prediction power.

Keywords: Data Mining, airtime credit, risk prediction

CHAPTER ONE

INTRODUCTION

1.1 Background

Data mining is a new kind of business information processing technology, which can extract interesting patterns or knowledge implicated in many incomplete, noisy, and ambiguous data that people do not know in advance but with potential application [1]. It aims to find out 'hidden' correlations among data by extracting, converting, analyzing, and modeling from huge amount of transaction data in business database. Simply it is the process of extracting information in order to discover hidden facts contained in the database using a combination of machine learning, statistical analysis, modeling techniques and database technology in the areas such as decision support, prediction, forecasting and estimating. Generally, the goal of data mining is to create models for decision making that predict future behavior based on analysis of past activity. To effectively exploit the potential of data mining, database should be first organized into a format that used for further data mining process.

Among a host of recent technology innovations, data mining is making changes to the entire makeup of our skills and comfort zones in information analysis. Not only has introduced an array of new concepts, methods, and phrases, it also departs from the well-established, traditional, hypothesis-based statistical techniques. According to [1], data mining is a new type of exploratory and predictive data analysis whose purpose is to describe systematic relations between variables when there are no (or incomplete) a priori expectations as to the nature of those relations. In recent years, data mining is becoming hot research area and has made some remarkable achievements in business and nowadays, telecommunication is one of those businesses which is benefiting from this.

Mainly data mining has two approaches, they are predictive – to predict what possibly could happen next from the prior trends and the other aspect is description – to describe the trend from the existing trends or facts.

It is common practice for telecommunication operators to record customer data, such as customer address, contractual details, usage detail, service detail and the likes, in various databases to

support their billing activity. Information on customer usage patterns, as well as their payment transaction patterns, is becoming critical for telecommunication companies to understand the behavior of their customers. Buried with this vast amount of data are all sorts of information that could make a significant difference to the ways in which telecom operators run their business and interact with their customers. However, the necessary information that exists within the company databases are too fragmented and complex for a human mind to support efficient conclusions upon. In addition, it is too inaccessible and time consuming to gather, because the information required to make strategic and timely decisions is hidden in complex database systems [2]. This can be simplified by applying data mining techniques to analyze such a large data to find out the new information. Hence, data mining helps to facilitate automated prediction of trends and behaviors as well as automated discovery of hidden patterns which improves decision making and minimize time required to understand our data.

Airtime Credit service is a new kind of service being provided by mobile network operators (MNO) loaning airtime which is introduced to increase customer satisfaction as well as to generate more revenue to the ISPs. Prepaid Airtime Advance drives consumption, as subscribers can use services when they need it the most. A customer's "call opportunity" is often lost due to the inability of customer to purchase airtime. Prepaid Airtime Advance allows a customer to acquire airtime when it is needed and reduces the lost "call opportunity". This in turn increases the customer's average revenue per user (ARPU) and affinity to the network. Average revenue per user is a measure used primarily by consumer communications, digital media, and networking companies, defined as the total revenue divided by the number of subscribers. Researches show that many Prepaid Airtime Advance lending transactions occur in the early morning or late evening which provides a clear indication of a customer's needs to get airtime to generate a call when it is not easily accessible to purchase. Therefore, it is right to say Airtime advance is the most convenient and effective way to distribute airtime and always stay connected.

In Ethiopian context, Airtime credit or advance is introduced in which is almost one year now. This service is provided to subscribers to get them money as an advance and they will be charged on their next top up or when they receive transfer from another subscriber. The pre-paid airtime advance is a self-service option that allows prepaid customers to receive airtime in advance and pay for it later. It is like the customer is getting a loan, get airtime now and pay later. The benefit

of this service is that customers can stay connected and continue making calls when they are unable to recharge or buy more airtime.

How Airtime Loan Service Works

There are three parties involved in the airtime loan services operated by any mobile operator, with reference to Ethio Telecom, although in other climes there are two depending on the risk appetite level of the operator. These are the mobile service operator, a VAS (Value Added Service) Provider and the subscriber.

Each of these parties plays different role in the airtime loan service value chain, the subscriber is the end- user interested in the loan service, the VAS provider is the middleman who provides the loan (the airtime) and bears the risk of loss if the subscriber defaults in the repayment of the loan and mobile operator is the one that provides the platform that links the other two parties in the value chain.

The way it works is that the VAS provider purchase airtime in bulk from the operator and this is done by paying cash into the operator's bank account in exchange for electronic airtime that will be disbursed for the loan service. When the subscriber request for a loan through the mobile operator, the operator passes the request to the VAS provider to credit the subscriber with the airtime loan requested for. Once this is credited into the subscribers account the airtime can then be used by the subscriber to purchase any network services offered by the operator. At the time of disbursing the loan, an interest fee is deducted ahead.

For example, if a subscriber request for a loan of 100 Birr, the VAS operator will credit the subscriber with 100 Birr. But the outstanding amount of the subscriber will be 110 Birr as 10% service charge is included. The point to note here is that the debt amount is 10% more than the loan request. To pay back the loan the subscriber will be debited the cumulative sum of 110 Birr once he/she top-up airtime or get airtime by transfer thereafter. The conclusion here is that the subscriber has paid 10 Birr in excess of the airtime received. The excess of 10 birr is what will be shared between the VAS provider and the mobile operator as their own interest in the loan disbursed based on the agreed terms. In some cases, the VAS provider will take 10% while the

operator takes 90% of the interest charged to the subscriber. If the subscriber defaults on paying back the loan, it becomes a bad debt which will be serviced by the VAS provider hence, the risk of non-payment of bad debt is solely the responsibility of the VAS provider, the mobile operator is not going to share in the loss.

From the discussions, we can deduce what the airtime is and how the airtime loan service offered by the operators are listed below:

1. Airtime is not a service, it is money. Services are the airtime you use to purchase, and these are voice, SMS and data services.
2. The airtime loan is not the same as the amount paid back by the subscriber.
3. The VAS provider and the operator has an advantage over the subscriber because the subscriber pays more than the amount loaned.
4. The operator gains at all times, they have collected the money for the airtime ahead of time from the VAS provider, when the loan is paid back, they get a certain percentage of the interest (also called service charge) and when the subscriber defaults they have nothing to lose.
5. The VAS provider is at the receiving end as they have everything to lose when the subscriber defaults.

Therefore, it is necessary to say that this air time credit requires intensive customer behavior analysis to score the customers before deciding the amount of loan a subscriber can get up on request so as to create a positive experience for operators, VAS providers and subscribers by reducing the defaulters as well as minimizing the risk of revenue loss. Hence, it is possible to use data mining techniques as a means of predicting defaulters and deny their request, which may lead to reducing the risk of revenue loss.

The potential of data mining is becoming very important and a wide range of companies around the globe has deployed successful applications of data mining [3]. However, there are attempts that were made in applying data mining in the areas of financial institutions and telecommunication in Ethiopia.

The first attempt made was a case study on prediction of non-payment for Ethiopian Telecom made by [4] on Using Data Mining techniques to Combat Infrastructure Inefficiencies. Their study focused on building a model to rank customers of the company according to their non-payment

likelihood using decision tree. They found that location was the strongest predictor of non-payment; however, billing changes were also significant predictors.

The other attempt was on the application of data mining on credit data [5] for predicting defaulters or inconsistent loan payers which was conducted on united bank customers. In which a model was built to identify trends of good and bad patterns from historic data and the classification model performed well using J48 decision tree algorithm.

Moreover, [6] has developed a prediction model of customer loyalty (Non loyal or Loyal) which supports microfinance institutions during loan decision making using different DM techniques, among them a classification model of J48 was used to generate the rule.

Hence, this research is a continuation of the data mining researches carried out so far, however, with a different area of application and goal, which is to predict defaulters, non-payments likely hood, of airtime credit subscribers by taking Ethio Telecom as a case.

1.2 Statement of the Problem

Price fairness, customer services and coverage are major factors which can highly affect the customer satisfaction in telecommunications industry [7]. One means of getting the customers happy and satisfied is by providing them services such as airtime credit. This service enables users to use any telecom service provided by the ISP at any time of need without worrying whether their account has enough balance. Most of the time subscribers are unable to recharge their account during late night and early in the morning [8]. During this time providing them with airtime credit loan can create convenience among the users as well as increasing the revenue generated from users.

This service has its own challenge when it comes to loan repayment. For instance, in the case of Ethio Telecom, airtime credit service is introduced before four months and has got a lot of subscribers using the service as it has solved their limitation of not using services when running out of balance. From the preliminary investigation of the researcher there are more than 4.5 million subscribers who have unpaid loan. From these subscribers, more than 180,000 are those who took

the loan before three months and yet do not settle their debt. The total loan amount of these subscribers stands at about 2.7 million Birr. Even though the service is recently implemented, the amount of outstanding debt is very high which can be a big worry for the lender. So, this requires a special attention as this may lead to further increase in defaulters i.e. those who do not tend to pay their debt on time. This will have unwanted consequence for the company as it will lead to big revenue loss which is against the objective of the service at all as there is a high possibility that these subscribers may not pay their debt at all. Therefore, this research will focus on tackling this specific problem by applying different data mining techniques which will help to analyze customers of the company for different decision making. One of these is to predict which customers are likely to request an advance loan and are likely to repay their debt on time based on historical data on the subscriber.

So far researches have been conducted on finical institutions focusing on credit risk. But due to the nature of the loan it is not possible to directly apply the models to airtime credit risk prediction. Some of the dissimilarity of the financial and airtime credit loans are: Financial loans are based collateral agreement, relatively high amount of money, not 100 percent risk as there will be physical and financial assessment before granting the loan. Whereas, airtime credit is technology dependent, micro loan or relatively small amount of money is lent, made at subscribers' convenience and purely based on trust that the subscriber will not default.

To the best of the knowledge of the researchers there is not any other such a research that is done before with the same objective specifically, airtime credit. The main purpose of this research is therefore, to build a predictive model that can accurately identify or predict subscribers of Ethio Telecom that are likely to pay and those who are likely not to pay their loan based on the usage data.

Research Question

The research answered the following major questions:

- What is the extent of problem in the existing airtime credit loan repayment?
- Is data mining technique (like classification) suitable for suggesting best features for airtime credit loan decisions and making prediction on customer airtime loan repayment likelihood?

- Is predictive data mining model can be built for airtime credit risk prediction based on the usage pattern of subscribers?
- What are the best features to consider in passing airtime credit loan decisions by the company?

1.3 Objective of the Study

The general and specific objectives of the research are described below.

1.3.1 General Objective

The general objective of this research is to build a model for predicting airtime credit risk of non-repayment by using data mining techniques.

1.3.2 Specific Objectives of the Study

In order to achieve the general objective, the specific objectives identified are the following:

- To assess the extent of problem associated with airtime credit repayment,
- To collect relevant customer dataset required for the mining, analysis and performance evaluation.
- To prepare the data for pattern mining by selecting, cleaning, reducing, summarizing and integrating.
- To select appropriate data mining technique to address the risk of non-payment
- To design a classification scheme in order to develop a predictive model for non-payment of airtime credit by customers
- To evaluate the model to effectively characterize the customers
- To report on the results and make recommendations based on findings.

1.4 Scope of the Study

The scope of this research was restricted to building data mining model for classifying the customers, interpreting the resulting models, and generating a classification rules for the selected model. The classification model prototype is implemented to the real scenario and tested further. For this purpose, and to simplify the implementation only selected attributes such as data usage, voice usage and network age are used. The selection of the attributes is done by applying WEKA

tool attribute evaluator function based on gain ratio feature ranking. Moreover, it is limited to the data mining aspects of classification using the proposed algorithm due to time limitation.

1.5 Application of the Study Result

The result of this study helps to characterize subscribers based on their usage behavior. Characterizing them support during different decision making. One of the goals that is supported by this decision-making process is airtime credit loan. During an airtime credit the subscriber is evaluated whether it is qualified for loan or not by considering how it was behaving. The model used previous knowledge of defaulters and non-defaulters before deciding the subscribers' fate. This helps to make sure the subscribers who are requesting the loan and gets the airtime are those who are likely to pay back their debt. Which in the other hand reduces the risk of defaulting as well as increasing revenue of the company.

1.6 Thesis Organization

This thesis is consisting of five chapters. The first chapter deals with the general overview of the study including background of the study, statement of the problem, objectives, scope, application and methodology of the research. The second chapter is devoted to literature review on data mining technology and machine learning. And it also covers review of applicable data mining techniques including related research works. The third chapter discusses about the methodology followed to conduct the study. This chapter comprises model building steps such as business understanding, data understanding, data preparation. The fourth chapter is dedicated to model building experiments using different data mining algorithms, evaluation, and deployment of the result. Results of the experiment are also analyzed and interpreted there. The last chapter which is Chapter five presents the conclusion that summarize the major points of the research and recommendations forwarded for practice and further research and adjustments in the organization on the ground of the research results.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Now a days it is common to find a huge amount of data stored in companies like telecom, banks and any other transaction related businesses [9]. The data can be simple numerical figures and text documents, or more complex information like spatial data, multimedia data, and hypertext documents. To take complete advantage of data stored in files, databases, and other repositories, data retrieval is simply not enough. It requires a technique or powerful tool for analysis and interpretation of such data and that could help in decision-making [10]. This technique or tool is data mining. Data mining is the extraction of hidden predictive and descriptive information from large databases or huge data typically stored in a data warehouse. It is a powerful technology with great potential to help organizations to focus on the most important information in their data warehouses [11]. Data mining techniques also enable to predict future trends and behaviors, and helps organizations to make proactive knowledge-driven decisions and it can solve the problems that traditionally were too much time consuming like preparing databases for finding hidden patterns, finding predictive information that experts may miss [12]. And, different techniques and algorithms are used to accomplish the tasks of data mining. The key properties of data mining are: automatic patterns discovery, prediction of likely outcomes, creation of actionable information and it focuses on large data sets and databases. Automatic discovery refers to the execution of data mining models [13].

This chapter discusses the potential of data mining to discover knowledge from huge databases. It also provides a brief historical development of the field. Besides, it presents a review of different data mining methods, common tasks and basic steps for data mining process.

2.2 Data Mining

This is an age often referred to as the information age. In this information age, information leads to power and success, and thanks to technologies such as computers, satellites, etc., it is possible to collect tremendous amounts of data/information. Initially, with the advent of computers and mass digital storage, collecting, storing, and counting using computers were started. Unfortunately, these massive collections of data stored on disparate structures very rapidly became increased [10]. This problem has led to the creation of structured databases and Database Management Systems

(DBMS) [9]. The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of information from a large collection of data. Today, there is huge amount of information (more information than can handle) from different sources such as from business transactions and scientific data, satellite pictures, text reports and military intelligence and so on [11]. Information retrieval is simply not enough anymore to use this enormous data to extract important information for decision-making purpose. Confronted with huge collections of data, new needs are created now to help us make better managerial choices [13].

These needs are automatic summarization of data, extraction of the "essence" of information stored, and discovery of patterns in raw data. Data mining techniques are the means to do these needs. Technologies such as data warehousing, data mining, and campaign management have greatly assisted companies to gain a competitive advantage [12].

Data mining enables the extraction of hidden predictive and descriptive information from large databases. As a result, business enterprises identify valuable customers, predict future customers' behaviors, and enable firms to make proactive knowledge driven decisions [13]. Thus, data mining is a means to preserve customers by understanding their needs proactively.

Data mining is defined as exploration and analysis of large quantities of data by automatic or semi-automatic means to discover meaningful patterns and rules and these patterns allow a company to better understand its customers, and improve its marketing, sales, and customer support operations [14]. Data mining is often done to analyze data in order to gain knowledge about the behavior patterns of customers and to identify key relationships that may help in decision making.

2.3 Knowledge Discovery Process Models

Data mining is "the principle of sorting through large amounts of data and picking out relevant information" [15]. It is usually used by business intelligence organizations, and financial analysts, but nowadays, it is increasingly used in the science fields to extract information from the enormous data sets generated by modern experimental and observational methods.

Data mining is also known as knowledge discovery. Even though data mining and Knowledge Discovery in Databases (KDD) are frequently treated as synonyms, knowledge discovery is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases, whereas data mining is one part of the knowledge discovery process [1]. The following figure shows data mining is one of the steps in knowledge discovery process.

KDD refers to the overall process of discovering useful knowledge from stored data, and data mining refers to a step-in knowledge discovery process. Data mining is a step that consists of applying data analysis and discovery algorithms that produce a enumeration of patterns (or models) over the data. Specifically, data mining is the application of specific algorithms for extracting of expected patterns from dataset [12]. Witten and Frank [16] also define, data mining step as the process of finding interesting patterns from raw data that are not clearly/explicitly part of the data. These interesting patterns can be used to tell us something new and to make predictions. It also analyzes large observational data sets to find hidden relationships and to summarize the data in new ways that are both understandable and useful to the user of the data [13]. Practically, data mining provides tools by which large quantities of data can be automatically analyzed and extracted.

2.3.1. The KDD Process Model

Knowledge discovery in data base is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [24].

As described in figure 2.1 below, Knowledge discovery in database (KDD) has five stages, such as selection, preprocessing, transformation, Data Mining and Interpretation or Evaluation.

Selection: This stage is concerned with creating a target data set or focusing on a subset of variables or data samples, on which discovery is to be performed by Understanding the data and the business area. Because, Algorithms alone will not solve the problem without having clear statement of the objective and understanding.

Pre-processing: This phase is concerned in removing noise or outliers if any, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes. On top of these tasks, deciding on DBMS issues, such as data types, schema, and mapping of missing and unknown values are parts of data cleaning and pre-processing.

Transformation: The transformation of data using dimension reduction or transformation methods is done at this stage. Usually there are cases where there are large numbers of attributes in the database for a case. With the reduction of dimension there will be an increase in the efficiency of the data-mining step with respect to the accuracy and time utilization.

Data Mining: This phase is the major stage in data KDD because it is all about searching for patterns of interest in a representational form or a collection of such representations. These

representations include classification rules or trees, regression, clustering, sequence modeling, dependency, and line analysis. Therefore, selecting the right algorithm for the right area is very important.

Evaluation: In this stage the mined data is presented to the end user in a Human viewable format. This involves data visualization, where the user interprets and understands the discovered knowledge obtained by the algorithms.

Using the Discovered Knowledge: Incorporating this knowledge into a performance system, taking actions based on the knowledge, or simply documenting it and reporting it to interested parties, as well as checking for and resolving for conflicts with previously acquired knowledge are tasks in this phase.

Knowledge discovery in database (KDD), as a process consists of an iterative sequence of steps as discussed above. It is also clear that data mining is only one step in the entire process, though an essential one, it uncovers hidden patterns for evaluation.

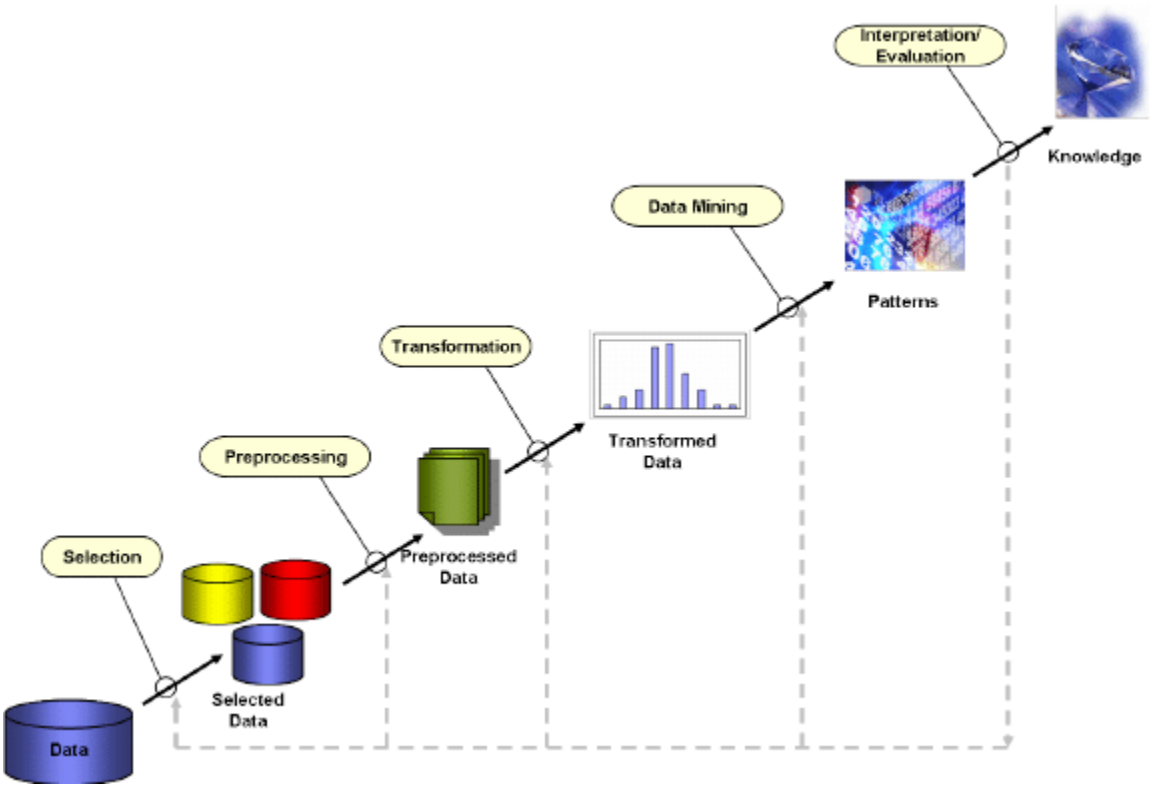


Figure 2.1 The five steps in KDD process [24]

2.3.2 CRISP-DM Process Model

A data mining process model defines the approach for the use of data mining, i.e. phases, activities and tasks that must be performed whereas data mining represents a complex and specialized field. So, a generic and standardized approach is needed for the use of data mining in order to help organizations. CRISP-DM (CRoss-Industry Standard Process for Data Mining) is a non-proprietary, documented and freely available data mining process model created in 1996. It was developed by the industry leaders and the collaboration of experienced data mining users, data mining software tool providers and data mining service providers [24]. To develop further and refine this process model and service the data mining community well Special Interest Group (CRISP-DM SIG) was formed. CRISP-DM version 1.0 was presented in 2000 and it is being accepted by business users [24].

According to [25], the life cycle (steps) of a data mining (KDD process), consists of six phases. In the transformation of raw data from business transaction and other sources to useful information that could help to make decisions, each step is built on the previous ones and the sequence of the phases is not rigid, moving back and forth between different phases is always required depending on the outcome of each phase. The main phases are:

- **Business Understanding Phase**: This phase focuses on understanding the data mining (KD process) objectives and requirements from business perspective, then converting this knowledge into a data mining problem definition and this is a preliminary plan designed to achieve the objectives. According to [18], this phase may also be termed as research understanding phase.
- **Data Understanding Phase**: This phase is concerned with data collection and understanding of the data using exploratory data analysis to get familiar with the data, to evaluate the quality of data, and to discover first approaching into the data or to detect interesting subsets or actionable patterns to form hypotheses for hidden information.
- **Data Preparation Phase**: This phase, which is labor intensive, covers all aspects of preparing the final data set, which will be used for subsequent phases, from initial, dirty and raw data. This stage includes operations like dimension reduction (such as feature selection and sampling), data cleansing (such as handling missing values, removal of noise or outliers), data transformation (such as Discretization of numerical attributes and attribute construction) and finally, clean the raw data so that it is ready for the modeling tools.

- **Modeling Phase**: In this phase, various modeling techniques (classification, regression, clustering and summarization) and algorithms are selected and applied or employed accordingly and their parameters are standardized or adjusted to optimal values. Often, several different data mining techniques may be applied for the same data mining problem.
- **Evaluation Phase**: In this stage, the model is methodically evaluated. Review and interpret the mined patterns for quality and effectiveness of the model before deploying it for use in the field. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use or task of the data mining results should be reached or decided.
- **Deployment Phase**: The purpose of the model is to increase knowledge gained from the data, and the knowledge gained need to be organized and presented in a way that the customer can understand and use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

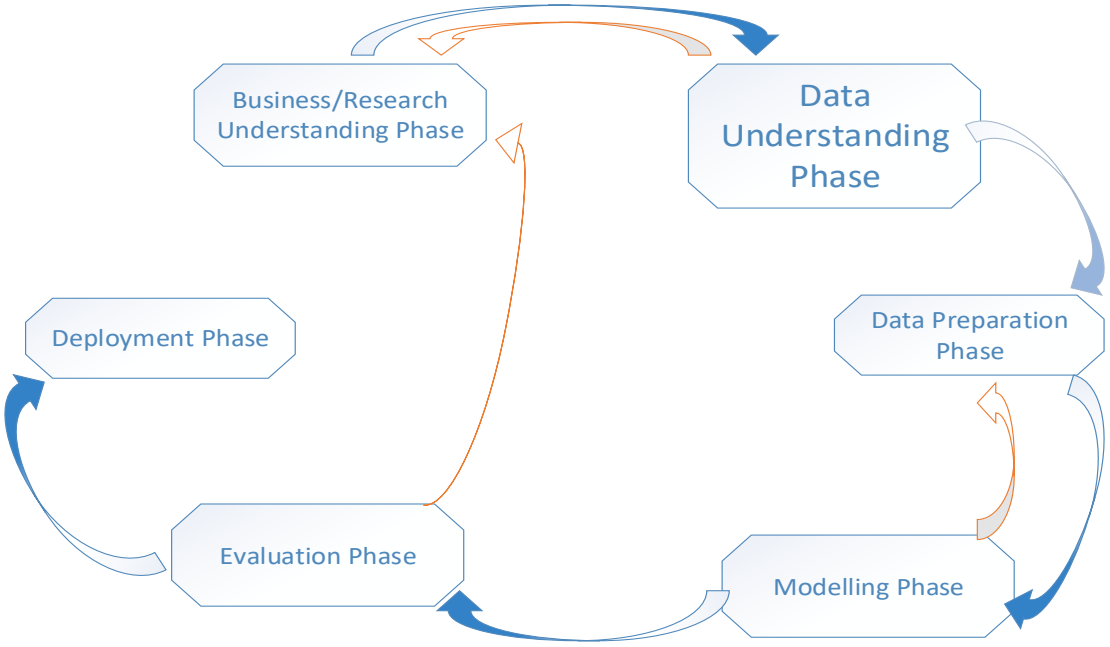


Figure 2.2 CRISP-DM Process model [27]

2.3.3. SEMMA Process Model

The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a DM project. The SAS Institute considers a cycle with five stages for the process [28]:

Sample - this stage consists on sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly;

Explore - this stage consists on the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas;

Modify - this stage consists on the modification of the data by creating, selecting, and transforming the variables to focus the model selection process;

Model - this stage consists on modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome;

Assess - this stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the DM process and estimate how well it performs.

The SEMMA process offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for his conception, creation and evolution, helping to present solutions to business problems as well as to find de DM business goals [28].

2.4 Tasks of Data Mining

The data mining tasks can be classified generally into two types based on what a specific task it tries to achieve. These two categories are descriptive tasks and predictive tasks. The descriptive data mining tasks characterize the general properties of data whereas predictive data mining tasks perform inference on the available data set to predict how a new data set will behave.

Data mining can be used to accomplish different tasks. But, the task of data mining is depending on the use of the data mining result that means for what purpose the result will be used [18]. The tasks are classified as follow:

- **Exploratory Data Analysis**: It is simply to exploring the data without any clear ideas of what we are looking for. These techniques are interactive and visual.
- **Descriptive Modeling**: It describe all the data, it includes models for overall probability distribution of the data, partitioning of the p-dimensional space into groups and models

describing the relationships between the variables. In short, it is applied to describe the existing data.

- **Predictive Modeling**: To predict the future having or based on the existing data or behavior. The model enables to predict the value of one variable from the known values of other variables.
- **Discovering Patterns and Rules**: It is concerned with pattern detection; the aim is spotting fraudulent behavior by detecting regions of the space defining the different types of transactions where the data points significantly different from the rest. Unlike descriptive modelling, pattern discovery deals with generating patterns and important hidden rules.
- **Retrieval by Content:** It is finding pattern similar to the pattern of interest in the dataset. This task is most commonly used for text and image datasets.

2.5 Types of Data Mining Systems

Data mining systems can be classified into different categories. The classification is based on: type of data source mined, data model, kind of knowledge discovered, mining techniques used, and also the degree of user interaction involved in the data mining process [19]. The classification of data mining systems according to the type of data source mined is based on the type of data handled for mining purpose such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

The classifications of data mining systems based on the data model are relational database, object-oriented database, data warehouse, transactional database, etc. Classification of data mining systems according to the kind of knowledge discovered; this classification is based on data mining functionalities or methods, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together. Classification of data mining systems based on mining techniques used in the data analysis approach are machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented [18], etc. Finally, the data mining system classifications based on the degree of user interaction in the data mining process are query-driven systems, interactive exploratory systems and autonomous systems. A comprehensive system would provide a wide variety of data mining techniques to fit different situations and options and offer different degrees of user interaction.

2.6 Data Mining Functions

Data mining techniques or methods are classifications of data mining systems based on the kind of knowledge discovered. The two main types of data mining methods are verification-oriented (the system verifies the user's pre-defined hypothesis) and discovery-oriented in which the system can finds new rules and patterns autonomously from the data [20].

Discovery methods are methods that automatically identify or recognize patterns in the dataset. The discovery method consists of prediction methods and description methods. Description-oriented data mining methods focus on understanding the way the underlying data operates, whereas, a prediction-oriented method aim to build a model that can be able to predict newly and unseen data according to the model developed by the sample (training set). However, some prediction-oriented methods can also provide understanding of the data. Most of the discovery-oriented techniques are based on inductive learning that means the model is constructed explicitly or implicitly by generalizing from enough training data. The underlying assumption of the inductive approach is that the trained model is applicable to future new and unseen variable.

The following figure shows the categories (classifications) of data mining methods:

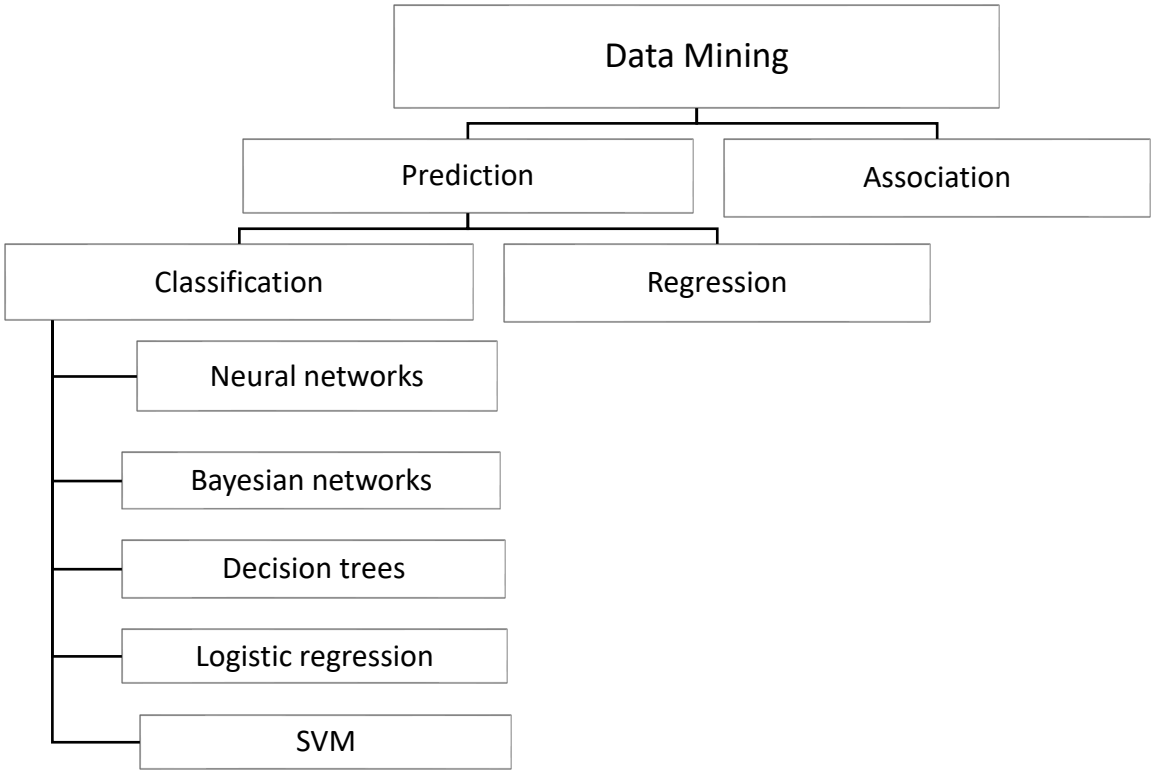


Figure 2.3. Classification of data mining techniques [20]

Verification methods, on the other hand, deal with the evaluation of a hypothesis proposed by an external source (an expert). These methods include the most common methods of traditional statistics like goodness-of-fit test, hypothesis testing, and analysis of variance. These methods have less connection with data mining than discovery-oriented methods because most data mining problems are related to predicting and selecting a hypothesis (out of a set of hypotheses) rather than testing a known one. Moreover, "the focus of traditional statistical methods is usually on model estimation as opposed to one of the main objectives of data mining: model identification". According to [20], the common data mining methods are classification, regression, clustering, summarization, dependency modeling, and change and deviation detection.

2.6.1 Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data, which is also the main goal of this paper. A classification task begins with a data set in which the class assignments are known. Classifications are discrete and do not imply order. The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, in our case as high credit risk (defaulters) and low credit risk (good credit) while multi class targets have more than two classes.

In the model build process or training, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques to find relationships. These relationships are summarized in a model, which can then be applied to different data set in which the class assignments are unknown. The models are tested by comparing the predicted values in a set of test data. The historical data used for the classification purpose is typically divided in to two data sets: one is to build the model; and the other is for testing the built model accuracy.

Some of the well-known classification algorithms are Bayesian classification (based on Bayes Theorem), decision trees, neural networks and backpropagation (based on neural networks), k-nearest neighbor classifiers (based on learning by analogy), and genetic algorithms [21].

According to [21], decision trees are the most known top-down approach for classification that divides the data into leaf and node divisions until the entire data set has been analyzed and evaluated. Neural networks are nonlinear predictive tools that learn from a prepared data set and then applied it to new and large sets. Genetic algorithms are like neural networks but incorporate

natural selection and mutation. Nearest neighbor utilizes a training set of data to measure the similarity of a group and then use the resultant information to analyze the test data.

2.6.2 Regression

Regression is a DM function that predicts a number. A regression task begins with a data set in which the target values are known. This method is used to make predictions based on existing data by applying formulas. For example, a regression model that predicts house values could be developed based on observed data for many houses over a period of time [17]. Using linear or logistic regression techniques from statistics, a function is learned from the existing data. The new data is then mapped to the function in order to make predictions [22]. According to [10], decision trees with averaged values at the leaves are a common regression technique.

2.6.3 Clustering

Clustering involves identifying a finite set of categories (clusters) to describe the data. The clusters can be mutually exclusive, hierarchical or overlapping [20]. Each member of a cluster should be very similar to the members within its cluster and dissimilar to other clusters members. Techniques used to create clusters on data stored include partitioning method like the simple k-means algorithm and hierarchical methods which group objects or data into a tree of clusters, such as density-based methods [1].

2.6.4 Summarization

According to Dunham [22], summarization is also called characterization or generalization. It derives summary data from the stored data or extracts actual portions of the data which briefly characterize (describes) the contents. Summarization maps data into subsets and then applies a compact description for that subset.

The mining methods discussed above form the basis for most data mining activities. Many variations on the basic approaches described above are found in the literature, including algorithms specifically modified to apply to spatial data, temporal data mining, multi-dimensional databases, text databases and the Web [17,22].

2.7. Predictive Data Mining

Data mining is the exploration of historical data (usually large in size) in search of a consistent pattern and/or a systematic relationship between variables; it is then used to validate the findings by applying the detected patterns to new subsets of data [22]. The roots of data mining originate

in three areas: classical statistics, artificial intelligence (AI) and machine learning [9]. Pregibon [23], described data mining as a blend of statistics, artificial intelligence, and database research, and noted that it was not a field of interest to many until recently.

According to [20] data mining can be divided into two tasks: predictive tasks and descriptive tasks. The aim of data mining is prediction; therefore, predictive data mining is the most common type of data mining and is the one that has the most application to businesses or life concerns. Prediction task predicts the possible values of missing or future data. Prediction involves developing a model based on the available data and this model is used in predicting future values of a new data set of interest. Predictive data mining has three stages, they are: data preprocessing, prediction and deployment. These stages are elaborated upon in Figure 2.3, which shows a more complete picture of all the aspects of data mining.

The data mining process starts with the collection and storage of data in the data warehouse. Data collection and warehousing is a whole topic of its own, consisting of identifying relevant features in a business and setting a storage file to document them. It also involves cleaning and securing the data to avoid its corruption. According to Kimball, a data warehouse is a copy of transactional or non-transactional data specifically structured for querying, analyzing, and reporting [12]. Data exploration, which follows, may include the preliminary analysis done to data to get it prepared for mining. The next step involves feature selection and or reduction. Mining or model building for prediction is the third main stage, and finally come the data post-processing, interpretation, and/or deployment.

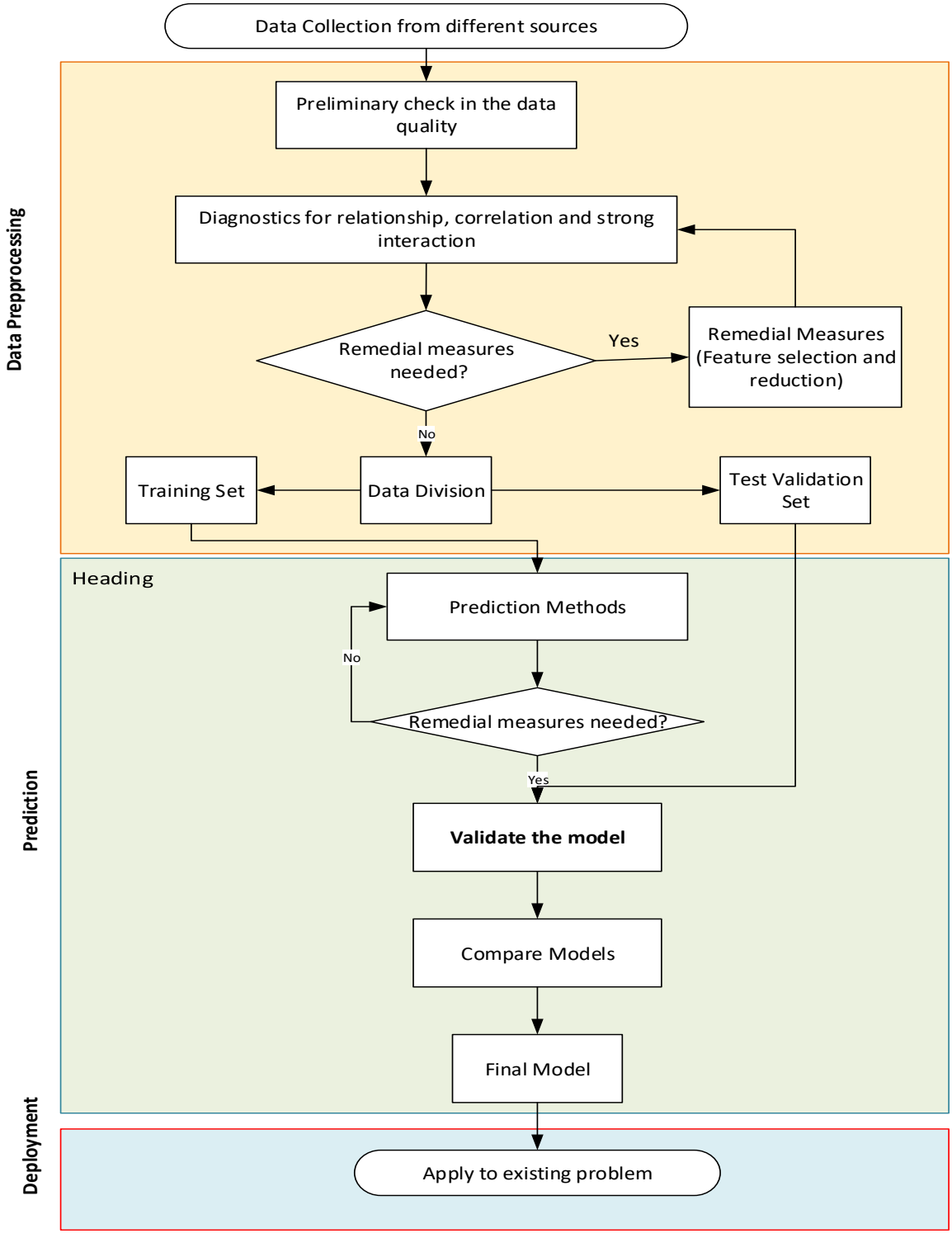


Figure 2.4. The stages of predictive DM [17]

2.8 Supervised and Unsupervised Learning

2.8.1. Supervised learning

Supervised learning is about uncovering the existence of a relationship between independent variables or attributes and a target attribute which is sometimes called a dependent variable [47]. The above relationship is established by identifying and using a selected learning algorithm which will identify those values of the target variable which form an association with predictor variables [48]. It is also worth noting that several supervised learning systems usually rely on a bigger pool of records which are already pre-labelled, and that this helps the selected algorithms to formulate models that could help in solving the identified problem within an organization [48].

In addition, as shown in Figure 2.5 below, supervised learning is used to estimate an unknown dependency from known input-output data [46]. This means that the selected input datasets are processed by a learning algorithm and whatever output results that come out of that are compared with those from the sample. Such an action allows for the error signals from the sample output to be at a minimum through repeated adjustments of the learning system [46]. Furthermore, it is worth noting that in supervised learning, the results are assessed, using intrinsic ways due to the presence of class labels which are pre-determined [47].

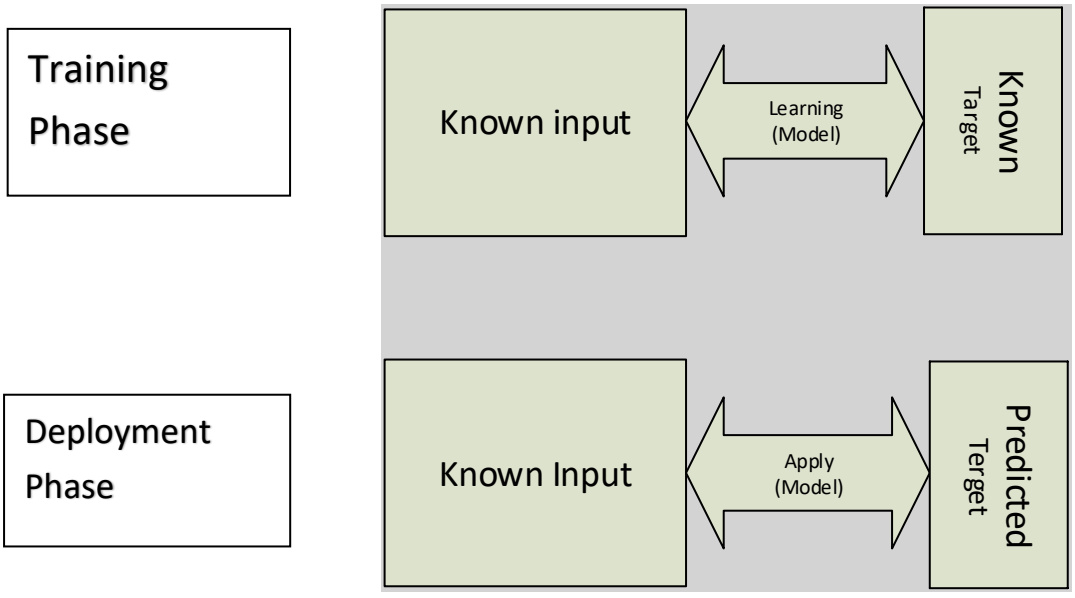


Figure 2.5 Supervised learning [48]

Supervised learning is popular in many fields of study such as in banking where they may be used to determine whether a housing bond application could be labelled as either bad or good credit risk. In telecommunications, for example, customers can, using supervised learning, be categorized as loyal or churner based on their usage behavior or pattern of their call detail record. Noticeable among the classification methods are the categorical variables which are segmented into pre-determined classes or categories [48]. Due to its success in predicting the value of a target attribute, classification has also been embraced in many areas such as financial firms and telecommunications [47].

2.2.2 Unsupervised learning

Contrary to supervised learning, unsupervised learning systems can search independently for new or previously unknown knowledge patterns. They are also known for not relying on any target or variables. The selected input data goes through a learning system without any validation against any output. Unsupervised learning systems can uncover 'natural' structures in the input data. Furthermore, unsupervised learning independently discovers and presents output results whose assessment is often regarded as intrinsic [47].

In addition, as shown in Figure 2.6. below, unsupervised learning can be described as a bottom-up approach where data can discover independently without pre-determined rules [49]. Unlike supervised learning, unsupervised methods do not have any targeted output, and the identified data is searched with the aim of discovering patterns [48].

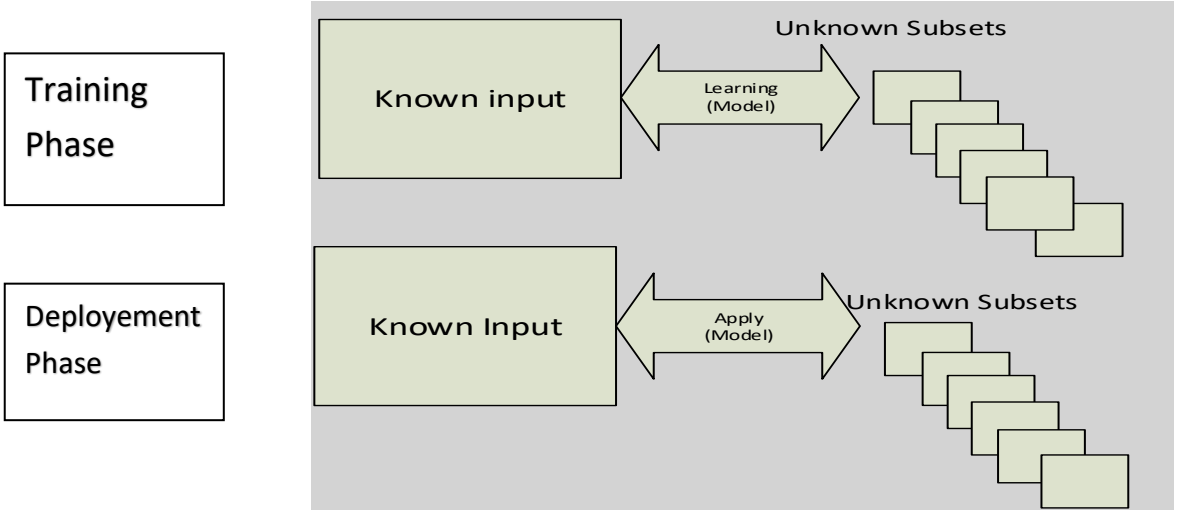


Figure 2.6 Unsupervised learning [48]

Unsupervised leaning, of which clustering learning algorithm forms a part, can be divided into three classes, namely metric distance-based methods, model-based methods and partition-based methods [55]. They can also be used as precursors for other DM activities. For example, in neural networks, it is always a good thing when dealing with large sets of data to do clustering first in order to eliminate problems further down the line [48].

In conclusion, unsupervised learning, of which clustering forms a part, is descriptive in nature whereas classification, a member of supervised learning, is predictive in nature. Although there is a thin line between predictive and descriptive learning systems, it is common for the models that come out of these learning systems to display each other's characteristics. Even though there are various reasons for choosing suitable learning systems, KDD tends to favor description rather than prescription [56].

2.9. Classification Learning Algorithms

According to [28], classification is the most important and popularly used technique in data mining. It is a process of finding a set of models or pre-defined conditions that describe and distinguish data classes or concepts.

Supervised learning method is alternative term to express the classification technique. In supervised learning (classification), we are provided with a collection of labeled (pre-classified) patterns and the problem are to label a newly encountered, yet unlabeled pattern. The given labeled (training) patterns are used to learn the descriptions of classes which in turn are used to label (classify) a new coming pattern. Classification technique maps data into predefined groups. The derived model of classification may be represented in various forms such as (IF-THEN) rules, decision tree, neural networking, Bayesian networks etc.

2.9.1. Decision Tree

A Decision tree is a flowchart like tree structure, where each internal node (non -leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision tree uses the traditional structure shown in Figure 2.6. It starts with a single root node that splits into multiple branches, leading to further nodes, each of which may further split or else terminate as a leaf node. Associated with each non-leaf node will be a test or question that determines which branch to follow. The leaf nodes contain the decisions [29][30].

Decision trees attempt to find a strong relationship between input values and target values in a group of observations that form a data set. When a set of input values is identified as having a strong relationship to a target value, then all these values are grouped in a bin that becomes a branch on the decision tree. These groupings are determined by the observed form of the relationship between the bin values and the target [31][32].

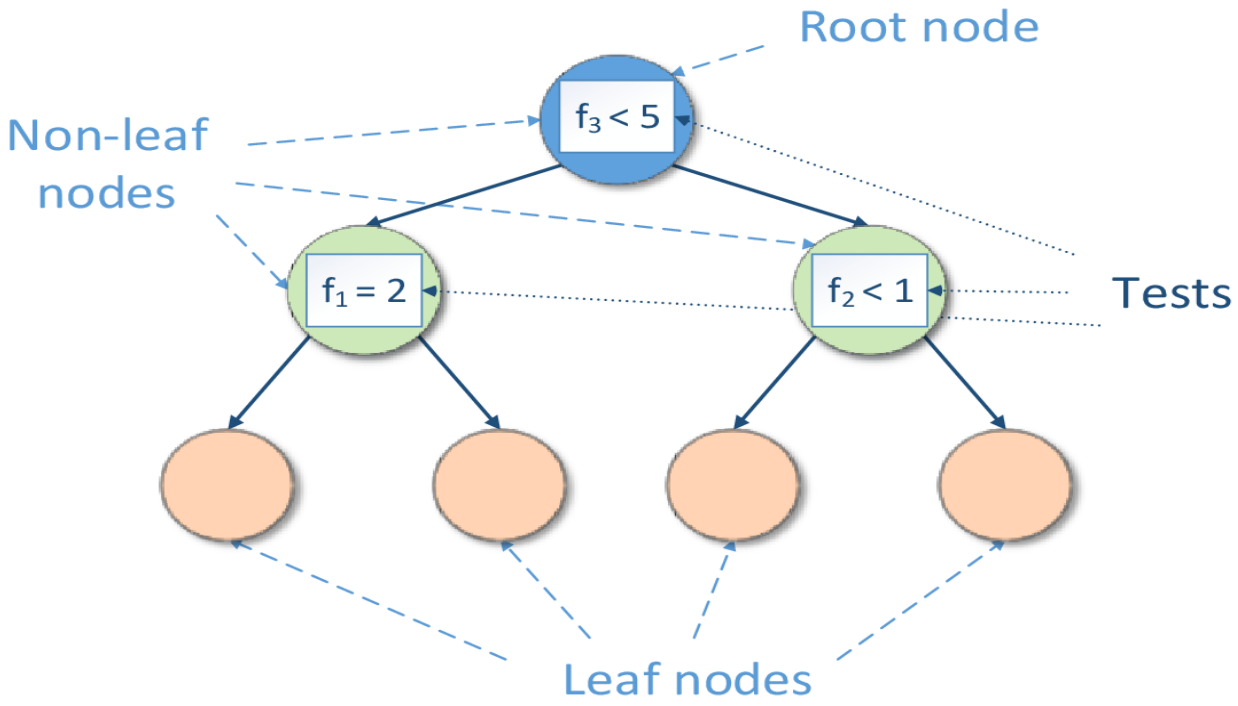


Figure 2.7. Traditional decision tree structure adopted from [30]

Construction of Decision Tree

The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle multidimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. Their attraction lies in the simplicity of the resulting model, where a decision tree (at least one that is not too large) is quite easy to view, understand, and, importantly explain. In general, decision tree classifiers have good accuracy. However, decision trees do not always deliver the best performance, and represent a trade-off between performance and simplicity of explanation and successful use may depend on the data at hand, [31][32].

A. Tree Size

Generally, decision makers prefer less complex decision trees, since it is considered more comprehensive and easier to understand. According to [33], the tree complexity has a crucial effect on its accuracy. The tree complexity is explicitly controlled by the stopping criteria used and the pruning method employed. Usually the tree complexity is measured by one of the following metrics: the total number of nodes, total number of leaves, tree depth and number of attributes used.

B. Rule Induction in Trees

As described in [34], there is a close relationship between decision tree induction and rule induction. Each path from the root of a decision tree to one of its leaves can be transformed into a rule simply by conjoining the tests along with the path to form the antecedent part and taking the leaf ‘s class prediction as the class value.

C. Splitting Criteria

As indicated in [29], there are two types of splitting criteria in decision tree: Univariate Splitting Criteria and Multivariate Splitting Criteria.

Univariate Splitting Criteria: Univariate means that an internal node is split according to the value of a single attribute. Consequently, the inducer searches for the best attribute upon which to split. The most common criteria in the literature includes: Impurity-based Criteria, Information Gain, Gini Index, Gain Ratio, etc.

Multivariate Splitting Criteria: In multivariate splitting criteria, several attributes may participate in a single node split test. Obviously, finding the best multivariate criteria is more complicated than finding the best univariate split. Furthermore, although this type of criteria may dramatically improve the tree’s performance, these criteria are much less popular than the univariate criteria. Most of the multivariate splitting criteria is based on the linear combination of the input attributes.

D. Stopping Criteria

The growing phase continues until a stopping criterion is triggered. Conditions that are common for stopping rules are: All instances in the training set belong to a single value of y; The maximum tree depth has been reached; The number of cases in the terminal node is less than the minimum number of cases for parent nodes; If the node were split, the number of cases in one or more child nodes would be less than the minimum number of cases for child nodes; The best splitting criterion is not greater than a certain threshold, [29][34].

E. Tree pruning

When a decision tree is crafted, countless of the divisions imitate anomalies in the training data due to noise or outliers. Tree pruning [35] methods report this setback of above fitting the data. Such methods normally use statistical measures to remove the least reliable divisions there are two common approaches to tree pruning: pre-pruning and post-pruning. Key motivation of pruning is ―trading accuracy for simplicity‖. There are assorted methods for pruning decision trees. Most of them present top down or bottom up traversal of the nodes. A node is pruned if the procedure improves precise conditions. 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.

Cost-Complexity Pruning Cost intricacy pruning (also renowned as weakest link pruning or error intricacy pruning) takings in two stages. In the early period, sequences of trees are crafted on the training datasets, whereas the early tree beforehand pruning is the root tree. In the subsequent period, one of these trees is selected as the pruned tree, established on its generality of error estimation.

Pessimistic pruning

Pessimistic pruning avoids the need of pruning set or cross validation and it uses the pessimistic statistical association test in its place. The basic idea is that the error ratio estimate during the training set is not consistent sufficiently. Instead a more practical measure known as continuity correction for binomial allocation should be used.

Reduced Error Pruning

As traversing above the inner nodes from the bottom to the top of a tree, the REP procedure Checks for every single internal node, whether substituting it alongside the most recapped class that does not cut the accuracy of trees. In this case, the node is pruned. The procedure endures till each more pruning would cut the accuracy. In order to guesstimate the accuracy Quinlan provides to use a pruning set. It can be shown that this procedure ends alongside the smallest accurate sub- tree alongside respect to a given pruning set.

Hence, decision tree is selected to be used in this study because of the following reasons: It is a good fit for two class label data, easy to understand and explain the result, it is fast to build a model which is very important for such a large data of telecom, and the accuracy of the model [36]. There are different decision tree algorithms like Id3, J48graft, AD tree, C4.5, J48 etc. J48 algorithm is WEKA’s improved implementation of C4.5 algorithm.

2.9.2. J48 Decision Tree Algorithm

Decision tree is a classification machine learning algorithm, as is one of the algorithms employed in this study, which is a most-used supervised learning due to its practical nature and its ability to influence several areas within the knowledge discovery field of study [20]. It plays a specific role, namely, to generate models which come from pre-determined rules. The pre-determined categories on which classification methods work are, for example, already labeled as either yes or no [37].

Classification, which is popular in many fields of study, ranging from scientific discoveries to financial engineering, is used by many experts to predict several areas of interest. Its popularity may be tied to the fact that the developed rules called IF…THEN make it easier for experts to obtain the desired results from the selected datasets in a seamless manner [38]. In a banking environment, for example, customers are granted or denied credit based on their risk profiles which may be classified as either low or high. The result during the use of classification algorithms on any selected dataset are models which are normally evaluated in order to give credibility to the selected data [39].

The whole process of classification, as it is the case with the use of C4.5 (j48) in this study, is preceded by the division of the chosen data into the training and test sets. According to [40], a training set is given to a learning algorithm, which derives a classifier. Then the classifier is tested with the test set, where all class values are hidden. If the classifier classifies most cases correctly, it can be assumed that it will also work accurately on future data. On the other hand, if a classifier makes too many errors, we can assume that it is a wrong model.

In a telecommunications environment, classification is popular because of its iterative nature which means that data can be manipulated several times until the desired model is generated [40]. It also can produce models which, in relation to customer behavior, may be categorized as either loyal or churner, likely to repay loan or not, likely to request loan or not, fraudulent or not. In addition, predicting the possibility of a customer requesting airtime credit depends on certain criteria which, when put to the test, would further predict whether they will repay the loan or becomes defaulter [41]. Furthermore, classification learning algorithms may be used by the company in order to group customers according to their usage behavior and their preferred service. It can also help the company to generate more revenue by granting loan to subscribers as per their repaying capability.

Hence, in a telecommunication sector, automatic classification becomes an unavoidable option during airtime credit advance loan. It is, therefore, common practice for the selected system to classify customers based on their current situation. This leads to the generation of a model which can predict the class value from other explanatory attributes [40].

For this study there will be additional two classification methods to be used in addition to J48 Decision tree. They are Naïve Bayes and ANN Multilayer Perception, which are discussed in the next sections.

2.9.3. Bayesian Network Classifiers

Bayesian networks are graphical models which are very useful for representing variables (as nodes of the graph) and the probabilistic relationships between them (as connections, or edges of the graph). By knowing the value at one of the nodes in a Bayesian network, one can infer the value of other nodes in the network. Bayesian network classifiers are used in many fields and one common class of classifiers are Naive Bayes classifiers. The induction of classifiers from data sets of pre-classified instances is a central problem in machine learning. Numerous approaches to this problem are based on various functional representations such as decision trees, decision lists, neural networks, decision graphs, and rules. One of the most effective Bayesian network classifiers, in the sense that its predictive performance is competitive classifiers, is the Naive Bayesian classifier [42]. This classifier learns from training data the conditional

probability of each attribute *Ai* given the class label *C*. Classification is then done by applying Bayes rule to compute the probability of *C* given the particular instance of *A*1,…., *An,* and then predicting the class with the highest posterior probability.

According to Elkan [43], on many real-world datasets naive Bayesian learning gives better test set accuracy than any other known method, including backpropagation and C4.5 decision trees. Also, these classifiers can be learned very efficiently.

Bayesian networks can have different advantages. Among those, some of them are provide probabilistic output, can operate with limited sensor data availability, more flexible relative to engineering development then traditional expert systems, used for both data qualification (state recognition) and anomaly reasoning, can operate in a central or distributed run-time environment either shore-side or ship-board.

The reason why use Bayesian networks is Bayesian inference methods have proven to be valuable for knowledge-based data mining applications and are based on a causal (explanation based) modeling framework. Because relationships between variables in a Bayesian network are defined probabilistically, trends can be detected and analyzed over a continuous scale, rather than in a Boolean fashion. The main reason to choose Naïve Bayes algorithm is, it is very fast [42].

Bayes Theorem

According to [44], Bayesian probability is named after Thomas Bayes, who was an eighteenth-century theologian. Bayesian probability allows prior knowledge and logic to be applied to uncertain statements.

As indicated in [45], let the training dataset D consist of n points of xi in a d-dimensional space, and let y_i denote the class for each point, with $y_i \in \{c1, c2\ldots ck\}$. The Bayes classifier directly uses the Bayes theorem to predict the class for a new test instance, x. It estimates the posterior probability $P(C_i|X)$ for each class ci and chooses the class that has the largest probability. The Bayes theorem allows inverting the posterior probability in terms of the likelihood and prior probability, which can be written as:

$$P(Ci/X) = \frac{P(X/Ci).P(Ci)}{P(X)} \qquad (2.1)$$

Where P(x|ci) is the likelihood, defined as the probability of observing x assuming that the true class is ci, P(ci) is the prior probability of class ci, and P(x) is the probability of observing x from any of the k classes, given as

$$\mathrm{P(x)} = \sum_{j=1}^{k} P(x/cj).P(cj) \qquad (2.2)$$

[46] stated Bayes' classification rule for multiclass objects with a multidimensional feature vector as:

"Given an object with a corresponding feature vector value x, assign an object to a class cj with the highest a posteriori conditional probability P (cj/x)."

In other words:

"For a given object with a given value **x** of a feature vector, assign an object to class cj when P (cj/**x)** > P (ci/**x),** i = 1, 2…, l; i ≠ j"

The conditional probability p (ci/x) is difficult to ascertain; however, using Bayes' theorem, we express it in terms of p (x/ci), P (ci) and P (x):

"A given object, with a given value x of a feature vector, can be classified as belonging to class cj when

$$\frac{P(X/cj)P(cj)}{P(X)} > \frac{P(X/ci).P(ci)}{P(X)}, i = 1,2, \dots; i \neq j \qquad (2.3)$$

After canceling the scaling probability p(x) from both sides, we obtain the following form of the Bayes classification rule:

"Assign an object with a given value x of a feature vector to class cj when $P(x/cj)P(cj) > P(x/ci)P(ci)$, $i = 1,2, \dots; i \neq j$"

In the sense of the minimization of the probability of classification, Bayes' error rule is the theoretically optimal classification rule. In other words, there is no other classification rule that yields lower values of the classification error.

2.9.4. Neural Network

Neural networks are computing models for information processing and are particularly useful for identifying the fundamental relationship among a set of variables or patterns in the data. They grew out of research in artificial intelligence; specifically, attempts to mimic the learning of the biological neural networks especially those in human brain which may contain more than 10^{11} highly interconnected neurons. They do share two very important characteristics with biological neural networks - parallel processing of information and learning and generalizing from experience [29].

The neural networks field was originally coined by psychologists and neurobiologists who sought to develop and test computational analogs of neurons. Roughly speaking, a neural network is a set of connected input/output units in which each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples. Neural network learning is also referred to as connectionist learning due to the connections between units [30].

As described in [47] the basic neural networks structure consists of two kinds of components: neurons - the processing elements and interconnections (synapses, links). Each link in the network is described by the weight parameter. Neurons can be classified into 3 groups: input, output and hidden neurons. Input neurons receive and process the signal from outside the networks, output neurons produce the out coming information (result) and neurons whose inputs and outputs are connected to other neurons are called hidden neurons.

[48], describes the network topologies and the paradigm of learning in neural network as follows:

A. Network topologies

Feed-forward networks where the data flow from input to output units is strictly feed-forward. The data processing can extend over multiple (layers of) units, but no feedback connections are present, that is, connections extending from outputs of units to inputs of units in the same layer or previous layers.

Recurrent networks that do contain feedback connections; contrary to feed-forward networks, the dynamical properties of the network are important. In some cases, the activation values of the units undergo a relaxation process such that the network will evolve to a stable state in which these activations do not change anymore. In other applications, the change of the activation values of the output neurons is significant, such that the dynamical behavior constitutes the output of the network.

Classical examples of feed-forward networks include the Perceptron and Adaline; and that of recurrent networks are the Hopfield network, the Elman network (where some of the hidden unit activation values are fed back to an extra set of input units), and the Jordan network (where output values are fed back into hidden units.

B. Paradigms of learning

We can categories the learning situations in two distinct sorts. These are:

Supervised learning or Associative learning in which the network is trained by providing it with input and matching output patterns. These input-output pairs can be provided by an external teacher, or by the system which contains the network (self-supervised).

Unsupervised learning or Self-organization in which an output unit is trained to respond to clusters of patterns within the input. In this paradigm the system is supposed to discover statistically salient features of the input population. Unlike the supervised learning paradigm, there is no a priori set of categories into which the patterns are to be classified; rather the system must develop its own representation of the input stimuli.

2.9.4.1. Multi-Layer Perceptron

Multilayer perceptions (MLPs) are the best known and most widely used kind of neural network. They are formed by units of the type shown in Figure 2.7, as an input layer. Each of these units forms a weighted sum of its inputs, to which a constant term is added. This sum is then passed through a nonlinearity, which is often called its activation function. Most often, units are interconnected in a feed-forward manner, that is, with interconnections that do not form any loops, as shown in Figure 2.7, as hidden part.

As described in [49], in the MLP structure, the neurons are grouped into layers. The first and last layers are called input and output layers respectively, because they represent inputs and outputs of the overall network. The remaining layers are called hidden layers. Typically, an MLP neural network consists of an input layer, one or more hidden layers, and an output layer, as shown in Figure 2.7.

As briefly described in [29][30], the advantage of neural networks include they involve long training times, high tolerance of noisy data as well as ability to classify patterns on which they have not been trained. They can be used when there is a little knowledge regarding the relationships between attributes and classes.

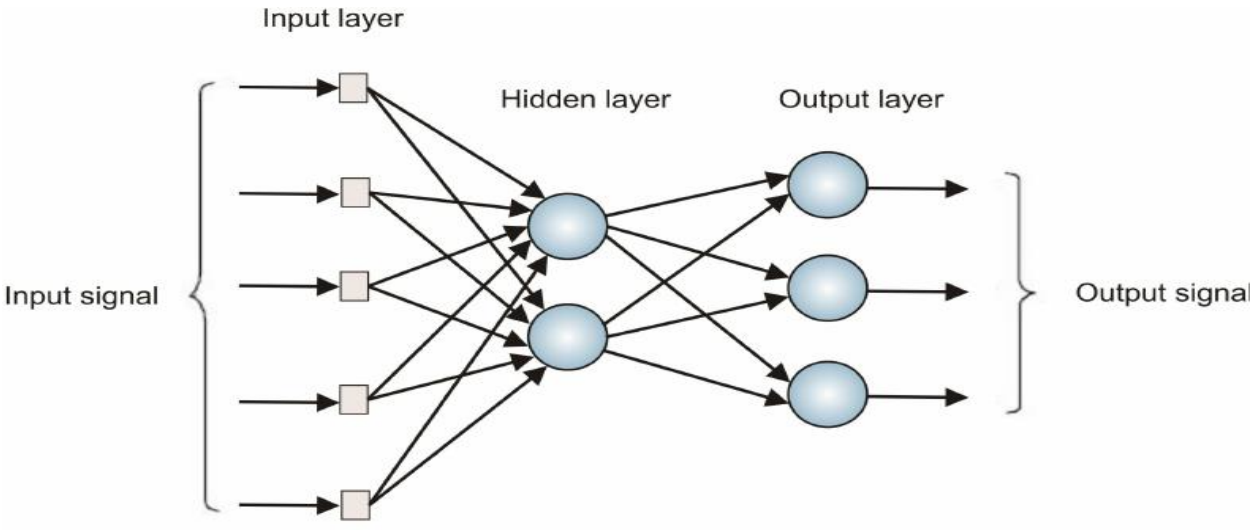


Figure 2.8. Structure of Multi-perceptron Layer

They are well suited for continuous-valued inputs and outputs, unlike most decision tree algorithms. They have been successful on a wide array of real-world data, including handwritten character recognition, pathology and laboratory medicine, and training a computer to pronounce English text. Neural network algorithms are inherently parallel; parallelization techniques can be used to speed up the computation process. In addition, several techniques have been recently developed for rule extraction from trained neural networks. These factors contribute to the usefulness of neural networks for classification and numeric prediction in data mining. However neural networks have been criticized for their poor interpretability. For example, it is difficult for humans to interpret the symbolic meaning behind the learned weights and of "hidden units" in the network.

2.9.5. Logistic Regression

Logistic regression: logistic regression measures the relationship between a response variable and independent variables, like linear regression, and belongs to the family of exponential classifiers [50]. Logistic regression classifies an observation into one of two classes [51], and this algorithm analysis can be used when the variables are nominal or binary. The data are analyzed after the discretization process for the continuous variables, like the Bayesian group.

2.9.6. Selection of Classification Algorithms

In data mining, it is crucial to use a comparison to determine the best classifier [52]. The classifier's performance is evaluated according to the following criteria [53]:

- **(i)** **Classification accuracy:** the ability of the model to correctly predict the label of class which is expressed as a percentage
- **(ii)** **Speed:** the speed refers to the time taken to set up the model
- **(iii)** **Robustness:** the ability to predict the model correctly even though the data has noisy observations and missing values
- **(iv)** **Scalability:** the ability of a model to be accurate and productive while handling an increasing amount of data
- **(v)** **Interpretability:** the level of understanding provided by the model
- **(vi)** **Rule Structure:** the understandability of the algorithms' rule structure

2.10. Applications of Data Mining

Today, different organizations are realizing the numerous advantages that come with data mining. And, organizations are using data mining to manage all phases of the customer life cycle (acquiring new customers, increasing revenue from existing customers, and retaining good customers) [23]. It provides clear and competitive advantage across a broad variety of industries by identifying potentially useful information from the huge amounts of data collected and stored.

Telecommunications and credit card companies are two of the leaders in applying data mining to detect fraudulent use of their services. Insurance companies are also interested in applying this technology to reduce fraud. Medical applications are another important area in which data mining can be used to predict the effectiveness of surgical procedures, medical tests or medications. Companies involved in financial markets use data mining to determine market and industry characteristics as well as to predict individual company and stock performance. Retailers are making more use of data mining to decide which products to stock in particular stores as well as to assess the effectiveness of promotions and coupons. Pharmaceutical firms are mining large databases of chemical compounds and of genetic material to discover substances that might be candidate for development as agents for the treatments of disease. In power utilities data mining can be used to forecasting power demand of customers.

2.10.1. Application of DM in Telecommunications

The data mining applications for any industry depend on two factors: the data that are available and the business problems facing the industry. Telecommunication companies around the world face escalating competition which is forcing them to aggressively market special pricing programs aimed at retaining existing customers and attracting new ones. The telecommunications industry has been one of the early adopters of data mining and has deployed numerous data mining applications [25]. This is most likely because telecommunication companies normally generate and store enormous amounts of high-quality data, have a very large customer base, and operate in a rapidly changing and highly competitive environment [25]. Telecommunication companies utilize data mining application to improve their marketing efforts, identify fraud, and better manage their telecommunication networks (network fault isolation and prediction). However, these companies also face several data mining challenges due to the enormous size of their data sets, the sequential and temporal aspects of their data, and the need to predict very rare events such

as customer fraud and network failures in real-time. And to respond to these challenges new methods have been developed and existing methods have been enhanced. The competitive and changing nature of the industry, combined with the fact that the industry generates enormous amounts of data, ensures that data mining will play an important role in the future in the telecommunications industry.

Some of the application areas of data mining in telecommunication industry, according to [9], are:

Fraud Detection

According to [26], [27], data mining helps to analyze customer's usage pattern to identify potentially fraudulent users which is also called subscription fraud. It can also support detection of attempts to gain fraudulent access to customers' accounts a method known as superimposed fraud. The other application of data mining in fraud detection is that it can discover unusual patterns that may seek special focus such as busy hour, frustrated repeated call attempts, switch and route congestion patterns.

The result of this study helps to characterize subscribers based on their usage behavior. Characterizing them support during different decision making. One of the goals that is supported by this is airtime credit loan. During an airtime credit the subscriber is evaluated whether it is qualified for loan or not by considering how it was behaving. The model used previous knowledge of defaulters before deciding the subscribers' fate. This helps to make sure the subscribers who are requesting the loan are those who are likely to pay back their debt. Which in the other hand reduces the risk of defaulting as well as increasing revenue of the company.

Network Analysis

Data mining techniques can support the provision of a good network management service. It helps to identify network fault by correlating multiple alarms to a single fault point, this makes the maintenance process much easier as the operator can focus on the root cause [27]. It can also simplify life by making network fault prediction. Some of the other important tasks that can be supported by data mining for network management purposes are identifying and comparing data traffic, system workload management, and resource usage management [26].

Customer Risk Prediction

Customer churn prediction helps the customer relationship management (CRM) to avoid customers who are expected to churn in future by proposing retention policies and offering better incentives or packages to attract the potential churners in order to retain them. Hence, the possible loss of the company's revenue can be prevented [28].

According to [29], there are two types of churners, they are voluntary and involuntary. Involuntary churners are those customers removed by the telecommunication service provider itself due to non-payment status, deception or non-usage. Meanwhile, voluntary churners are the customers that decide to terminate their service with the respective telecommunications service provider. Involuntary churners are easy to be recognized; however, the voluntary churners are more difficult to be identified. Generally, the customer churn problem in the Telecommunications industry is voluntary. Fuzzy algorithm is used to predict voluntary churners, a technique which gained popularity in telecom industry [30].

2.10.2. Credit Risk

Credit risk is the volatility of earnings of lenders [59]. This volatility is the result of financial losses recorded by a financial institution, due to the borrower's failure to repay a loan or otherwise meet a contractual obligation [60]. The term of risk has been introduced in many areas of social, economic and science life. First, the term of risk can be found in finance, banking, insurance and medicine [59]. Some of the authors are trying to build a common risk theory, but there is still the risk theory placed in the specific context e.g. insurance, banking [5]. There is a need of risk research for business related issues concerning crucial business processes. Especially the risk research has become valid for companies which the core of its business is a provision of services e.g. telecommunication services.

Credit Scoring

Data mining models can be used to mine the data on which they are built, but most types of models are generalizable to new data. The process of applying a model to new data is known as scoring. Credit scoring is the use of predictive modelling techniques to support decision making in lending. According to [61], credit scoring can be defined as a method that is used to predict the probability that a loan applicant or existing borrower will default or become delinquent [59]. The objective of

credit scoring is to help credit providers quantify and manage the financial risk involved in providing credit so that they can make better lending decisions quickly and more objectively.

Credit scores help to reduce discrimination because credit scoring models provide an objective analysis of a consumer's creditworthiness. This enables credit providers to focus on only information that relates to credit risk and avoid the personal subjectivity of a credit analyst [60]. They also help to increase the speed and consistency of the loan application process and allows the automation of the lending process [59]. As such, it greatly reduces the need for human intervention on credit evaluation and the cost of delivering credit [61]. With the help of the credit scores, financial institutions can quantify the risks associated with granting credit to an applicant in a shorter time.

Most credit scoring articles focused on enterprise credit score: using audited financial accounts variables and other internal or external, industrial or credit bureau variables, the enterprise score is extracted, rather than individual (consumer) credit score: the individual credit score uses variables like applicant age, marital status, income and some other variables and can include credit bureau variables.

Other articles on credit scoring compare data mining algorithms in order to identify the best performing model in terms of predictive capacity. Therefore, [59] concluded that the linear discriminant analysis is superior to the genetic algorithm, and neural networks have a lower predictive ability than the linear discriminant analysis.

The studies stated above focused on credit scoring on customers of bank or financial institutions. But this study will focus on telecom airtime credit loan and tries to develop a model that can accurately predict mobile prepaid user's airtime credit loan repayment likelihood.

Credit Risk in Telecommunication

For telecommunication sector especially crucial became the churn analysis and proper using methods of data analysis. In the recent years also significant from the point of view of operational processes for telecommunication companies became the credit risk analysis for the individual and business customers in the activation process. According to [62], using machine learning and data mining methods in pattern recognition as a new approach in order to find the riskiest customers, they presented credit scoring in a form of activation models, which are used to predict customers' debt as well as indicate clients with the highest, medium and smallest credit risk.

Technological development within telecommunications area has significantly progressed over recent years. This relates to the growth in the competitiveness within this sector. Telecommunication companies are trying to get as many customers as they can by offering them a lot of attractive deals and products [63]. Finding new customers becomes more difficult though when the market gets more saturated. Customers who respond to those offers are very valuable, because they generate profit for the company. Unfortunately, among them there are some who fail to pay their bills and put the company at risk of making considerable losses. To minimize this risk companies can take precautions by using data mining methods [59].

Since there was no prior study related to airtime credit in telecommunications, this study applies data mining techniques based on their usage patterns to predict airtime credit risk.

2.11. Review of Related Research Works

Although not fully related to airtime credit loan risk prediction, there are research works conducted in the context of Telecom risk prediction and customer classification in other areas such as financial institutions which are reviewed here. In this section, research works which have relevance to this study has been presented in two sections as local and international works.

Local research works

According to Melaku [23], a study that focused on Ethiopian Telecommunication Corporation CDMA (Code Division Multiple Access) telephone customers, applied data mining techniques to make behavioral segmentation. The researcher used classification and clustering data mining techniques on customers‟ database and adopted CRISP-DM model. The applied data mining tools are K-means clustering, decision tree (J48) and artificial neural network (feed forward backward propagation). For his research he used the CDMA CDR, bill data and customers profile data from "USHACOM" system of the corporation. As a result, using decision tree he managed to get 98.97% accurately classified and 98.62% using neural network. The numbers of customers wrongly classified are 103 and 139 using decision tree and neural network respectively. For both high valued and low valued customers decision tree resulted in better accuracy than that of neural network [23].

Another research work by Yigzaw [4], a study which focused on predicting postpaid user's usage bill non-payment for Ethiopian Telecom, indicated that location is the strongest predictor of non-

payment. They used different classification techniques such as decision trees, naive Bayes and logistic regression to predict nonpayment. They applied data mining classification techniques such as decision trees, nave Bayes, and logistic regression to predict nonpayment. The applied data mining techniques showed that a change in call usage pattern has a strong relationship with nonpayment of bill in future. In other words, they found out that a change in behavior or bill consumption is a key indicator of future non-payment. They also proposed prediction and early prevention of default could save the corporations revenue loss They reported that decision tree performed well, and the result of their model shows that billing changes were also significant predictors.

Henock [55], conducted his study targeted at testing the application of data mining techniques to support CRM activities at Ethiopian Airlines. He considers the Ethiopian Airlines frequent flyer programs database, which contains individual flight activity and demographic information of more than 22,000 program members. He applied the K-means clustering algorithm was used to segment individual customer records into clusters with similar behaviors and then the decision tree classification techniques were employed to generate rules that could be used to assign new customer records to the segments. The study revealed encouraging results of DM techniques in supporting CRM. Denekew [56] has also conducted his study in the same area to fill the gaps, which are left as a further study by [55]

Kumneger [5], conducted his study on the application of DM techniques to support CRM for the case of the Ethiopian Shipping Lines. He used customer profile file of ESL having more than 20,000 records. He applied K-Means clustering algorithm to segment individual customer records into clusters with similar behaviors and then decision tree classification techniques were employed to generate rules that could be used to assign new customer record to the segments. The study showed encouraging results in the applicability of DM techniques in supporting CRM.

Another research [58], on the application of data mining in credit data for predicting defaulters or inconsistent loan payers was conducted on united bank customers. In which a model was built to identify trends of good and bad patterns from historic data and the classification model performed well using J48 decision tree algorithm. Also, similar research by [59], has developed a prediction model of customer loyalty (Non loyal or Loyal) which supports microfinance institutions during

loan decision making using different DM techniques, among them a classification model of J48 is used to generate the rule.

International Research Works

Monica [26], assessed telecom customers credit risk from the moment of opening a customer account to the moment of closing an account due to non-payment. It used customers data base in a telecommunications company to apply different algorithms to identify and insolvency of debtors. Accordingly, models allow calculating profitability for a telecommunication company through measuring customer value and determining level of the risk. By using these models, we can divide customers into groups with high, medium, low risk and examine their features are as well as choices they make. In addition, these models can be used to prevent financial debt by, for example, implementing deposit policy. The deposit is usually imposed on certain groups where customers were assigned after data analysis.

Also, according to [26], activation models were used to predict customer's failure to pay after the debt collection process had finished. The most significant variable is a tariff, chosen by an individual customer. This variable divide whole population of customers into two groups where one is estimated to be ninety percent of good payers and the other one ten percent of bad payers. The best features that identified the risky subscribers according to this study are, customers who fail to pay their bill are those who activate for phone restriction and do not apply for phone installation address. And the existing do not need to show their financial reliability document while during contract agreement signature which shows lack of risky customers genuineness. This research was basically focused on applying different data mining techniques for postpaid subscribers [26].

Björkegren [29], used call record to predict credit defaulting of middle-income customers in developing countries. It demonstrated a method to predict default among borrowers without formal financial histories, using behavioral patterns revealed by mobile phone usage. This study [29], suggested that nuances captured in the use of mobile phones themselves can alleviate information asymmetries, and thus can form the basis of new forms of low-cost lending. Oliver [36], on the other hand used mobile phone activity data to compute credit score to predict customers defaulting

with no prior financial history. It has concluded that mobile phone data can be enough to generate a reliable customer score.

As it can be seen from the list of researches conducted so far none of them have focused on airtime credit service. So, it is a big research gap to explore the potential risk that comes with telecom airtime credit service loan non repayment likelihood. This study uses different data mining techniques to build a model that predicts the likelihood of customers non repayment.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1. Introduction

In this chapter the study describes the data mining goals, sources of data and the techniques that have been used in preprocessing and model building phases.

As it is mentioned at the beginning, the goal of this study is to build a model that can help to predict payment likelihood of telecom airtime credit loan users using data mining techniques. Thus, the experiment is conducted based on selected process model which is a six step CRISP-DM, which is an industry standard mining methodology discussed in section 2.3.2.

According to the six step CRISP-DM methodology, the life cycle of a data mining task consists of the following six phases [24]:

- Business understanding
- Data understanding
- Data preparation
- Modelling
- Model evaluation
- Implementation

3.1. Business Understanding

3.1.1. Research Context

Telecommunication is a dynamic business which requires a clever leadership to stay in such highly demanding environment. One of the recently common mechanisms to keep customers happy is providing advance airtime credit service to their subscribers. As discussed in chapter one, airtime advance credit service does not only increase customers satisfaction, but it also boosts revenue by providing convenience to users. Being said that there is also a risk that comes with the service, that is, the subscribers being default defaulters. The term defaulter is common in other financial businesses such as banking, micro credit associations, and any loan related services. Defaulters are those who took their granted loan but fails to repay it on the specified time limit.

Therefore, this study attempted to apply data mining technique to predict airtime credit risk. The result of the study answers the problem of defaulting by supporting during decision making of the loan provision.

3.1.1.1. Airtime Credit Service

Airtime advance credit service is a solution that creates convenience and flexibility allowing pre-paid subscribers to receive airtime value in credit to extend their use time once they run out of balance. The borrowed credit can be used like usual recharge to make voice calls, purchase data or SMS bundles, use value added services or transfer balance for another subscriber. Subscribers will repay loan on their next top up until their debt and service charge is fully cleared.

Airtime credit service increases user's satisfaction by monetizing out of credit subscribers. There are three parties involved in credit service. They are the operator or solution provider, the loaner also called VAS, and the subscriber. From the operators' point of view there are some benefits including but limited to, such as boosting average revenue per user also known as ARPU, churn reduction by increasing user loyalty in multi operator environment. Subscribers advantage by using this service are getting airtime any time anywhere, making important and urgent calls anytime, consistent sense of security (can reach out emergencies anytime), convenient, safe way to get airtime when traditional channels such as VC (scratch-able voucher cards) are not available.

3.1.1.2. Airtime Credit Service Flow

The following diagram shows a service flow of airtime advance credit service flow adopted in telecommunications [31], the case of Ethio Telecom.

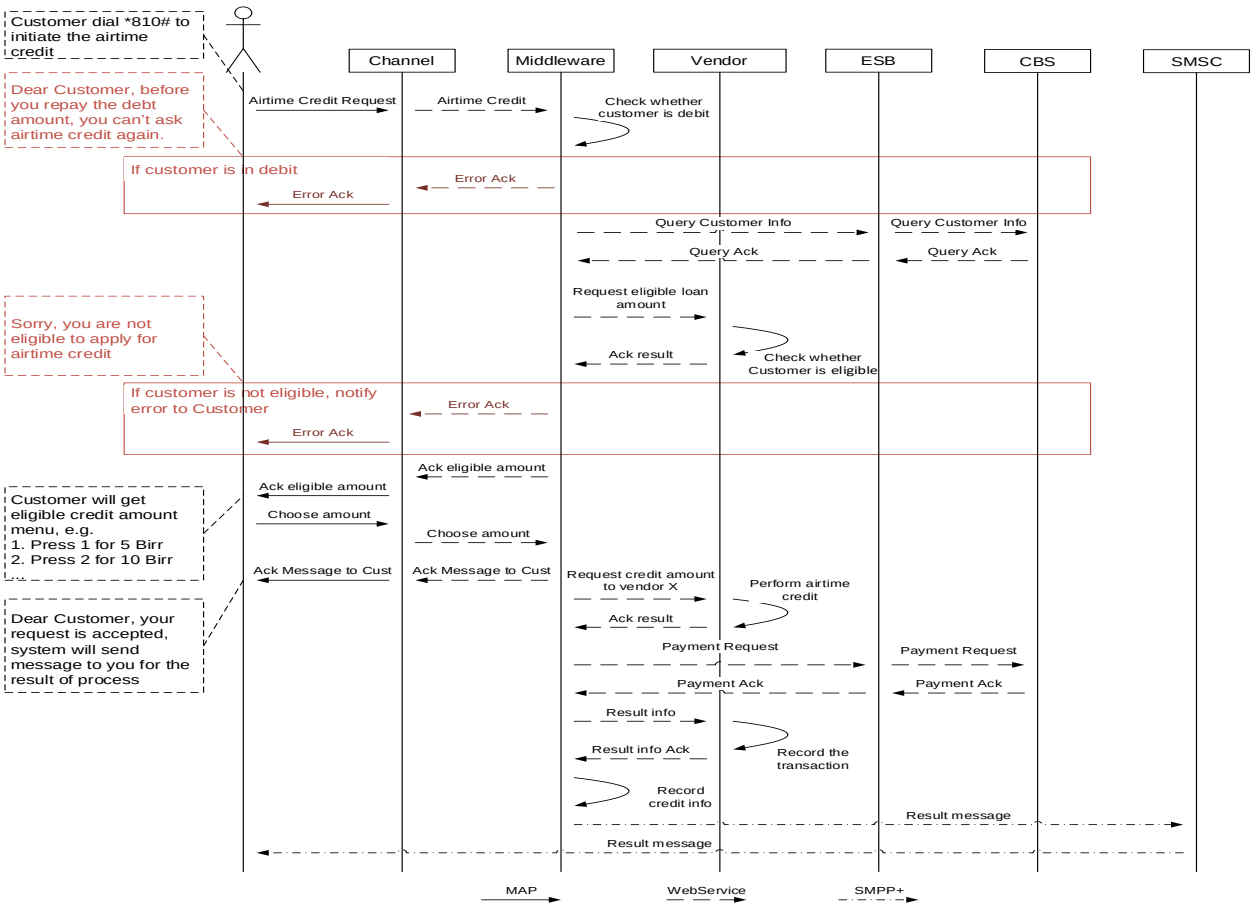


Figure 3.1. Service flow of airtime credit purchase

Customer initiate "Airtime Recharge" request, CN will trigger USSD/IVR command to USSDC/IVR

Steps for Airtime Credit service

1) The request will be forward to Middleware system
2) Middleware system will check whether customer reach the limit of debts, if it is, middleware will response error message to channel, then channel forward it to customer; if it isn't, go to step 3).
3) Middleware system send "Query Customer info" request to ESB.
4) EBS forward request to CBS.
5) CBS retrieve subscriber's "payment type", "balance", "status", "First activation date", "language ID", information, then response to ESB.
6) ESB forward response message to Middleware.

7) Middleware will send the customer information, such as: “payment type”, “balance”, “status” “First activation date” to the vendor X based on robin algorithm. If customer is not eligible, vendor will response to middleware, and middleware will response error message to channel, then channel forward it to customer; if customer is eligible, go to step 8)
8) Middleware system send eligible credit amount to channel
9) Channel show eligible credit amount to customer
10) Customer choose the credit amount
11) Channel forward customer’s chosen to Middleware
12) Middleware will response to the customer the request is accepted, and the result will be sent upon the process is completed. Middleware system send credit request to the same vendor X.
13) Vendor will confirm the requested credit amount to middleware.
14) Middleware system send “Payment” request to ESB
15) ESB forward request to CBS.
16) CBS perform the payment, then response the result to ESB
17) ESB forward response to Middleware
18) Middleware inform the payment result to the same vendor X.
19) Vendor X do the record for the credit request, and response to the middleware.
20) Middleware record credit information
21) Middleware will send short message to SMSC, and SMSC deliver the short message to customer for the result.

3.1.2. Problem Understanding

Currently more than half a million subscribers have not repaid their credit on time. those customers who failed to pay back their loan within three months after they get credit are considered as defaulters. Also, due to this the company has a risk of losing a lot of revenue by two means. One these customers may not need their number anymore which means they will not top up again. And the other is if they are not paying their debt it is a huge loss for the company as it does not have any guarantee to collect the loan. So, the company needs a solution to address this issue and protect its revenue as well as keeping its customers satisfied by the service it provides. This can be

achieved by implementing a prediction mechanism during loan decision while the customer's request airtime credit.

To tackle this and other related research questions raised in Section 1.2, various data mining techniques are applied to minimize the risk of airtime credit. Different classification techniques are implemented to classify customers as eligible (non-defaulter) or ineligible (defaulter) at the time of requesting airtime credit based on their usage pattern such as voice usage, internet usage, loan history, and customer profile. Furthermore, it is hoped that the company revenue could be protected, and decision making can be simplified. Finally, it will positively influence customers as they can only get loan that they can pay back easily which intern helps to create a better relationship between the customer and service provider

The business goal to be achieved is that to help the operator or the service provider during decision making by providing the required tool and technique. The data mining goal of this research is to build a predictive model that can support telecom operators who are providing airtime credit loan service in order to minimize subscribers defaulting by analyzing their usage behavior.

3.2. Data Understanding

As mentioned in section 1.4, this study used Ethio Telecom prepaid mobile service subscriber's data. The data used in this study was collected from the company business support systems (BSS) databases. BSS includes customer relationship management (CRM) system, convergent billing systems (CBS), and other subsystems supporting the business. Customer profile data was collected from CRM database which consists information such as customer name, address, billing information, contact number, subscriptions, birthdate, contract and other details about the customer. Likewise recharge information, loan information and loan repayment information was gathered from CBS BMP databases. Whereas, customer usage details such as call history including voice usage, internet usage and short message service (SMS) usage were collected from CBS CDR (Call detail record) databases. The detail of initially collected data is presented in the following subsequent subsections.

3.2.1. Customer profile data

As indicated above, the data used in this study was mainly collected from the company business databases. The systems consist of various databases each taking care of different customer and business information data.

Basic customer data is stored in CRM databases. The data consists of more than 63 fields. Some of them are described below:

Customer Type: Individual or Enterprise

Date of Birth: This is where customer birth date is filled

Language: This is a preference selected by customer in what language to receive notifications sent by ET

Education: This shows academic rank of the user it includes illiterate up to doctorate

Identification number: It can be a passport or identification card number received from kebele

Customer Name: Customer identifier as first name, last name and middle name

Gender: Gender of the customer, if individual, as Male or Female

Age: Age of the customer at the time of subscription

Place of Birth: Where the customer is born

Religion: Shows the religion of the customer.

Income: Monthly income of the customer

Address: subscriber residence which includes Zone/Region, woreda, kebele, house number

Primary Offer: Primary service subscribed, like 2G/3G/4G (Network type)

Service number: An identification that is used to use the service, usually a ten-digit number in incase of ET.

Payment Type: Prepaid, postpaid, hybrid,

The following attributes given in table 3.1 were selected for this study based on their relevance and data completeness.

Attribute	Description	Type
Service Number	Primary Identifier	Number
Subscription Date	To Calculate Network Age	Date
Date of Birth	Used to calculate customer age	Date
Gender	Sex	Nominal
Address	Location of subscriber	Nominal
Primary Offering	Subscribed service	Nominal

Table 3.1 Information of customer profile data detail from CRM database

3.2.2. Loan information data

Loan information shows the list of subscribers who take advance airtime loan which includes the time the loan is granted, the amount they received, repaid loan amount, remaining loan amount, poundage fee and other important parameters that is used to manage the loan.

The loan information data is collected from billing system databases. The data is stored both on physical database and memory database also known as GMDB (Global Memory Database). The data in GMDB shows real time paid and unpaid amount and other necessary fields. It is necessary to map with the physical database to get full information about the loan.

From loan data we get the list of customers who have paid their loan on time (according to the company timeline, a subscriber who repaid its loan in seven days after receiving the airtime credit is categorized as good loan score. Similarly, from the same data source we identify bad loan. Bad loan is considered when a subscriber fails to repay its loan fully with in three months after they were granted loan.

Loan data description is given below,

Service number: This is a primary identity used to uniquely identify a subscriber

Loan amount: advance amount taken by the subscriber

Loan occurrence date: The date loan is given to the subscriber

Repaid amount: Shows how much is repaid from the total loan taken

Operation type: Means of latest repayment. It can be by transfer or recharge.

Initial Loan poundage: The total amount the subscriber is subject to pay as a commission on the top of original loan

Loan balance: Remaining unpaid loan amount

Entry Date: latest loan repayment date

The attributes in table 3.2 are selected for this study, to identify defaulter (bad loan record) and good loan records. Here, loan occur date is used to calculate the time passed since the loan is granted. Whereas, loan balance shows subscribers remaining unpaid amount and used to identify bad and good loan records by considering entry date.

Attribute	Description	Type
Service number	Primary subscriber identifier	Number
Loan Occur Date	Loaned date	Date
Loan Balance	Unpaid amount	Number
Entry Date	Latest repayment date	Date

Table 3.2 Attributes of loan information detail data

3.2.3. Call Detail Records Data

Call detail records data consists of information's such as recharge log including amount recharged, channel used for recharge, and the date it occurred. Also, some of the other information stored in CDR database includes call transactions logs, data usage log, and SMS usage log. Some of the data stored in CDR database is described below.

Recharge Amount: The amount a customer recharged in a top-up.

Recharge Channel: The means of recharge, it can be via electronic means including e top-up and bank systems or by stretchable voucher card.

Date of recharge: The date balance recharge occurred.

The following table will give a detailed attributes description for recharge data which are used in the study.

Attribute	Description	Type
Service Number	Primary identifier	Number
Recharge amount	Amount recharged at an attempt	Birr (number)
Recharged channel	How the balance is recharged	ETOP VC
Date of recharge	Date of each top up	Date
Total Recharge amount	Amount in particular time frame	Total birr

Call Transaction log	Used to analyze subscriber voice call usage behavior	Nominal
Internet Transaction log	Used to analyze subscriber internet data usage behavior	Nominal
SMS transaction log	Used to analyze subscriber SMS usage behavior	Nominal

Table 3.3 Attributes of recharge history data detail

3.3. Data Preparation

As it was mentioned in chapter 1.4.3, more time was allocated to the preparation of the collected data set. In line with the requirements of the CRISP-DM methodology, this section form part of the entire data preparation which was done on the selected data. According to [1], the data preprocessing phase of DM process includes data selection, data cleaning, data construction, data integration and data formatting.

3.3.1 Data Selection

The data selection for this study is based on recommendation of domain experts of the company as well as the researchers. The data collected for analysis in this study was Ethio Telecom prepaid mobile subscriber's usage pattern of over the period of six months from July 2018 to January 2019. It consists of the following two data sets:

- Good loan (Non-Defaulters) data and
- Bad loan (Defaulters) data.

Final subscriber's data set is generated by combining both non-defaulters and defaulter's data. The data set consists of eleven columns and 86,024 rows. From these, 43,012 represents those of bad loan records and the other half, 43012, represents the good loan records. Equal data records are taken for both defaulters and non-defaulters to make the balancing equal.

Further the data instances are split into two sets. The splitting is made by thumb rule which says more than two third of data can sufficiently represent the whole data for training. Also, this idea is supported by WEKA data mining tool which sets default splitting to 66% for training and the remaining for testing. This is illustrated in the following table 3.5.

Total Data Set	**Data Set Allocated**	
86,024	Number of training data instance	Number of test data instance
	56775	29249

Table 3.4 Data set allocation into 66% training and 34% testing

The training data set is used to develop a model for predicting subscriber's loan repayment likelihood. The test data set is used for testing the accuracy of the model generated during the training phase.

3.3.2. Data cleaning

Data cleaning is one of the activities in data preparation phase of any data mining task and it must be done before going to drive new attributes from the basic ones. Data cleaning is removing of records that had incomplete, missing, duplicated, inconsistent data and irrelevant data under each attribute column [65]. There are different methods used to handle the missing values, such as ignoring the tuples, filling the missing values by using the modal value (for nominal and ordinal variables) and the mean (for continuous variable). And, in this research the missing data was filled by using WEKA 3.8.3 preprocessing facility "replace missing values with modes and means from the training data" and removed manually, duplicated attribute values that do not vary at all or that vary too much had been also removed. There are two strategies for dealing with outliers: detect and eventually remove outliers as a part of the preprocessing phase or develop robust modeling methods that are insensitive to outliers.

In this study, the cleaning and removal of incomplete data was done manually as part of a preprocessing phase. Income, occupation and religion are deleted based on expert's advice from the company. Income and occupation are only filled for less than five percent of customers during their subscription so it is logical to discard it, as it cannot represent most of the users. The attributes like customer type and payment mode are also removed as they do not have any contribution for study. Since the payment mode under consideration for this study is only prepaid and customer type on which the study focuses are only residential or individual customers. Due to such reasons these attributes are considered as outliers and not included in the data set as an attribute

3.3.3. Data Construction

The data collected consists of attributes such as birth date from which the researcher calculated the age of subscriber for those who have a missing age value. The original collected data consists of a field called 'age', but only some instances has been filled during contract agreement or at the time of subscription. Due to this it is important to calculate the age of remaining customers from their

birth date data. The other attributes constructed are voice usage, data usage and short message service usage behaviors of individuals.

These usage behaviors are analyzed over the period of six months, as previously mentioned. There is a KPI used in the company which is based on revenue generated from each service type (i.e. internet, voice and SMS) and taking into consideration the expert's advice the data was categorized into three groups based on their usage nature of each service. This categorization can simplify our classification model building. For example, a subscriber is considered as voice inclined or high voice service user if it uses 3G service and spends more than 60 percent of its recharged amount for voice calls, but its usage falls under the category of average if it uses between 40% to 60% percent of the money for voice calls and the usage is considered as low if only less than 40 percent is spent on this service. Whereas, if it is 4G LTE subscriber, which is expected to use more data service the percentage share is different. For example, the subscriber is categorized as high voice user if the money spent on this service is more than 39%, as average if between 16% to 39% and as low if less than 16% is spent on voice usage. The summary of categories for voice, internet/data and SMS usages is shown in Table 3.5.

Subscription	Service	Category (Based on amount spent in percentage)		
		LOW	AVERAGE	HIGH
3G	Voice	<40%	40% to 70%	>70%
	Data	<25%	25% to 55%	>55%
	SMS	<4%	4% to 10%	>10%
4G	Voice	<16%	16% to 39%	>39%
	Data	<40%	40% to 60%	>60%
	SMS	<3%	3% to 5%	>5%

Table 3.5 Categorizing subscriber's usage based on amount spent on a service

To brief Table 3.5 with an example: Let's say a 4G LTE subscriber spent 1000 Birr in six months, from which 710 Birr is spent on internet usage, 260 Birr on voice usage and the remaining 30 Birr is spent on SMS usage. The categorization of this subscriber usage is shown in Table 3.6.

Subscription	Total Spent	Service	Spent percentage	Category
3G	1000 Birr	Voice	71%	High
		Data	26%	Average
		SMS	3%	Low

Table 3.6 Categorizing a subscriber based on its usage

The other attribute changed to category was the age of subscribers. The category was made according to [39] which classified telecom customers into three age groups. The category is given in table 3.6.

Item	**Age of subscriber**		
Group	Up to 25 years old	From 25 to 55 years old	Above 55
Category Label	Young	Adult	Old

Table 3.7 Categorizing customers according to their age

Network age is the time calculated from date of activation or subscription of the service to the date the data is extracted for experiment. Since the data is in number and continuous, it is better to put in category for our classification purpose based on expert's recommendation as follows, Table 3.8.

Item	**Years since activation (Net Age/Network Age)**		
Group	Up to 2 years	Between 3 and 5 Years	Above 5 Years
Category Label	Small	Medium	Long

Table 3.8 Categorizing customers according to their service age

Average recharged amount is the amount spent in one month to use telecom services such as data, voice and SMS. The average amount (Recharged) is calculated from the money spent over the period of six months. Since the data is continues like age, we put it into three different categories according to their average top up. The categorization in Table 3.9 is based on recommendation of domain experts of the company. Here customers with less than 30 Birr average monthly recharge are excluded from the final data set as they are not eligible to request loan.

	Amount spent in a month (Average)		
Group	Below 50 Birr	Between 50 to 100	Above 100 Birr
Category Label	Low	Average	High

Table 3.9 Category of subscriber based on average recharge

3.3.4. Data Integration

Since the data are obtained from different databases, it is necessary to merge them and have a unified data to be used for this study. In order to merge the data, it is important to have a linking attribute between the data we collected from different sources. Hence, service number, also called access number, was used as a primary key to link all the information and put them together in one table. It is also inevitable to identify and remove the attributes that have no significance for the output of this investigation.

Some of the attributes that are removed from listed fields in section 3.2.1, are customer name as it doesn't affect the experiment by any means and for the sake of privacy of subscribers. The other parameters discarded are account number, place of birth, education, language preference can also be safely disregarded as they have no significance. Education level would have been useful, but for more than 96 percent of the subscribers the data is empty, due to this it was removed. Religion was also removed as it may cause unwanted result and it is basically sensitive case.

In a similar fashion there are attributes that were removed from the list in tables 3.2.2 and 3.2.3, remaining loan poundage, and reaming loan amount can be discarded as they can be calculated by deducting the repaid amount from the initial value which is the loan balance. Also, date of recharge for each top up can be removed as the total amount is used for the experiment

The merged data consists two sets. The first set is for those who didn't settle their loan within three months since they lent. These group of subscribers can be called as defaulters or bad loan records and the second set includes those subscribers who paid their loan on time also called good loan records (non-defaulters).

Non defaulters were identified as those subscribers who took loan at least three times and repaid all within seven days after they get the loan. These set of subscribers do not have unpaid loan. And the number of days is taken by the recommendation of the company experts based on their internal working guideline. Whereas, defaulters are identified as those who did not repay their loan within three months after they get airtime advance loan.

Thus, from the original data that had more than 150 attributes collected from the three databases, only eleven important parameters are taken for this study. The inclusion of the address in the study

is to highlight whether the geographic location of a subscriber can affect the loan repayment likelihood.

Service number is used as a primary key to collect and combine the data. But due to privacy reasons and its insignificance for this study, it is finally discarded.

Therefore, the customer data collected and integrated from different tables of the database and ready for data mining techniques to be undertaken in this research looks like the following:

Attribute	**Data Type**	**Description**
Service Number	Number	The number to identify the customer uniquely
Gender	String	Describes the gender of customer
Age	Number	Describes the age of customer
Network Age	Number	Describes how long it is since the service is activated
Location	String	A region or zone (in case of Addis Ababa) in which the service is provided to the customer
Recharge Channel	String	Airtime recharge method used by subscribers
Average Recharge amount	Number	Average recharge amount over the period of six months
Date Usage	String	Amount spent for internet service usage
Voice Usage	Number	Amount spent for voice service usage
SMS Usage	Number	Amount spent for SMS service usage
Loan Status	Number	The loan status of subscriber

Table 3.10 Final data set attributes to be used for experiment

The data collected for analysis in this study was Ethio Telecom prepaid mobile subscriber's usage pattern of over the period of six months from August 2018 to January 2019.

3.3.5. Data formatting

In order to meet the WEKA's data formatting requirements, some the extracted data from memory database for this study has been prepared on a excel sheet. The data which is in a excel has been saved using a comma separated value (csv) format. While the other data which is collected from physical database has been provided in excel format with many unwanted attributes. Then it is imported to oracle database for further processing and exported in csv format.

Although WEKA's data storage method is ARFF format, it can also read csv format directly, thus saves time as there is no need to create the ARFF file. All the selected data from the BSS billing databases and CRM database has been saved using the csv format.

3.4. Modelling

Model building phase of data mining is the process of providing the preprocessed data to the selected classification algorithm and select the model that shows better performance. There are several tasks involved in this phase. Some of the tasks include selection of modeling technique, test design and building and assessing of the best model.

3.4.1. Selection of Modelling Techniques

For this study, 3.8.3 version of Waikato Environment for Knowledge Analysis (WEKA) has been selected to generate models using Ethio Telecom prepaid mobile subscriber's data. Tied to the choice of WEKA further meant an identification of one supervised and three unsupervised method are used in this study. WEKA has a variety of algorithms which can be specifically chosen to solve and suit the needs of any company such as KD tree, BF tree, J48, J48graft, Naïve Bayes, SVM, MLP, K-Means and others. WEKA's ability to operate on csv format made is easy to convert the selected subscriber's data set to suit this important requirement. The variety, quality, and flexibility of visualization tools have strongly influenced the usability, interpretability, and attractiveness of a data mining system like WEKA for this study [66].

In this study, a C4.5 decision tree classifier using j48 algorithm, Naïve Bayes algorithm, multilayer perception (MLP) – ANN and Logistic regression are implemented, and their classification accuracy is compared against each other.

The motivation for choosing C4.5 decision tree learner using j48 algorithm, Naïve Bayes, Logistic Regression and Multilayer Perception (MLP) was discussed in section 2.10., also they are mostly used in many telecommunication related data mining studies.

3.4.2 Data Mining Test Design

In this study, for classification purpose the data is divided into training and testing data sets. Using the data set allows the generated models to be compared with each other. Classification model that shows better performance accuracy has been selected from the three classification algorithms namely J48, Naïve Bayes, Logistic Regression and MLP.

In this research, the analysis and interpretation of the results was made by the researcher and domain experts. Domain experts have been involved in the whole process of classification and segmentation results. Hence, the final classification provides a good knowledge for designing and implementation of better predictive strategies in the company that enhances the airtime loan service and improves risk of defaulting.

3.4.3. Model Building

The model building process of this study also consists of three activities namely: attribute selection, applying model building techniques, and identifying the best performing model.

The data collected has attributes which were selected as shown in Table 3.9. The data prepared based on the selected attributes are shown in Table 3.10. This data is not transformed to categorial basis for some columns. But for classification analysis it is a good idea to prepare data in a categorical form which is presented in Table 3.11. to use it for the model building process.

NETAGE	GEND.	AGE	RECH	LOC.	VOIC.	DATA	SMS	CHANL	CLASS
27-Dec-17	Male	27	75	EEAZ	23.73%	61.09%	15.18%	VC	DEF
16-Aug-08	Female	39	30	NR	89.00%	0.00%	11.00%	ETOP	ND
9-Oct-08	Male	32	175	NWR	61.00%	29.78%	9.22%	BOTH	ND
25-Nov-08	Female	39	150	WR	78.87%	11.62%	9.51%	VC	ND
7-Jun-15	Male	58	35	EAAZ	42.19%	57.81%	0.00%	ETOP	DEF
23-Jun-10	Male	23	740	NAAZ	92.00%	3.90%	2.10%	BOTH	ND
29-Jun-10	Male	36	50	WAAZ	52.00%	48.00%	0.00%	VC	ND

23-May-11	Female	24	60	CAAZ	43.00%	51.12%	3.88%	ETOP	ND
15-Dec-17	Female	36	155	EAAZ	10.00%	87.12%	2.88%	BOTH	DEF
18-Sep-17	Male	36	85	NR	47.93%	0.00%	52.17%	VC	DEF

Table 3.11 Before the Data is categorized for 3G WCDMA subscribers.

After some attributes such as network age, age, voice, data and SMS usage are transformed to category, the final data set used for building classification models looks like the following Table 3.11.

NETAGE	GENDER	AGE	RECH.	LOC.	VOICE	DATA	SMS	CHANL	CLASS
2	Male	Adult	AVG.	EEAZ	LOW	HIGH	HIGH	VC	DEF
3	Female	Adult	LOW	NR	HIGH	LOW	HIGH	ETOP	ND
12	Male	Adult	HIGH	NWR	AVG.	AVG.	AVG.	BOTH	ND
12	Female	Adult	HIGH	WR	HIGH	LOW	AVG.	VC	ND
4	Male	Old	LOW	EAAZ	AVG.	HIGH	LOW	ETOP	DEF
9	Male	Youth	HIGH	NAAZ	HIGH	LOW	LOW	BOTH	ND
9	Male	Adult	LOW	WAAZ	AVG.	AVG.	LOW	VC	ND
8	Female	Youth	AVG.	CAAZ	AVG.	AVG.	LOW	ETOP	ND
1	Female	Adult	HIGH	EAAZ	LOW	HIGH	LOW	BOTH	DEF
2	Male	Adult	AVG.	NR	AVG.	LOW	HIGH	VC	DEF

Table 3.12 After data is categorized for classification for 3G WCDMA subscribers

3.4.4. Building and assessing models using data mining

From the classification models that were generated using the classifier J48, MLP and Naïve Bayes on training data sets, visual representation is used to analyze the models. In addition, the testing data sets, which produced results from the training models, make use of the confusion matrix to test the accuracy of the results as per the subscriber score (defaulter or non-defaulter). In this study, the output emanates from the different model building activities which are iterative in their nature. They are, going to be analyzed and presented for decision making.

3.5 Evaluation Phase

At this stage, the models that were generated after the application of DM tools and techniques on subscriber's data are evaluated in terms of achieving the initially stated aims and objectives.

Models generated are measured against each other to select the better performing one. The evaluation is done using confusion matrix by comparing accuracy, recall, precision, ROC area, and f-measure.

3.5.1. Evaluating Data Mining Results

The study investigates the generated models and determine the generated knowledge from them with respect to the goal set at the beginning. In addition, the test result is used to determine the prediction accuracy of the generated models. There are four classification algorithms used in this study. Each model was tested several times to choose the best performing representative model based on accuracy. Then each best model from the four classifiers are compared against one another by using performance measurement techniques mentioned earlier such as accuracy, precision, recall and f-measure.

3.6 Application of discovered knowledge

After conducting the experiment and selecting a better performing model, the discovered knowledge was implemented by integrating to the airtime loan service system. The result discovered shows there are some important features that can be used to identify airtime credit defaulters. This helps during decision making of the loan whether to grant the request or decline it. Accordingly, a Java implementation code is developed to support this decision and it shows that knowledge discovered can successfully predict defaulters as well as the non-defaulters.

3.7 Summary

This study is guided using CRISP-DM methodology, an industrial-standard mining methodology which ensures that all the stages in the mining process are followed and adhered to [65]. Further to this, WEKA tools rich algorithms means it is possible to conduct different experiments using different classifiers as much as needed. The next chapter gives detailed account of the output results obtained from the application of data mining processes on Ethio Telecom Prepaid mobile service subscribers data.

CHAPTER FOUR

EXPERIMENTATION

4.1. Introduction

This chapter presents the output of both the model building and evaluation phases of the CRISP-DM methodology followed by this study. These results include the output of the J48 Decision Tree, Naïve Bayes, Logistic Regression and Multilayer Perceptron algorithms that were used for classification. The classification models were trained using the training data sets as discussed in Chapter 3.4. These models were then tested against the test datasets for which the confusion matrixes were used to determine the accuracy of each model. This was an iterative process and the final models that were analyzed and interpreted are presented in this chapter.

4.2. Classification Model Building

In this study, experimentation of classification model is conducted using three different algorithms: J48 decision tree classifier, Naïve Bayes classifier, Logistic Regression, and multilayer Perceptron classifier. These algorithms are experimented by tuning different parameters in WEKA to get the best classification model. Thus, a data set of 86024 instances extracted to equally represent both classes (defaulter and non-defaulter) and preprocessed to train and test the model is used.

Among the four test options in WEKA tool, 10-fold cross validation and percentage split have been used to build and test the models. The output of the experiments is compared against one another with respect to performance measures such as Accuracy, Precession, Recall, ROC Area (AUC), and F-Measure.

The following snapshot, Figure 5.1, shows when the prepared data is loaded to WEKA tool for the first time.

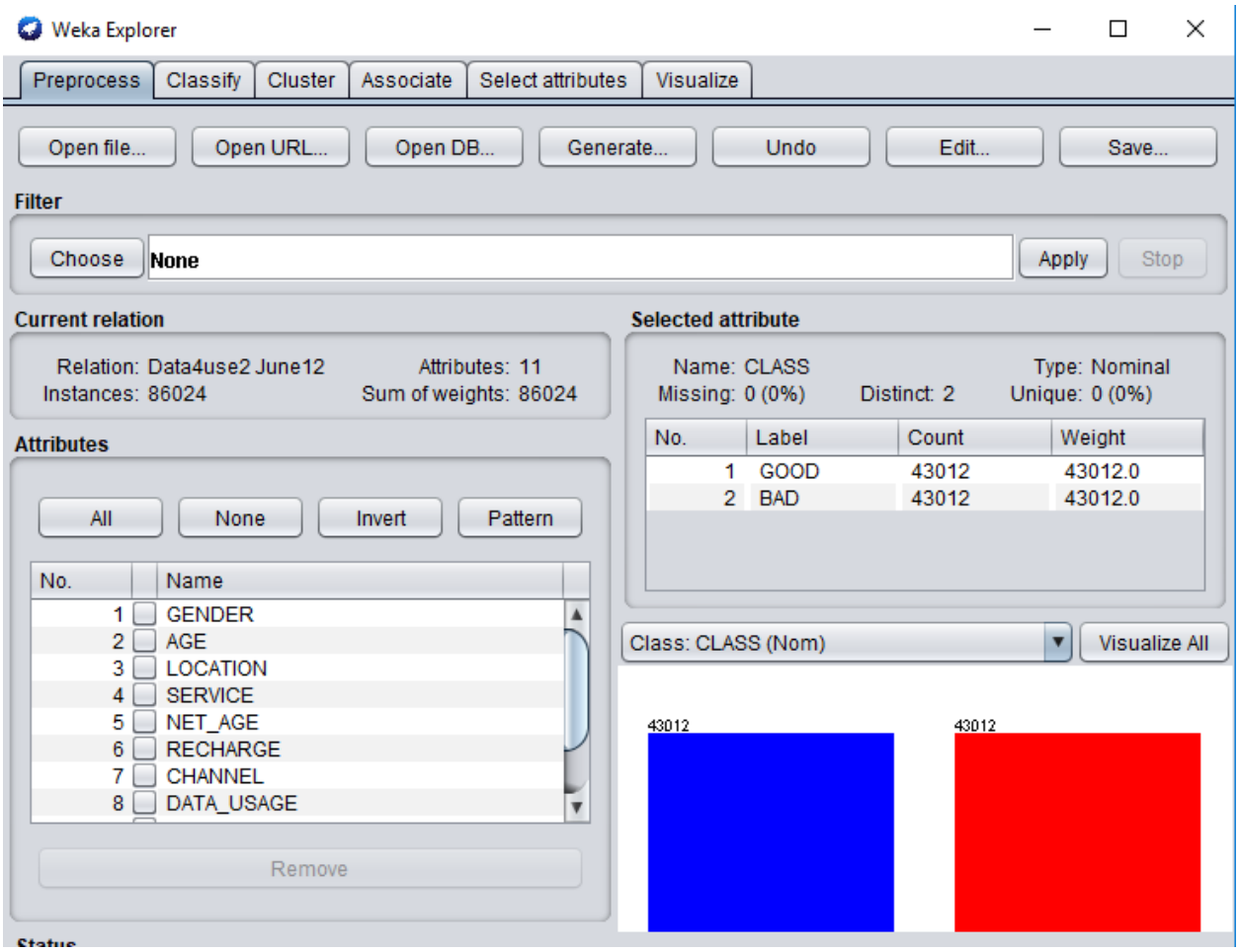


Figure 4.1 Snapshot showing when first time data is loaded to WEKA tool

4.2.1. J48 Decision Tree Classifier

Decision tree is known as state of the art in data mining. The reason this classifier is chosen is that it has a very good performance for large data size analysis [29], the result can be easily interpreted and they can be applied to categorical data and also most researches have chosen it as the best classifier when it is compared with other algorithms [30].In order to come up with the best J48 classification model, both the post pruning and pre pruning methods are implemented. These pruning mechanisms are labeled in Weka as ConfidenceFactor (CF) and MinNumObj (MNO) respectively. In this study these two parameters shown on Table 5.1 are tuned for different values including the default to optimize the classification model.

Parameter	Description	Default Value
ConfidenceFactor (CF)	The confidence factor used for pruning (smaller values incur more pruning)	0.25

MinNumObj (MNO)	The minimum number of instances per leaf	2

Table 4.1. Description of parameters to be tuned in J48 classification modeling

This experiment is conducted in the J48 decision tree using the 10-fold cross validation and the percentage split classification test option in Weka, with different ConfidenceFactor (CF) and MinNumObj (MNO) values including the default parameters shown in Table 5.1. The default value of percentage split, which is 66% for training and 34% for testing is applied. A data set of eleven selected attributes and 86026 instances has been used in this experiment.

The following parameters are used throughout the experiments [10],

Accuracy is the percentage of correct predictions. According to confusion matrix, it can be calculated as

$$AC = \frac{TN+TP}{\text{TP+FP+TN+FN}} \quad \text{..... (4.1)}$$

Where

TN is the true negative, i.e., instances that are correctly classified as negative.

TP is the true positive, i.e., instances that are correctly classified as positive

FP is the false positive, i.e., instances that are predicted to be positive but should have been classified as negative.

FN: false negative, i.e., instances that are predicted to be negative but should have been classified as positive.

Precision is the ratio of correct prediction to the sum of true and wrong prediction. It can be calculated as,

$$Precision = \frac{TP}{\text{TP+FP}} \text{..... (4.2)}$$

Recall is the ratio of correct prediction to the sum of true prediction and false negative prediction. It is calculated can be calculated as,

$$Recall = \frac{TP}{\text{TP+FN}} \text{ ……...… (4.3)}$$

F-Measure can be calculated as follows,

$$F - Measure = \frac{2(Precision*Recall)}{\text{Precision+Recall}} \text{ …….. (4.4)}$$

EXPERIMENT 1.

By setting the value of CF=0.25and MNO=2 to their default values and using 10-fold cross validation test option in WEKA, the snapshot of this experiment is shown in Annex 1. Here and for all the next subsequent tests the class attribute is set to CREDIT (having values of BAD=Defaulter, GOOD=Non-Defaulter).

CF= 0.25 (Default)

MNO =2 (Default)

Seed = 1

The Experiment Result is as shown Below (WEKA Output: Annex 1):

```
=== Summary ===

Correctly Classified Instances       84510               98.24    %
Incorrectly Classified Instances      1514                1.76    %
Kappa statistic                          0.9648
Mean absolute error                      0.03
Root mean squared error                  0.1246
Relative absolute error                  6.0046 %
Root relative squared error             24.9279 %
Total Number of Instances            86024

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.983    0.018    0.982      0.983    0.982      0.965    0.991     0.989     GOOD
                 0.982    0.017    0.983      0.982    0.982      0.965    0.991     0.989     BAD
Weighted Avg.    0.982    0.018    0.982      0.982    0.982      0.965    0.991     0.989

=== Confusion Matrix ===

     a     b   <-- classified as
 42287   725 |     a = GOOD
   789 42223 |     b = BAD
```

Figure 4.2. Experiment result of J48 Decision tree.

Number of Leaves: 292

Size of the tree: 365

Time taken to build model: 0.34 seconds

Correctly Classified Instances 84510 98.24 %

Incorrectly Classified Instances 1514 1.76 %

According to equation 4.1.,

Accuracy = (TP+TN)/ (TP+FP+TN+FN)

= (42287+42223)/ (42287+725+789+42223) = 0.9824 =**98.24%**

Therefore, the overall accuracy of the model was measured at 98.24 %. As illustrated in the confusion matrix below.

=== Confusion Matrix ===

a b <-- classified as

TP=42287 FP=725 | a = GOOD

FN=789 TN=42223 | b = BAD

The detailed accuracy by class, having parameters such as Precision, Recall, F-Measure, ROC-Area, and weighted result are given below.

=== Detailed Accuracy by Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.983	0.018	0.982	0.983	0.982	0.965	0.991	0.989	GOOD
	0.982	0.017	0.983	0.982	0.982	0.965	0.991	0.989	BAD
Wd Avg.	0.982	0.018	0.982	0.982	0.982	0.965	0.991	0.989	

This result shows that the J48 learning algorithm as per this set up scored an accuracy of 98.24%. From the total training set 84510 instances were correctly classified, while 1514 (1.76%) instances were incorrectly classified.

After changing the value of seed to 2,10,100… the result is still the same. This shows that the seed does not affect the performance of the model. Therefore, seed is set to its default value (1) for the remaining experiments made.

EXPERIMENT 2:

By setting the value of CF=0.25 (Default) and MNO=5 values and using 10-fold cross validation test option in WEKA. The snapshot of this experiment is shown **Annex 2**.

CF=0.25 (Default), MNO=5

The Experiment Result is as shown Below:

Number of Leaves: 292

Size of the tree: 365

Time taken to build model: 0.36 seconds

Correctly Classified Instances 84464 98.1866 %

Incorrectly Classified Instances 1560 1.8134 %

The overall accuracy of the model was measured at 98.1866 %. As illustrated in the confusion matrix below.

=== Confusion Matrix ===

a	b	<-- classified as
42256	756	a = GOOD
804	42208	b = BAD

The detailed accuracy by class, having parameters such as Precision, Recall, F-Measure, ROC-Area, and weighted result are given below.

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.982	0.019	0.981	0.982	0.982	0.964	0.990	0.989	GOOD

	0.981	0.018	0.982	0.981	0.982	0.964	0.990	0.987	BAD
Wd Avg.	0.982	0.018	0.982	0.982	0.982	0.964	0.990	0.988	

Other experiments are done by setting CF=0.25 (default) and changing the value of MNO for 10-fold cross validation test method, and the result is as follows;

CF=0.25 (Default), MNO=10

Number of Leaves: 176

Size of the tree: 224

Time Taken to build Model: 0.47 seconds

Accuracy: 98.0936 %, Weighted Precision= 0.981, Recall= 0.981, ROC Area = 0.990,

F-measure= 0.981.

CF=0.25 (Default), MNO=100

Number of Leaves: 79

Size of the tree: 98

Time taken to build model: 0.33 seconds

Overall Accuracy: 97.474 %, Weighted Precision= 0. 975, Recall= 0. 975, ROC Area= 0.988, F-measure= 0. 975.

As it can be seen from the above tests, increasing the value of MNO while keeping CF at its default value does not further improve the accuracy of the model.

EXPERIMENT 3:

By setting the value of CF=0.5 and MNO=2 (Default) to their default values and using 10-fold cross validation test option in WEKA, the snapshot of this experiment is shown in **Annex 3**.

CF=0.5 MNO=2

Number of Leaves: 439

Size of the tree: 538

Time taken to build model: 0.41 seconds

Correctly Classified Instances 84622 98.3702 %

Incorrectly Classified Instances 1402 1.6298 %

The overall accuracy of the model was measured at **98.3702%.** As illustrated in the confusion matrix below.

=== Confusion Matrix ===

a b <-- classified as

42397 615 | a = GOOD

787 42225 | b = BAD

The detailed accuracy by class, having parameters such as Precision, Recall, F-Measure, ROC-Area, and weighted result are given below.

=== Detailed Accuracy by Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.986	0.018	0.982	0.986	0.984	0.967	0.995	0.993	GOOD
	0.982	0.014	0.986	0.982	0.984	0.967	0.995	0.993	BAD
Wg Avg.	0.984	0.016	0.984	0.984	0.984	0.967	0.995	0.993	

EXPERIMENT 4

By setting the value of CF=0.75 and MNO=2 (Default) and using 10-fold cross validation test option in WEKA, the snapshot of this experiment is shown in **Annex 4.**

CF = 0.75 and MNO =2

Number of Leaves: 765

Size of the tree: 910

Time taken to build model: 7.43 seconds

Correctly Classified Instances 84788 98. 5632%

Incorrectly Classified Instances 1236 1. 4368%

The overall accuracy of the model was measured at 98**.** 5632%. As illustrated in the confusion matrix below.

=== Confusion Matrix ===

a b <-- classified as

42563 449 | a = GOOD

787 42225 | b = BAD

The detailed accuracy by class, having parameters such as Precision, Recall, F-Measure, ROC-Area, and weighted result are given below.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.990	0.018	0.982	0.990	0.986	0.971	0.996	0.995	GOOD
	0.982	0.010	0.989	0.982	0.986	0.971	0.996	0.996	BAD
Wd Avg.	0.986	0.014	0.986	0.986	0.986	0.971	0.996	0.995	

The other experiments done using 10-fold cross validation test option are as follows:

CF=0.5, MNO=5

Time taken to build model: 0.4 seconds

Overall Accuracy: 98.3144%, Weighted Precision= 0. 983, Recall= 0. 983, ROC Area= 0.995, F-measure= 0. 983.

CF=0.5, MNO=10; Time taken to build model: 0.53 seconds

Overall Accuracy: 98.1947%, Weighted Precision= 0.98982, Recall= 0. 982, ROC Area= 0.995, F-measure= 0. 982.

Scheme: C 0.75 -M 5

Time taken to build model: 6.75 seconds

Overall Accuracy: 98.3458 %, Weighted Precision= 0.983, Recall= 0. 983, ROC Area 0.996, F-measure= 0.983.

CF=0.75 and M=10

Time taken to build model: 5.14 seconds

Overall Accuracy: 98.2156 %, Weighted Precision= 0. 982, Recall= 0. 982, ROC Area = 0.995, F-measure= 0. 982.

EXPERIMENT 5

By setting the value of CF=0.25 and MNO=2 (Both Default), and using percentage split test option in WEKA, the snapshot of this experiment is shown in **Annex 5.**

Setting WEKA parameters as CF=0.25 and MNO=2, the result shows:

Number of Leaves: 292

Size of the tree: 365

Time taken to build model: 0.44 seconds

Correctly Classified Instances 28708 98.1537 %

Incorrectly Classified Instances 540 1.8463 %

The overall accuracy of the model was measured at 98.1537 **%**. As illustrated in the confusion matrix below.

=== Confusion Matrix ===

a b <-- classified as

14321 248 | a = GOOD

292 14387 | b = BAD

The detailed accuracy by class, having parameters such as Precision, Recall, F-Measure, ROC-Area, and weighted result are given below.

=== Detailed Accuracy by Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.983	0.020	0.980	0.983	0.981	0.963	0.990	0.985	GOOD
	0.980	0.017	0.983	0.980	0.982	0.963	0.990	0.986	BAD
Wd Avg.	0.982	0.018	0.982	0.982	0.982	0.963	0.990	0.986	

EXPERIMENT 6.

By setting the value of CF=0.5 default and MNO=2 (Default), and using Percentage split test option in WEKA, the snapshot of this experiment is shown in **Annex 6**.

Number of Leaves: 439

Size of the tree: 538

Time taken to build model: 0.42 seconds

Correctly Classified Instances 28736 98.2495 %

Incorrectly Classified Instances 512 1.7505 %

The overall accuracy of the model was measured at 98.7486 %. As illustrated in the confusion matrix below.

=== Confusion Matrix ===

a	b	<-- classified as
14349	220	a = GOOD
292	14387	b = BAD

The detailed accuracy by class, having parameters such as Precision, Recall, F-Measure, ROC-Area, and weighted result are given below.

=== Detailed Accuracy by Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class

	0.985	0.020	0.980	0.985	0.982	0.965	0.994	0.991	GOOD
	0.980	0.015	0.985	0.980	0.983	0.965	0.994	0.992	BAD
Wd.Avg.	0.982	0.017	0.983	0.982	0.982	0.965	0.994	0.992	

EXPERIMENT 7.

By setting the value of CF=0.5 and MNO=5, and using Percentage split test option in WEKA, the snapshot of this experiment is shown in **Annex 7**.

Number of Leaves: 385

Size of the tree: 476

Time taken to build model: 0.4 seconds

Correctly Classified Instances 28722 98.2016 %

Incorrectly Classified Instances 526 1.7984 %

The overall accuracy of the model was measured at 98.2016 %. As illustrated in the confusion matrix below.

=== Confusion Matrix ===

a	b	<-- classified as
14338	231	a = GOOD
295	14384	b = BAD

The detailed accuracy by class, having parameters such as Precision, Recall, F-Measure, ROC-Area, and weighted result are given below.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.984	0.020	0.980	0.984	0.982	0.964	0.994	0.992	GOOD
	0.980	0.016	0.984	0.980	0.982	0.964	0.994	0.993	BAD
WdAvg.	0.982	0.018	0.982	0.982	0.982	0.964	0.994	0.992	

Other experiment conducted with the percentage split test option are as follows,

CF=0.75, M=5; Time taken to build tree = 6.33 seconds

Overall accuracy=98.1845%, Weighted precision=0.982, Recall=0.982, F-measure=0.982 and ROC Area=0.994

As it can be seen from these experiments, the overall accuracy starts to drop as the value of both parameters increases. This shows that changing the two values further does not improve the performance of the model.

The summary of the experiment results is presented in Table 4.2 to compare and select the best classification model of the J48 classification algorithm.

Experiment No.	Tuned Parameters		Test Mode	Number Of Leaves	Size of Tree	Time Taken For Build (seconds)	Accuracy
	CF	MNO					
1	0.25	2	10-Fold	292	365	0.34	98.24 %
			Percentage split	292	365	0.44	98.1537 %
2	0.25	5	10-Fold	293	365	0.36	98.1866 %
			Percentage split	293	365	0.38	98.1127 %
3	0.25	10	10-Fold	176	224	0.47	98.0936 %
			Percentage split	176	224	0.39	98.0341 %
4	0.5	2	10-Fold	439	538	0.41	98.3702 %
			Percentage split	439	538	0.42	98.2495
5	0.5	5	10-Fold	385	476	0.4	98.3144 %
			Percentage split	385	476	0.4	98.2016 %
6	0.5	10	10-Fold	308	381	0.53	98.1947 %
			Percentage split	308	381	0.43	98.0272 %
7	**0.75**	**2**	**10-Fold**	765	910	7.43	**98. 5632 %**
			Percentage split	765	910	8.91	98. 4819 %

8	0.75	5	10-Fold	632	756	6.75	98.3458 %
			Percentage	632	756	6.33	98.1845 %
9	0.75	10	10-Fold	460	553	5.14	98.2156 %
			Percentage split	460	553	5.22	98.0341 %

Table 4.2. Summary of experiment for the J48 algorithm using various parameter setting

As we can see from the above table, different parameters are tuned in order to optimize the J48 classifier. On this experiment, the J48 decision tree algorithm has a better classification accuracy which is 98.4028 % when MinNumObj = 2 and ConfidenceFactor = 0.75 with 10-fold cross validation test option is used and this model can be selected as the best classification model to represent the J48.

Algorithm	Test Option	Accuracy	Time (Sec)	Precision	Recall	ROC Area	F-Measure	Class
J48 Decision Tree	10-fold cross validation	98.5632%	7.43	0.982	0.990	0.996	0.986	GOOD
				0.989	0.982	0.996	0.986	BAD
Weighted Average		**98.5632%**	**7.43**	**0.986**	**0.986**	**0.996**	**0.986**	

Table 4.3. Confusion Matrix for the selected J48 classifier (CF=0.75, MNO=2 and Test Option = 10-fold cross validation)

As shown in Table 4.3, the overall accuracy of the selected classifier is 98.5632%. This means J48 Decision tree algorithm with 10-fold cross validation test option has a better classification performance in identifying the ‘Bad’ class or defaulters.

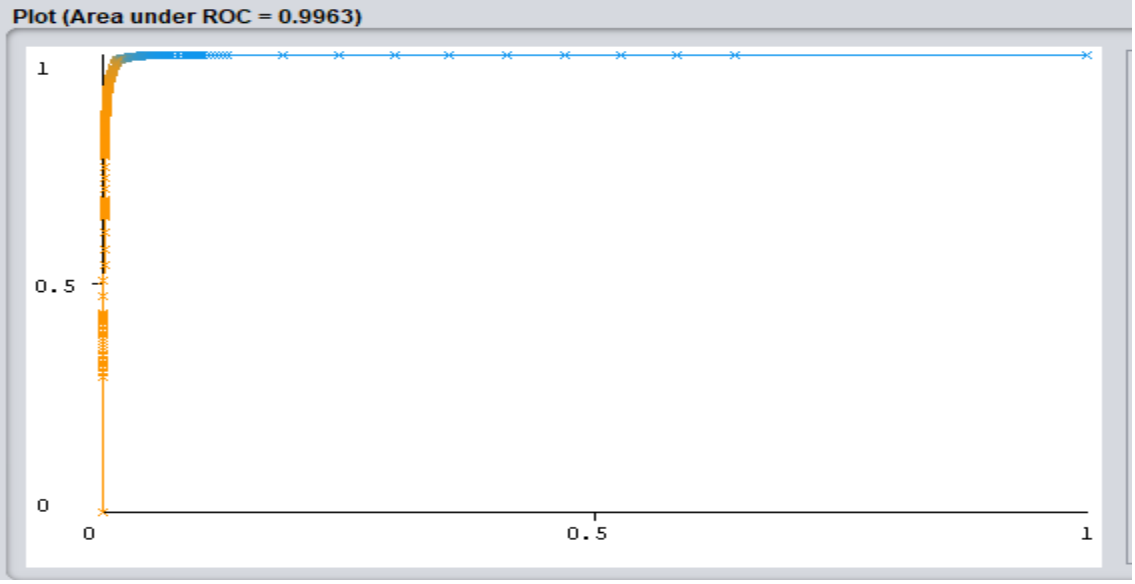


Figure 4.3. Visualization of the selected J48 classifier threshold curve

As shown in Figure 4.2, the area under the curve happened at ROC threshold =0.996 for both classes. The y-axis shows the true positive rate or the correctly classified instances where each point on the curve represents one model induced by the classifier. The x-axis shows the false positive rate points.

4.2.2. The Naïve Bayes Classifier

The second classification model experiment is conducted using the Naïve Bayes statistical classifier. The reason for choosing this classifier is that it has a good performance for large data size and time taken to build the models is very fast [29]. Also, this algorithm is mostly used to define classes and predict future behavior of existing instances. The WEKA default classification test option, which is 10-fold cross validation and the percentage split with the default distribution of instances, 66% for training and 34% for testing are used.

EXPERIMENT 1.

By setting WEKA default classification test option to 10-fold cross validation, the following result is obtained, the snapshot for this experiment is shown in Annex 9.

Time taken to build model: 0.06 seconds

Correctly Classified Instances	81324	94.5364 %
Incorrectly Classified Instances	4700	5.4636 %

The overall accuracy of the model was measured at 94.5364 %. As illustrated in the confusion matrix below.

```
    a     b   <-- classified as
 42067   945 |     a = GOOD
  3755 39257 |     b = BAD
```

The detailed accuracy by class, having parameters such as Precision, Recall, F-Measure, ROC-Area, and weighted result are given below.

=== Detailed Accuracy by Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.978	0.087	0.918	0.978	0.947	0.893	0.991	0.992	GOOD
	0.913	0.022	0.976	0.913	0.944	0.893	0.991	0.991	BAD
Wd Avg.	0.945	0.055	0.947	0.945	0.945	0.893	0.991	0.991	

EXPERIMENT 2.

By setting WEKA default classification test option to percentage split, the following result is obtained.

Time taken to test model on test split: 0.12 seconds

Correctly Classified Instances	27677	94.6355 %
Incorrectly Classified Instances	1571	5.3645 %

The overall accuracy of the model was measured at 94.6355 %. As illustrated in the confusion matrix below.

```
    a     b   <-- classified as
 14262   307 |     a = GOOD
  1264 13415 |     b = BAD
```

The detailed accuracy by class, having parameters such as Precision, Recall, F-Measure, ROC-Area, and weighted result are given below.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.979	0.086	0.919	0.979	0.948	0.895	0.991	0.991	GOOD
	0.914	0.021	0.978	0.914	0.945	0.895	0.991	0.990	BAD
WdAvg.	0.946	0.053	0.948	0.946	0.946	0.895	0.991	0.991	

The following Table 4.4. summarizes the result of the experiment using this classifier in both classification test options.

Test Option	Time taken	Classified Instances	
		Correctly classified	Incorrectly classified
10-fold cross validation	0.04 seconds	81324 (94.5364%)	4700 (5.4636%)
Percentage split	0.04 seconds	27679 (94.6355%)	1569 (5.3645%)

Table 4.4. Summary of experiments for Naïve Bayes classifier

As shown from the above summary of experiments, both models take equal time to build the models. But percentage split has a better correctly classified instances (94.6355 %) in percentage than the model with the 10-fold cross validation test option. Therefore, the model with the percentage split test option could represent the best Naïve Bayes classifier in this experiment.

Algorithm	Test Option	Accuracy	Time	Recall	Precision	ROC Area	F-Measure	Class
Naïve Bayes	Percentage split	94.6355%	0.04 sec	0.979	0.919	0.993	0.948	GOOD
				0.914	0.978	0.993	0.945	BAD
Weighted Average		94.6355%	0.04Sec	0.946	0.948	0.993	0.946	

Table 4.5. Confusion matrix for the selected Naïve Bayes classifier

As shown in Table 4.5, the Naïve Bayes classifier has the best classification result, where 94.6355% of the instances are correctly classified. This means Naïve Bayes classifier with percentage split test option has a better classification performance in identifying the ‘Bad’ class or defaulters.

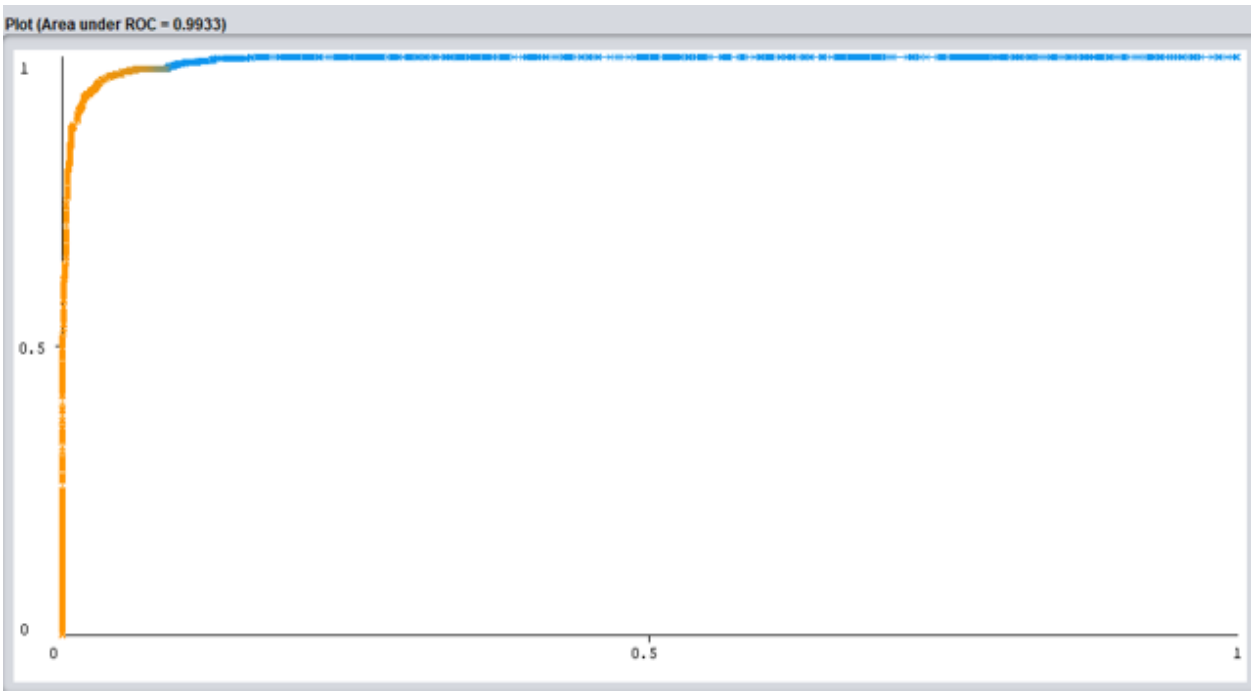


Figure 4.4. Visualization of threshold curve for the selected Naïve Bayes classifier

As it is shown on Figure 4.3, the threshold for selected Naïve Bayes classifier occurred at ROC=993.

4.2.3. Multilayer Perceptron Classifier

Multilayer Perceptron (MLP) classifier is a neural network classifier that classify instances using backpropagation. The reason this algorithm is selected is that it is accurate and efficient classifier [39]. In order to come up with the best MLP model, the WEKA tool has various setups some of them are described in Table 4.6.

Parameter	Description
Learning Rule	Learning Rate for the backpropagation algorithm. (Value should be between 0 - 1, Default = 0.3).
Hidden Layer	The hidden layers to be created for the network. (Value should be a list of comma separated Natural numbers or the letters 'a' = (attribs + classes) / 2, 'i' = attribs, 'o' = classes, 't' = attribs + classes) for wildcard values, Default = a). Comma separated numbers for nodes on each layer

Table 4.6. Description of parameters to be tuned in MLP classification modeling

This experiment is conducted based on two classification test options, 10-fold cross validation and percentage split. The default value of percentage split, which is 66% for training and 34% for testing have been used. An attempt was made to optimize the classification model through parameter setup of the 'Hidden Layer' and 'Learning Rate' that are implemented on this study. Comparison of best model is done in reference with the time taken to build the model and the percentage accuracy of correctly classified instances.

EXPERIMENT 1.

Setting WEKA classification test option to 10-fold cross validation, Learning Rate =0.3 (Default), Hidden Layers=4, Seed=0. The detailed summary snapshot for this experiment is shown in **Annex 10**.

Time taken to build model: 273.47 seconds

Correctly Classified Instances	83008	96.4929 %
Incorrectly Classified Instances	3017	3.5071 %

The overall accuracy of the model was measured at 96.4929 %. As illustrated in the confusion matrix below.

a	b	<-- classified as
41011	2008	a = GOOD
1009	41997	b = BAD

The detailed accuracy by class, having parameters such as Precision, Recall, F-Measure, ROC-Area, and weighted result are given below.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.953	0.023	0.976	0.953	0.965	0.930	0.989	0.986	GOOD
	0.977	0.047	0.954	0.977	0.965	0.930	0.989	0.987	BAD
WdAvg.	0.965	0.035	0.965	0.965	0.965	0.930	0.989	0.987	

EXPERIMENT 2.

Setting WEKA classification test option to Percentage split, Learning Rate =0.3(Default), Hidden Layers=4, Seed=0

Time taken to build model: 213.43 seconds

Correctly Classified Instances 28301 96.7622 %

Incorrectly Classified Instances 947 3.2378 %

The overall accuracy of the model was measured at 96.7622 %. As illustrated in the confusion matrix below.

```
    a     b   <-- classified as
 14025   575 |     a = GOOD
   372 14276 |     b = BAD
```

The detailed accuracy by class, having parameters such as Precision, Recall, F-Measure, ROC-Area, and weighted result are given below.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.961	0.025	0.974	0.961	0.967	0.935	0.983	0.970	GOOD
	0.975	0.039	0.961	0.975	0.968	0.935	0.983	0.985	BAD
WAvg.	0.968	0.032	0.968	0.968	0.968	0.935	0.983	0.978	

The following Table 4.7. summarizes the result of the experiment using MLP classifier in 10-fold cross validation and percentage split classification test options.

Parameter Setup	Test Option	Time taken to build model	Classification	
			Correct	Incorrect
Hidden Layer =4 Learning Rate=0.3	10-fold cross validation	269.9 seconds	96.4929 %	3.2378 %

Seed=0	Percentage split	213.43 seconds	96.7622 %	3.5071 %
Hidden Layer =3 Learning Rate=0.3 Seed=0	10-fold cross validation	216.17 seconds	94.3237 %	5.6763 %
	Percentage split	246.17 seconds	94.3237 %	5.6763 %

Table 4.7. Summary of experiments for Multilayer Perceptron classifier

As shown from Table 4.7, the best MLP classification model in this experiment that have the highest classification accuracy was the one with Hidden Layers = 4, Learning Rate = 0.3 and Seed = 0 using percentage split test option. The selected MLP classifier classified 96.7622 % of the instances correctly into their expected class. The time taken for all MLP classifiers is relatively slower than others

Algorithm	Test Option	Accuracy	Time (sec)	Recall	Precision	F-Measure	ROC Area	Class
Multilayer Perceptron	Percentage split	96.7622 %	213.47	0.961	0.974	0.967	0.983	GOOD
				0.975	0.961	0.968	0.983	BAD
Weighted Average		96.7622 %	213.47	0.968	0.968	0.968	0.983	

Table 4.8. Confusion matrix for the selected Multilayer Perceptron classifier

As shown in Table 4.8, the Multilayer Perceptron classifier has the best classification result, where 96.7622% of the instances are correctly classified. Which means, multilayer perceptron classifier with percentage split test option has a better performance in identifying non-defaulters. As it can be seen from the above result, the time taken to build the classifier using MLP, is the slowest of all for this experiment. This implies it is not as fast as it was claimed on other papers and it might not be suitable for airtime credit risk prediction which requires relatively better response time [47].

4.2.4. Logistic Regression

Logistic regression measures the relationship between a response variable and independent variables, like linear regression, and belongs to the family of exponential classifiers. Logistic regression classifies an observation into one of two classes, and this algorithm analysis can be used when the variables are nominal or binary. The main reason for choosing Logistic Regression for

this study is that it gives insight into variables that are important in prediction. Also, it is a good fit for large sample size [66].

In order to get the best possible result, the default WEKA test options with Percentage split and 10-fold cross validation are used. Then the result is compared to pick the better performing model and analyze it.

EXPERIMENT 1:

By using WEKA's percentage split test option at its default values, the following result is obtained.

The detailed summary for this experiment is shown in **Annex 11**.

Time taken to build model: 5.75 seconds

Correctly Classified Instances 28415 97.1519 %

Incorrectly Classified Instances 833 2.8481 %

The overall accuracy of the model was measured at 97.1519 %. As illustrated in the confusion matrix below.

=== Confusion Matrix ===

a	b	<-- classified as
14214	357	a = GOOD
476	14203	b = BAD

The detailed accuracy by class, having parameters such as Precision, Recall, F-Measure, ROC-Area, and weighted result are given below.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.975	0.032	0.968	0.975	0.972	0.943	0.996	0.995	GOOD
	0.968	0.025	0.975	0.968	0.972	0.943	0.996	0.996	BAD
Wd Avg.	0.972	0.028	0.972	0.972	0.972	0.943	0.996	0.996	

EXPERIMENT 2:

By using WEKA's 10-fold cross validation test option at its default values, the following result is obtained.

Time taken to build model: 6.07 seconds

Correctly Classified Instances 83591 97.1717 %

Incorrectly Classified Instances 2433 2.8283 %

=== Confusion Matrix ===

a b <-- classified as

41978 1034 | a = GOOD

1399 41613 | b = BAD

The detailed accuracy by class, having parameters such as Precision, Recall, F-Measure, ROC-Area, and weighted result are given below.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.976	0.033	0.968	0.976	0.972	0.943	0.996	0.995	GOOD
	0.967	0.024	0.976	0.967	0.972	0.943	0.996	0.996	BAD
Wd Avg.	0.972	0.028	0.972	0.972	0.972	0.943	0.996	0.996	

The following table summarizes the result of the experiment using this classifier in both classification test options.

Test Option	Time Taken (sec)	Classified Instances	
		Correctly classified	Incorrectly classified
10-fold cross validation	6.07	83591 (97.1717%)	2433 (2.8283%)
Percentage split	5.75	28415 (97.1519%)	833 (2.8481%)

Table 4.9. Summary of Experiments for Logistic regression classifier

As shown from Table 4.9, the best logistic regression classification model in this experiment that have the highest classification accuracy is the one with percentage split test option. The selected

logistic regression classifier classifies 97.1717% of the instances correctly into their class and the time taken to build the model is better than the other test option.

Algorithm	Option	Accuracy	Time	Recall	Precision	ROC	F-measure	Class
Logistic	10-fold cross v.	97.1717 %	6.07 sec	0.976	0.968	0.996	0.972	Good
				0.967	0.976	0.996	0.972	Bad
Weighted Average			6.07	0.972	0.972	0.996	0.972	

Table 4.10. Confusion matrix for the selected Logistic classifier

As shown in Table 4.10, Logistic regression with 10-fold cross validation test option has a better performance in classifying non defaulters correctly.

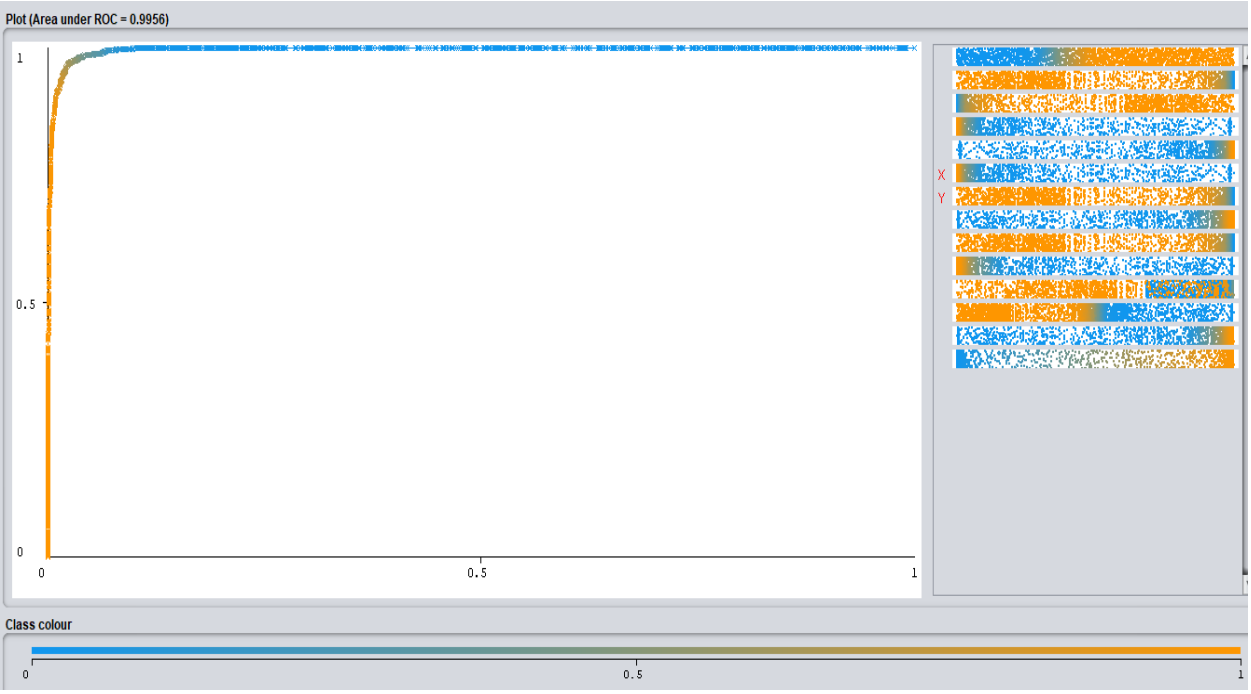


Figure 4.5. Visualization of the ROC Area curve for the selected Logistic classifier.

As shown on Figure 4.5, the threshold points for the ROC curve occurred at a threshold point = 0.996.

4.2.5. Comparison of J48, Naïve Bayes, Multilayer Perceptron and Logistic Regression Models

In order to select a data mining model for classification tasks in the context of this study, it is necessary to evaluate the selected best model from J48 decision tree algorithm, Naïve Bayes, Multilayer Perceptron (MLP) and logistic regression classifiers.

Classifier	Test option	Time (sec)	Accuracy	Precision	Recall	ROC	F-measure
J48	10-fold	7.43	98.5632%	0.986	0.986	0.996	0.986
Naïve Bayes	Percentage split	0.12	94.6355%	0.948	0.946	0.991	0.946
MLP	Percentage split	213.47	96.7622%	0.968	0.968	0.983	0.968
Logistic	10-fold	6.07	97.1717%	0.972	0.972	0.996	0.972

Table 4.11. Performance comparison of the selected models

The best results of the four algorithms are compared with each other by their overall classification accuracy (performance). And, as it is shown in Table 4.11, the overall performance of the decision tree model was 98.5632% with 86024 datasets and 10-fold cross validation test option. The precision recall and F-measure for this classifier is 9.86 while the ROC Area (AUC) is 0.996. The classification accuracy of the Naive Bayes model with the same data size and percentage split test option was 94.6355%. On the other hand, the classification accuracy of Multilayer Perceptron with 10-fold cross validation test option is 96.7622%. The fourth classification algorithm tested is logistic regression, and it has shown overall performance of 97.1717% with percentage split test option.

The J48 decision tree has shown better classification performance with 10-fold cross validation technique. Hence, it is reasonable to conclude that the J48 decision tree model is the best classifier model for implementing airtime credit risk prediction of the company.

By traversing through the selected J48 decision tree classifier we can observe the following important rules (Appendix 8).

If DATA_USAGE = HIGH and CHANNEL = VC and Gender=MALE, then CLASS = BAD (Defaulter)

This rule shows that high data consumption given recharge channel is VC (scratchable voucher card), there is a high prediction power that a subscriber may end up defaulting.

If DATA_USAGE = LOW and VOICE_USAGE = HIGH, then CLASS = GOOD (Non-defaulter).

This rule shows, unlike high data usage, high voice usage is a predictor of a non-defaulting subscribers.

Also, from the above two sample rules, it is possible to say that data usage and voice usage have high prediction power when considering airtime credit loan decision.

If VOICE_USAGE = LOW and NETWORK_AGE = SMALL, then DEFAULTER (BAD)

If VOICE_USAGE = HIGH or DATA_USAGE = LOW and NETWORK_AGE = LONG, then GOOD (NON-DEFAULTER)

From the above two rules we can see that network age is another important parameter in airtime loan decision. LONG network age means subscribers are likely to keep their number and LOW network age can be a predictor of defaulting subscriber given voice usage is LOW.

4.3. Evaluation

On the first experiment, it has been tried to create the best classification model from the J48 decision tree algorithm by testing it on different values of post-pruning and pre-pruning parameters. As a result, the model built on this classification algorithm could accurately classify 98.5632% of the instances into their right class. The knowledge extracted from the J48 decision tree classifier has identified a unique and new pattern that data usage pattern and voice usage pattern can play a great role in predicting defaulters and non-defaulters. Therefore, based on domain expert's evaluation, it can be deduced that when DATA_USAGE is HIGH with channel used for topping up is VC then there is a high risk of defaulting.

The second experiment is conducted on the Naïve Bayes classification algorithm that results the least classification accuracy of all the experimented classifiers, which is 94.6355 %. The time taken to build the model is the smallest of all other algorithms. But its accuracy is far below than the compared classifiers.

The third experiment was done by using Multilayer Perceptron classification algorithm which has an accuracy of 96.6767 %. This classifier is the nearest to J48 and Logistic in terms of accuracy. But it takes longer time to build the model when compared to other classification algorithms in all experiment setups. This algorithm takes 213.47 seconds to build the best MLP classifier, which is unimaginable to use for a very huge data such as that of Ethio Telecom.

On the other experiment conducted by logistic regression, 97.1717% accuracy is achieved. This classifier performed better than Naïve Bayes and MLP classifiers, but its performance is less than J48 Decision tree.

Therefore, J48 decision tree with parameters set to CF=0.75 and MNO=2 is selected as the best fit for this study both in terms of time required to build the model and the accuracy of prediction as well as it fits better than other classifiers into the data set.

4.4. Deployment of the result

The main aim of data mining is to give an insight and increase knowledge gained about a stored data. Deployment is the last step in data mining process which means applying the result of classification. The knowledge gained from data need to be organized and presented in a way that the company can understand and use it for successful airtime credit risk prediction.

In this study the result is deployed at a point of decision making. As the result shows there are important parameters such as data usage, channel used for topping up, network age and voice usage can predict the risk of customer that end up with defaulting. Therefore, the company can consider these important parameters at the time of decision making of granting airtime credit loan by integrating the decision system with airtime credit service provision. Sample implementation using attributes such as data usage, SMS usage, network age, data usage and average recharge amount is developed by java and the resulting output is shown in figure 4.5. The sample implementation Java script is presented in Annex 12.

The rules are deployed in Java by picking important features and ranking them using WEKA tools attribute evaluator ‘ClassifierAttributeEval’ which helps to evaluate the worth of an attribute and assigns weight accordingly. So, based on this the researcher discussed with domain experts to give the final weight for each attribute. After the weight for each attribute is given a threshold is set by using the average value of each evaluated attribute weight. The maximum achieved value according to this scoring is 225. Any score that shows less than 150 are categorized as defaulters. This result was again cross checked with the domain experts and real customers scenario and proved to be correct.

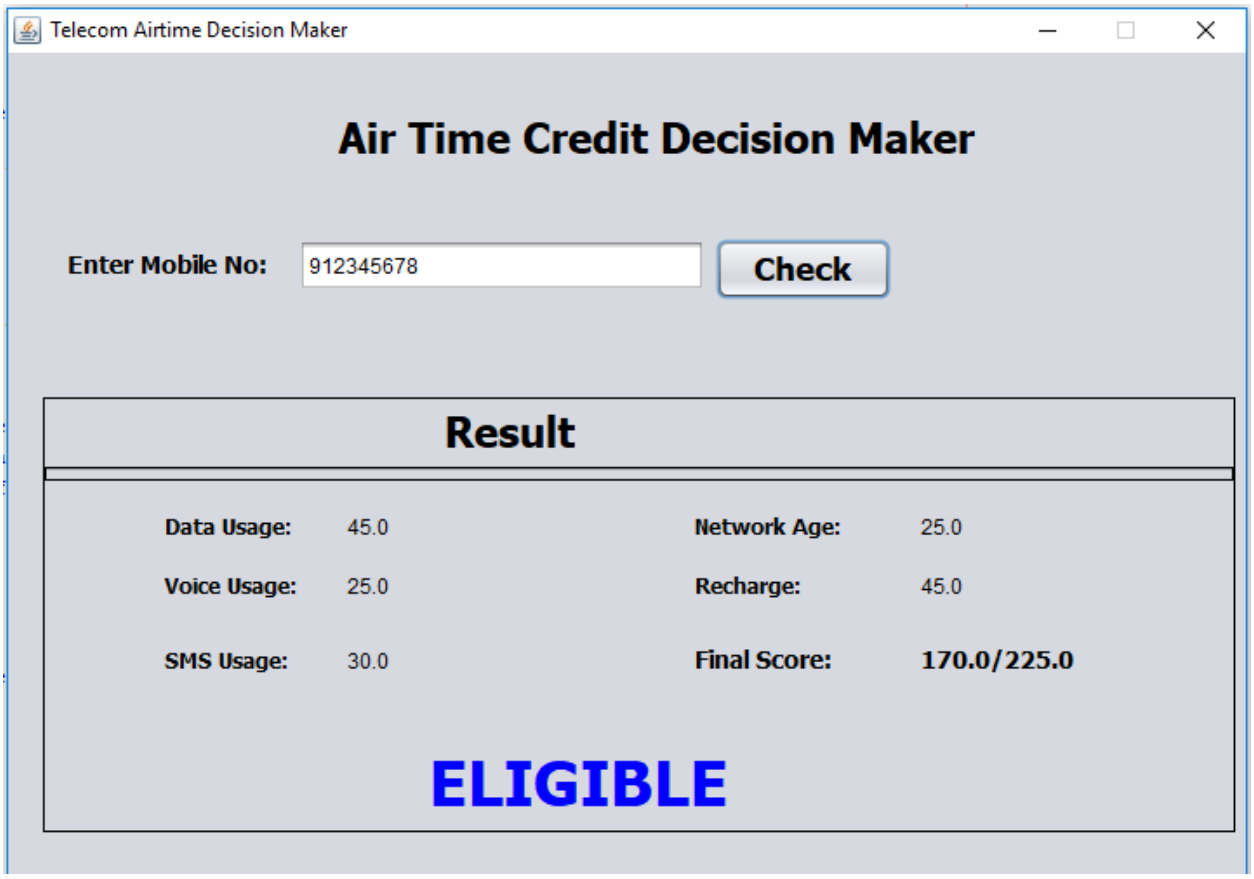


Fig 4.6. A decision made for a subscriber according to the input parameters as ‘ELIGIBLE’

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1. Conclusion

In this chapter a consolidated report on the whole study is presented. It presents the concluding remarks on how the whole process links theory and practice with reference to the data obtained from Ethio Telecom Prepaid Mobile Subscribers. The results that came from applying the DM tools and techniques are then used as reference points to make recommendations that will assist in further studies and in assisting the decision makers in the company to tackle the risk of defaulting airtime credit.

This study aimed at placing into practice the theory on the application of DM tools and techniques on Ethio Telecom Subscribers using airtime credit service. The primary question for the whole study is: How can the application of data mining tools and techniques assist in predicting the risk of airtime credit service defaulting?

The use of both theory and practice assisted in answering this pertinent question. The information gathered from theory confirms that data generation in companies like Ethio Telecom (ET) is increasing at an alarming rate with a huge amount. Such data which is stored in different databases must be used to improve the overall service of the company. But, without proper DM tools and techniques it would be impossible for companies like ET to extract valuable information from such big data. For the companies to get important knowledge out of the data generated, they need to get a better understanding of what benefits would accrue from an application of such DM tools and techniques. Ethio Telecom prepaid mobile subscriber data, the focus of this study, has been studied and put into practice by following the universally accepted standards.

It is from the above primary question and its subsidiary questions that the selected data was subjected to the universally accepted CRISP-DM principles. It is through the application of this methodology that new knowledge was uncovered. The new knowledge was drawn from the identified patterns that different models displayed from the ET prepaid mobile subscribers selected data. What make the methodology popular is its step by step approach and the ability to repeat the generation of models until the accepted results are obtained. As a new field of study DM promises to be an important tool in assisting companies like ET make sense of the huge data that they generate daily. In Telecommunications DM can be applied to different needs of the company such as fraud detection,

customer relationship management, network management, and others. In this study a new application area is explored, which is airtime credit service.

The goal to be achieved in this study was to predict airtime credit loan risk. The criteria to classify customers was based on their usage pattern and their registered attributes during subscription. This usage patterns are data usage, voice usage, recharged amount, topping up channel and SMS usage. The other attributes such as location and gender are taken from the already existing record of customers. Based on the preprocessed data, different classification algorithms were used to build a classifier model. The model built with decision tree J48 algorithm with 10-fold cross validation where its parameters are set to CF=0.75 and MNO=2 has an accuracy of 98.5632%. This model has the best accuracy compared to Naïve Bayes, MLP and Logistic Regression classification algorithms.

As the rules taken by traversing the selected J48 decision tree algorithm from root node to leaf shows, among the studied attributes of a subscribers, data usage pattern could predict the potential defaulting of an airtime credit service user. Channel used for topping up can also be a predictor next to the data usage as shown in the rules generated from selected decision tree classifier.

Effective analysis of prepaid mobile subscriber's data is challenging because of large volume of data collected from every transaction made every second and the data are in various databases which makes it more complex. Therefore, it is essential to apply data mining to extract the relevant information to a level which can be easily managed and used for decision making.

In undertaking of this research, considerable time of the research has been spent on the data preparation phase and consulting the domain expert's in the organization on the interpretation and selection of the appropriate model developed, which were built for classification tasks of data mining.

5.2. Recommendations

It is the researchers' belief that the contribution of this work could be used as a base to further improve and implement it to the airtime credit service. Also, the findings in the study would encourage the company to work on applications of data mining techniques to minimize the risk of defaulting, while protecting its revenue.

Finally, the researcher makes the following recommendations based on findings of the study.

- ✓ This study used different classification algorithms in which J48 performed better than others such as Naïve Bayes, MLP and Logistic Regression. However, other classification algorithms might reveal a better accuracy. Therefore, further research must be conducted using other algorithms.
- ✓ Based on the findings of this study, attributes such as data usage, topping up channel and voice usage can be used at a time of decision making as they had shown strong prediction power which can help to reduce the risk of airtime credit defaulting.
- ✓ The company can use the results of this study as an input and by integrating it with its airtime credit system to improve its service provision by minimizing risk of defaulting.
- ✓ The data preparation stage is mostly time taking, it is recommended for big companies like Ethio Telecom to implement or own a data warehouse where all their data can be formally stored over a longer period and ease their day to day service provisioning by implementing different data mining or machine learning techniques.

5.3. Future Works

In the future the researcher has a plan to further enhance the models built by incorporating different features such as the loan history of the subscriber.

REFERENCES

[1] Zaki, M. J., Meira Jr, W., & Meira, W. "Data mining and analysis: fundamental concepts and algorithms". Cambridge University Press. 2014.

[2] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. "Data Mining: Practical machine learning tools and techniques". Morgan Kaufmann. 2016.

[3] Rangra, K., & Bansal, K. L. "Comparative study of data mining tools". International journal of advanced research in computer science and software engineering, 4(6). 2014.

[4] Yigzaw, M., Hill, S., Banser, A., & Lessa, L., "Using data mining to combat infrastructure inefficiencies: The case of predicting nonpayment for Ethiopian telecom". In 2014 AAAI Spring Symposium Series.

[5] Tesfaye. M., "The Application of Data Mining in Credit Risk Assessment: The Case of United Bank Sc", Mengistu Tesfaye, 2014, AAU Masters Thesis

[6] Solomon, S., & Beshah, T. (2013). "Predicting Customer Loyalty Using Data Mining Techniques". HiLCoE Journal of Computer Science and Technology, 108., 2015.

[7] Shafei, I., & Tabaa, H. "Factors affecting customer loyalty for mobile telecommunication industry". EuroMed Journal of Business, 11(3), 347-361. 2016.

[8] Venable, J., Pries-Heje, J., & Baskerville, R. "FEDS: a framework for evaluation in design science research". European journal of information systems, 25(1), 77-89. 2016.

[9] Ali, S. M., & Tuteja, M. R. Data Mining Techniques. 2014.

[10] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J.," Data Mining: Practical machine learning tools and techniques." Morgan Kaufmann., 2016.

[11] Larose, D. T., & Larose, C. D. "Discovering knowledge in data: an introduction to data mining". John Wiley & Sons. 2014.

[12] Gupta, R. "Journey from data mining to Web Mining to Big Data". arXiv preprint arXiv:1404.4140. 2014.

[13] Rangra, K., & Bansal, K. L. "Comparative study of data mining tools". International journal of advanced research in computer science and software engineering, 4(6). 2014.

[14] Raju, P. S., Bai, D. V. R., & Chaitanya, G. K. "Data mining: techniques for enhancing customer relationship management in banking and retail industries". International Journal of Innovative Research in Computer and Communication Engineering, 2(1), 2650-2657. 2014.

[15] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. "Data Mining: Practical machine learning tools and techniques". Morgan Kaufmann. 2016.

[16] Witten, I. H. and E. Frank. "Data mining: Practical machine learning tools and techniques. San Francisco: Elsevier. 2016.

[17]Zaiane, O. R. "Principles of knowledge discovery in databases: University of Alberta. Pp. 1&3. 2015.

[18] Larose, D. T. “Discovering knowledge in data: An introduction to data mining” New Jersey: John Wiley & Sons, Inc. 2014.

[19] Aggarwal, C. C. “Data mining: the textbook”. Springer. 2015.

[20] Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. “From data mining to knowledge discovery in database”. AI magazine, American Association for Artificial Intelligence, Pp. 37-44. 2014.

[21] Qin, S. J. “Process data analytics in the era of big data”. AIChE Journal, 60(9), 3092-3100. 2014.

[22] Tan, P. N. “Introduction to data mining”. Pearson Education India. 2018.

[23] Larose, D. T., & Larose, C. D. “Discovering knowledge in data: an introduction to data mining”. John Wiley & Sons. 2014.

[24] Shafique, U., & Qaiser, H. “A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)”. International Journal of Innovation and Scientific Research, 12(1), 217-222. 2014.

[25] Piatetsky, G. CRISP-DM, still the top methodology for analytics, data mining, or data science projects. KDD News. 2014.

[26] Szczerba, M., & Ciemski, A. “Credit Risk Handling in Telecommunication Sector”. In Industrial Conference on Data Mining (pp. 117-130). Springer, Berlin, Heidelberg. 2016.

[27] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. “CRISP-DM 1.0 Step-by-step data mining guide. Crisp DM Consortium (Updated 2010) (1999). 2016.

[28] Azevedo, A. I. R. L., & Santos, M. F. “KDD, SEMMA and CRISP-DM: a parallel overview”. IADS-DM. 2015.

[29] Björkegren, D., & Grissen, D. “Behavior revealed in mobile phone usage predicts loan repayment”. Available at SSRN 2611775. 2018.

[30] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. “Data Mining: Practical machine learning tools and techniques”. Morgan Kaufmann. 2016.

[31] Williams, G. “Data mining with rattle and R: The art of excavating data for knowledge discovery”. Springer (2017).

[32] Weiss, G. “Data Mining in the Telecommunications Industry”. USA: Global Fordham University. (2014).

[33] Breiman, L. “Classification and regression trees”. Routledge. 2017.

[34] Gandini, G., Bosetti, L., & Almici, A. “Risk management and sustainable development of telecommunications companies”. SYMPHONYA Emerging Issues in Management, (2). 2014.

[35] Rokach, L. “Decision forest: Twenty years of research”. Information Fusion, 27, 111-125.2016.

[36] San Pedro, J., Proserpio, D., & Oliver, N. “MobiScore: towards universal credit scoring from mobile phone data”. In International Conference on User Modeling, Adaptation, and Personalization (pp. 195-207). Springer, Cham. 2015.

[37] Krawczyk, B., Woźniak, M., & Schaefer, G. “Cost-sensitive decision tree ensembles for effective imbalanced classification”. Applied Soft Computing, 14, 554-562. 2014.

[38] Kimball, R., Ross, M., Thornthwaite, W., Mundy, J. & Becker, B. “The Data Warehouse Lifecycle Toolkit”. 2nd ed. Indianapolis: John Wiley & Sons. 2016.

[39] L. Wang, "Data Mining, Machine Learning and Big Data Analytics", International Transaction of Electrical and Computer Engineers System, vol. 4, no. 2, pp. 55-61, 2017.

[40 Goeschel, K. Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive Bayes for off-line analysis. In Southeast Con 2016 (pp. 1-6). IEEE. (2016).

[41] Zhao, Y., & Zhang, Y. “Comparison of decision tree methods for finding active objects”. Advances in Space Research, 41(12), 1955-1959. (2015).

[42] Bielza, C., & Larrañaga, P. “Discrete Bayesian network classifiers: a survey”. ACM Computing Surveys (CSUR), 47(1), 5. 2014.

[43] Li, T., Li, J., Liu, Z., Li, P., & Jia, C. “Differentially private Naive Bayes learning over multiple data sources”. Information Sciences, 444, 89-104. 2018.

[44] Van Der Aalst, W. “Data science in action. In Process Mining” (pp. 3-23). Springer, Berlin, Heidelberg. 2016.

[45] Zaki, M. & Meira, J. W. “Data Mining and Analysis: Fundamental Concepts and Algorithms”. Cambridge University Press. (2014).

[46] Li, R., Zhang, W., Suk, H. I., Wang, L., Li, J., Shen, D., & Ji, S. Deep learning-based imaging data completion for improved brain disease diagnosis. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 305-312). Springer, Cham. 2014.

[47] Sorokosz, L. & Zieniutycz, W. “Artificial Neural Networks in Microwave Components and Circuits Modeling”. Przeglad Elektrotechniczny (Electrical Review). (2014).

[48] Gurney, K. (2014). “An introduction to neural networks”. CRC press. 2014.

[49] Haykin, S. S., Haykin, S. S., Haykin, S. S., Elektroingenieur, K., & Haykin, S. S. “Neural networks and learning machines” (Vol. 3). Upper Saddle River: Pearson education. (2014).

[50] S. Chen, Y.-J. J. Goo, and Z.-D. Shen, “A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements,” The Scientific World Journal, vol. 2014, Article ID 968712, 9 pages, 2014.

[51] Y. Kumar and G. Sahoo, “Analysis of parametric and non-parametric classifiers for classification technique using WEKA,” International Journal of Information Technology and Computer Science, vol. 4, pp. 43–49, 2015.

[52] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. “Data Mining: Practical machine learning tools and techniques”. Morgan Kaufmann. 2016.

[53] Hossin, M., & Sulaiman, M. N. “A review on evaluation metrics for data classification evaluations”. International Journal of Data Mining & Knowledge Management Process, 5(2), 1. 2015.

[54] Melaku, G. Applicability of Data Mining Techniques to Customer Relationship Management (CRM): The Case of Ethiopian Telecommunication Corporation's (ETC) Code Division Multiple Access (CDMA) Telephone Service. (Master’s thesis, Addis Ababa University). 2014.

[55] Henock, W. The application of data mining to support customer relationship Management at Ethiopian Airlines. Addis Ababa University, Unpublished Master's Thesis. 2014.

[56] Denekew, A. Application of data mining to support Customer Relationship Management (CRM) at Ethiopian Airline. Addis Ababa University, Unpublished master's thesis. 2014.

[57] Kumneger, F. Application of data mining techniques to support CRM in Ethiopian Shipping Lines: Addis Ababa University, Unpublished Master's Thesis. 2014.

[58] Solomon, S., & Beshah, T. Predicting Customer Loyalty Using Data Mining Techniques. HiLCoE Journal of Computer Science and Technology, 108. (2014).

[59] Yap, B. W., Ong, S. H., & Husain, N. H. M. "Using data mining to improve assessment of credit worthiness via credit scoring models". Expert Systems with Applications, 38(10), 13274-13283. (2015).

[60] Ge, Z., Song, Z., Ding, S. X., & Huang, B. "Data mining and analytics in the process industry: The role of machine learning". IEEE Access, 5, 20590-20616. 2017

[61] Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., & Wasinger, R. "Investigation and improvement of multi-layer perceptron neural networks for credit scoring". Expert Systems with Applications, 42(7), 3508-3516. 2015.

[62] Brânduşoiu, I., Toderean, G., & Beleiu, H. (2016, June). "Methods for churn prediction in the pre-paid mobile telecommunications industry". In 2016 International conference on communications (COMM) (pp. 97-100). IEEE. 2016.

[63] Weiss, G. M. "Data mining in telecommunications". In Data Mining and Knowledge Discovery Handbook (pp. 1189-1201). Springer, Boston, MA. (2015).

[64] García, S., Luengo, J., & Herrera, F. "Data preprocessing in data mining" (pp. 59-139). New York: Springer. 2015.

[65] Jovic, A., Brkic, K., & Bogunovic, N. "An overview of free software tools for general data mining". In 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 1112-1117). IEEE. 2014.

[66] Nataša Š., Ana B., "Logistic regression modelling: procedures and pitfalls in developing and interpreting prediction models", 631–652, 2017

APPENDICES

Annex 1: J48 Decision Tree Experiment 1

Experiment 1: J48 Decision Tree Experiment with WEKA parameters set to: 10-fold cross validation test option with CF =0.25 and MNO=2.

```
=== Summary ===

Correctly Classified Instances       84510               98.24    %
Incorrectly Classified Instances      1514                1.76    %
Kappa statistic                          0.9648
Mean absolute error                      0.03
Root mean squared error                  0.1246
Relative absolute error                  6.0046 %
Root relative squared error             24.9279 %
Total Number of Instances            86024

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.983    0.018    0.982      0.983    0.982      0.965    0.991     0.989     GOOD
                 0.982    0.017    0.983      0.982    0.982      0.965    0.991     0.989     BAD
Weighted Avg.    0.982    0.018    0.982      0.982    0.982      0.965    0.991     0.989

=== Confusion Matrix ===

     a     b   <-- classified as
 42287   725 |     a = GOOD
   789 42223 |     b = BAD
```

Annex 2: J48 Decision Tree Experiment 2

Experiment 2:

```
=== Summary ===

Correctly Classified Instances       84464               98.1866 %
Incorrectly Classified Instances      1560                1.8134 %
Kappa statistic                          0.9637
Mean absolute error                      0.031
Root mean squared error                  0.1262
Relative absolute error                  6.192  %
Root relative squared error             25.2331 %
Total Number of Instances            86024

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.982    0.019    0.981      0.982    0.982      0.964    0.990     0.989     GOOD
                 0.981    0.018    0.982      0.981    0.982      0.964    0.990     0.987     BAD
Weighted Avg.    0.982    0.018    0.982      0.982    0.982      0.964    0.990     0.988

=== Confusion Matrix ===

     a     b   <-- classified as
 42256   756 |     a = GOOD
   804 42208 |     b = BAD
```

- Snapshot for Output of experiment 2 with J48 Decision Tree with 10-fold cross validation test option where its parameters are set as CF=0.25 and M=5.

Annex 3**: J48 Decision Tree Experiment 3**

Experiment 3:

```
=== Summary ===

Correctly Classified Instances       84622               98.3702 %
Incorrectly Classified Instances      1402                1.6298 %
Kappa statistic                          0.9674
Mean absolute error                      0.0273
Root mean squared error                  0.119
Relative absolute error                  5.4605 %
Root relative squared error             23.8032 %
Total Number of Instances            86024

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.986    0.018    0.982      0.986    0.984      0.967    0.995     0.993     GOOD
                 0.982    0.014    0.986      0.982    0.984      0.967    0.995     0.993     BAD
Weighted Avg.    0.984    0.016    0.984      0.984    0.984      0.967    0.995     0.993

=== Confusion Matrix ===

     a     b   <-- classified as
 42397   615 |     a = GOOD
   787 42225 |     b = BAD
```

Snapshot for Output of experiment 3 with J48 Decision Tree with 10-fold cross validation test option where its parameters are set as CF=0.5 and M=2.

Annex 4**: J48 Decision Tree Experiment 4**

Experiment 4:

```
=== Summary ===

Correctly Classified Instances       84788               98.5632 %
Incorrectly Classified Instances      1236                1.4368 %
Kappa statistic                          0.9713
Mean absolute error                      0.0222
Root mean squared error                  0.109
Relative absolute error                  4.4492 %
Root relative squared error             21.8009 %
Total Number of Instances            86024

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.990    0.018    0.982      0.990    0.986      0.971    0.996     0.995     GOOD
                 0.982    0.010    0.989      0.982    0.986      0.971    0.996     0.996     BAD
Weighted Avg.    0.986    0.014    0.986      0.986    0.986      0.971    0.996     0.995

=== Confusion Matrix ===

     a     b   <-- classified as
 42563   449 |     a = GOOD
   787 42225 |     b = BAD
```

Snapshot for Output of experiment 4 with J48 Decision Tree with 10-fold cross validation test option where its parameters are set as CF=0.75 and M=2.

Annex 5**: J48 Decision Tree Experiment 5**

Experiment 5:

```
=== Summary ===

Correctly Classified Instances       28708               98.1537 %
Incorrectly Classified Instances       540                1.8463 %
Kappa statistic                          0.9631
Mean absolute error                      0.0313
Root mean squared error                  0.1284
Relative absolute error                  6.2504 %
Root relative squared error             25.6808 %
Total Number of Instances            29248

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.983    0.020    0.980      0.983    0.981      0.963    0.990     0.985     GOOD
                 0.980    0.017    0.983      0.980    0.982      0.963    0.990     0.986     BAD
Weighted Avg.    0.982    0.018    0.982      0.982    0.982      0.963    0.990     0.986

=== Confusion Matrix ===

     a     b   <-- classified as
 14321   248 |     a = GOOD
   292 14387 |     b = BAD
```

Snapshot for Output of experiment 5 with J48 Decision Tree with percentage split test option where its parameters are set as CF=0.25 and M=2.

Annex 6**: J48 Decision Tree Experiment 6**

Experiment 6:

```
=== Summary ===

Correctly Classified Instances       28882               98.7486 %
Incorrectly Classified Instances       366                1.2514 %
Kappa statistic                          0.975
Mean absolute error                      0.0199
Root mean squared error                  0.1034
Relative absolute error                  3.9811 %
Root relative squared error             20.6799 %
Total Number of Instances            29248

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.992    0.017    0.983      0.992    0.987      0.975    0.997     0.995     GOOD
                 0.983    0.008    0.992      0.983    0.987      0.975    0.997     0.996     BAD
Weighted Avg.    0.987    0.012    0.988      0.987    0.987      0.975    0.997     0.995

=== Confusion Matrix ===

     a     b   <-- classified as
 14447   122 |     a = GOOD
   244 14435 |     b = BAD
```

Snapshot for Output of experiment 6 with J48 Decision Tree with percentage split test option where its parameters are set as CF=0.5 and M=2.

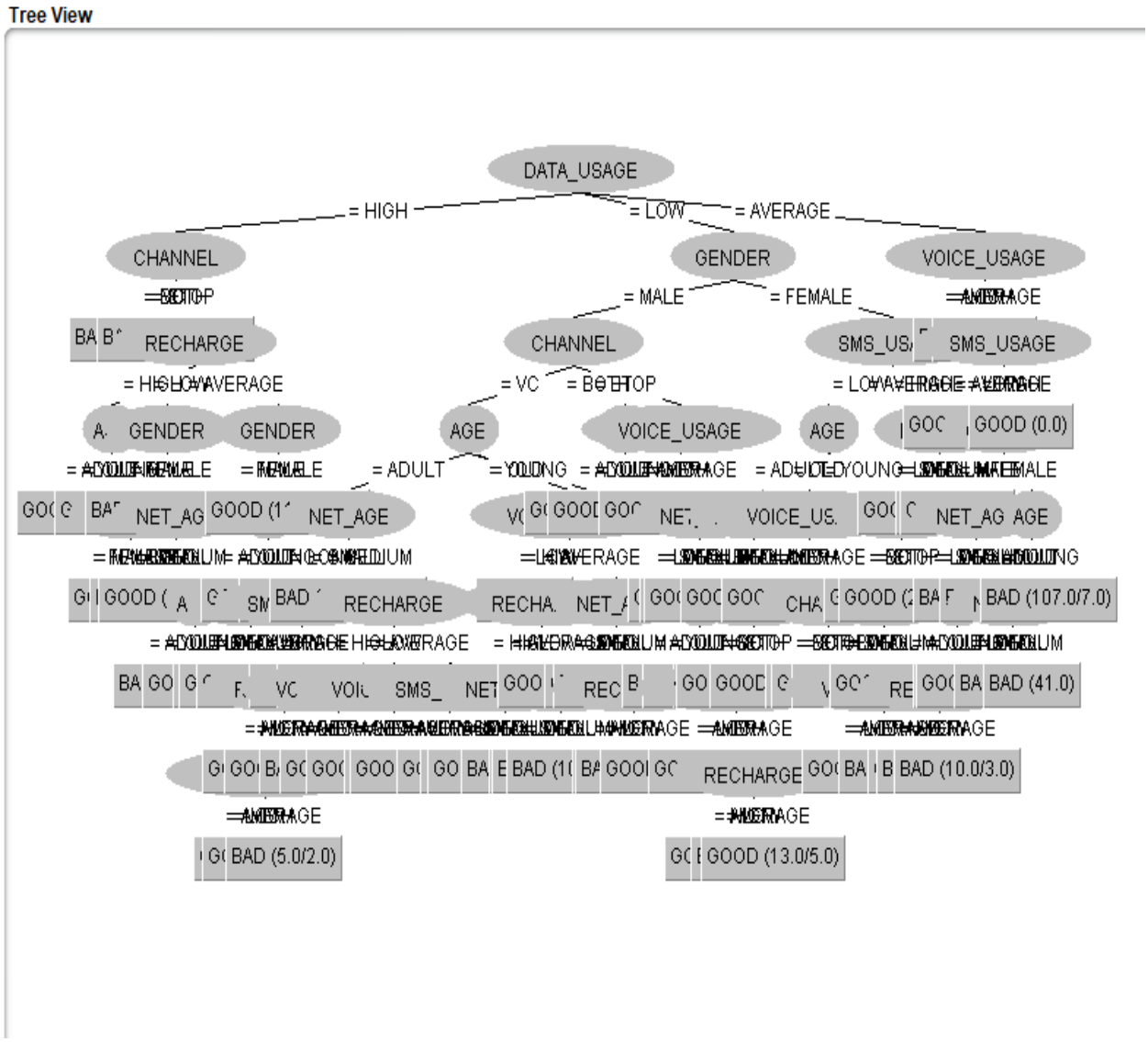
Annex 7: J48 Decision Tree Experiment 7

Experiment 7:

```
=== Summary ===

Correctly Classified Instances       28722               98.2016 %
Incorrectly Classified Instances       526                1.7984 %
Kappa statistic                          0.964
Mean absolute error                      0.0285
Root mean squared error                  0.1241
Relative absolute error                  5.705  %
Root relative squared error             24.8264 %
Total Number of Instances            29248

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.984    0.020    0.980      0.984    0.982      0.964    0.994     0.992     GOOD
                 0.980    0.016    0.984      0.980    0.982      0.964    0.994     0.993     BAD
Weighted Avg.    0.982    0.018    0.982      0.982    0.982      0.964    0.994     0.992

=== Confusion Matrix ===

     a     b   <-- classified as
 14338   231 |     a = GOOD
   295 14384 |     b = BAD
```

Snapshot for Output of experiment 7 with J48 Decision Tree with percentage split test option where its parameters are set as CF=0.5 and M=5.

Annex 8: Tree view and sample rules generated for the selected J48 classifier.



Generated Tree for the selected J48 classifier

Sample rules generated for selected decision tree algorithm.

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

DATA_USAGE = HIGH
|   CHANNEL = VC
|   |   GENDER = MALE: BAD (24631.0/8.0)
|   |   GENDER = FEMALE
|   |   |   AGE = ADULT
|   |   |   |   NET_AGE = LONG
|   |   |   |   |   LOCATION = SWAAZ: BAD (143.0)
|   |   |   |   |   LOCATION = NAAZ
|   |   |   |   |   |   RECHARGE = HIGH
|   |   |   |   |   |   |   SERVICE = WCDMA 3G: BAD (101.0/28.0)
|   |   |   |   |   |   |   SERVICE = 4G LTE: GOOD (2.0)
|   |   |   |   |   |   RECHARGE = LOW: BAD (121.0/8.0)
|   |   |   |   |   |   RECHARGE = AVERAGE: BAD (100.0/3.0)
|   |   |   |   |   LOCATION = CAAZ: BAD (272.0/3.0)
|   |   |   |   |   LOCATION = WAAZ
|   |   |   |   |   |   RECHARGE = HIGH: GOOD (41.0/13.0)
|   |   |   |   |   |   RECHARGE = LOW: BAD (37.0/12.0)
|   |   |   |   |   |   RECHARGE = AVERAGE: BAD (35.0/2.0)
|   |   |   |   |   LOCATION = NWR: BAD (422.0/8.0)
|   |   |   |   |   LOCATION = SSWR: BAD (51.0/2.0)
|   |   |   |   |   LOCATION = SER: BAD (130.0/7.0)
|   |   |   |   |   LOCATION = SR: BAD (591.0/3.0)
|   |   |   |   |   LOCATION = ER: BAD (76.0/2.0)
|   |   |   |   |   LOCATION = NER: BAD (140.0/4.0)
|   |   |   |   |   LOCATION = JIGJIGA: BAD (5.0)

DATA_USAGE = LOW
|   CHANNEL = VC
|   |   VOICE_USAGE = AVERAGE: GOOD (13371.0)
```

By traversing the generated tree from root to leaf, it is possible to derive important rules that facilitate business decision-making. Some sample rules generated from the decision tree are described as follows:

Rule 1: IF DATA_USAGE=HIGH and CHANNEL=VC, then DEAULTER (BAD)

RULE 2: IF DATA_USAGE=AVERAGE and CHANNEL=VC and AGE=ADULT and LOCATION=SWAAZ and GENDER=MALE, the DEFAULTER (BAD)

Rule 3: IF DATA_USAGE=LOW, CHANNEL_USED=VC and LOCATION=WAAZ, then NON-DEFAULTER (GOOD)

Rule 4. IF DATA_USAGE=LOW and CHANNEL_USED=BOTH and GENDER=MALE and Age=ADULT and Recharge=HIGH, then NON-DEFAULTER (GOOD)

Rule 5. IF DATA_USAGE=LOW, CHANNEL=ETOP, RECHARGE=HIGH, then NON-DEFAULTER (GOOD)

Rule 6. IF DATA_USAGE=LOW and CHANNEL=ETOP and RECHARGE=LOW and LOCATION=SWAAZ and NET_AGE=SMALL and AGE=OLD, then BAD

ANNEX 9: **Naïve Bayes Experiment Outputs**

Experiment 1 – 10-fold cross validation

```
=== Summary ===

Correctly Classified Instances       81324               94.5364 %
Incorrectly Classified Instances      4700                5.4636 %
Kappa statistic                          0.8907
Mean absolute error                      0.0532
Root mean squared error                  0.2135
Relative absolute error                 10.6499 %
Root relative squared error             42.7087 %
Total Number of Instances            86024

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.978     0.087     0.918      0.978     0.947      0.893     0.991     0.992     GOOD
               0.913     0.022     0.976      0.913     0.944      0.893     0.991     0.991     BAD
Weighted Avg.  0.945     0.055     0.947      0.945     0.945      0.893     0.991     0.991

=== Confusion Matrix ===

     a     b   <-- classified as
 42067   945 |     a = GOOD
  3755 39257 |     b = BAD
```

Output of Naïve Bayes with 10-fold cross validation test option

Experiment 2: Percentage split

Output of Naïve Bayes experiment with Percentage split test option

```
=== Summary ===

Correctly Classified Instances       27677               94.6287 %
Incorrectly Classified Instances      1571                5.3713 %
Kappa statistic                          0.8926
Mean absolute error                      0.0525
Root mean squared error                  0.2117
Relative absolute error                 10.4928 %
Root relative squared error             42.3399 %
Total Number of Instances            29248

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.979    0.086    0.919      0.979    0.948      0.895    0.991     0.991     GOOD
                 0.914    0.021    0.978      0.914    0.945      0.895    0.991     0.990     BAD
Weighted Avg.    0.946    0.053    0.948      0.946    0.946      0.895    0.991     0.991

=== Confusion Matrix ===

     a     b   <-- classified as
 14262   307 |     a = GOOD
  1264 13415 |     b = BAD
```

Annex 10. Summary of MLP experiment

Experiment 1: Using percentage split validation test option

```
=== Summary ===

Correctly Classified Instances       28301               96.7622 %
Incorrectly Classified Instances       947                3.2378 %
Kappa statistic                          0.9352
Mean absolute error                      0.0482
Root mean squared error                  0.1655
Relative absolute error                  9.6453 %
Root relative squared error             33.0908 %
Total Number of Instances            29248

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.961    0.025    0.974      0.961    0.967      0.935    0.983     0.970     GOOD
                 0.975    0.039    0.961      0.975    0.968      0.935    0.983     0.985     BAD
Weighted Avg.    0.968    0.032    0.968      0.968    0.968      0.935    0.983     0.978

=== Confusion Matrix ===

     a     b   <-- classified as
 14025   575 |     a = GOOD
   372 14276 |     b = BAD
```

Output of MLP experiment with Percentage split test option Hidden layers=4

```
=== Summary ===

Correctly Classified Instances       83008               96.4929 %
Incorrectly Classified Instances      3017                3.5071 %
Kappa statistic                          0.9299
Mean absolute error                      0.0501
Root mean squared error                  0.1676
Relative absolute error                 10.0256 %
Root relative squared error             33.5228 %
Total Number of Instances            86025

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.953    0.023    0.976      0.953    0.965      0.930    0.989     0.986     GOOD
                 0.977    0.047    0.954      0.977    0.965      0.930    0.989     0.987     BAD
Weighted Avg.    0.965    0.035    0.965      0.965    0.965      0.930    0.989     0.987

=== Confusion Matrix ===

     a     b   <-- classified as
 41011  2008 |     a = GOOD
  1009 41997 |     b = BAD
```

Output of MLP experiment with 10-fold cross validation test option with H=4

Annex 11: logistic Regression

Experiment 1: Percentage split test option

```
Correctly Classified Instances       28415               97.1519 %
Incorrectly Classified Instances       833                2.8481 %
Kappa statistic                          0.943
Mean absolute error                      0.0432
Root mean squared error                  0.1482
Relative absolute error                  8.6322 %
Root relative squared error             29.6453 %
Total Number of Instances            29248

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.975    0.032    0.968      0.975    0.972      0.943    0.996     0.995     GOOD
                 0.968    0.025    0.975      0.968    0.972      0.943    0.996     0.996     BAD
Weighted Avg.    0.972    0.028    0.972      0.972    0.972      0.943    0.996     0.996

=== Confusion Matrix ===

     a     b   <-- classified as
 14212   357 |     a = GOOD
   476 14203 |     b = BAD
```

Output of logistic experiment with Percentage split test option

Experiment 2: 10-fold cross validation test option

```
=== Summary ===

Correctly Classified Instances       83591               97.1717 %
Incorrectly Classified Instances      2433                2.8283 %
Kappa statistic                          0.9434
Mean absolute error                      0.0433
Root mean squared error                  0.147
Relative absolute error                  8.6593 %
Root relative squared error             29.4007 %
Total Number of Instances            86024

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.976    0.033    0.968      0.976    0.972      0.943    0.996     0.995     GOOD
                 0.967    0.024    0.976      0.967    0.972      0.943    0.996     0.996     BAD
Weighted Avg.    0.972    0.028    0.972      0.972    0.972      0.943    0.996     0.996

=== Confusion Matrix ===

     a     b   <-- classified as
 41978  1034 |     a = GOOD
  1399 41613 |     b = BAD
```

Output of logistic experiment with 10-fold cross validation test option

Annex 12: Java Implementation code

……

```
for(int i=1;i<=5;i++)
    {
    switch(i)
    {
    case 1: //CASE DATA_USAGE
    if(dataUsage<=0.25)
    {
      DATA_SCORE=LOW_VALUE;


    }
```

```
 else if(dataUsage>0.25 && dataUsage <= 0.55)
{
  DATA_SCORE=AVG_VALUE;

}
    else
{
  DATA_SCORE=HIGH_VALUE;

}
break;
 case 2: //CASE_VOICE
if(voiceUsage<=0.4)
{
  VOICE_SCORE=LOW_VALUE;
   }

 else if(voiceUsage>0.4 && voiceUsage <= 0.7)
{
  VOICE_SCORE=AVG_VALUE;

}
    else
{
  VOICE_SCORE=HIGH_VALUE;

}
break;
 case 3: //CASE SMS_USAGE
```

```
 if(smsUsage<=0.04)
{
  SMS_SCORE=LOW_VALUE;

}

  else if(smsUsage>0.04 && smsUsage <=0.1)
{
  SMS_SCORE=AVG_VALUE;

}
    else
{
  SMS_SCORE=HIGH_VALUE;

}
break;
 case 4://CASE NET_AGE
if(NETAGE<=2)
{
  NETAGE_SCORE=LOW_VALUE;

}

  else if(NETAGE>2 && NETAGE <=5)
{
  NETAGE_SCORE=AVG_VALUE;

}
    else
```

```
{
  NETAGE_SCORE=HIGH_VALUE;

}
break;
 case 5:
 if(RECHRAGE<=50)
{
  RECHRAGE_SCORE=LOW_VALUE;

}

 else if(RECHRAGE>50 && RECHRAGE < 100)
{
  RECHRAGE_SCORE=AVG_VALUE;

}
 else
{
 RECHRAGE_SCORE=HIGH_VALUE;

}
break;

default:
    FINAL_SCORE = 0.0;
    break;
…..
```