Addis Ababa University
College of Natural Sciences
School of Information Science

Applying data mining techniques for predicting telecommunication service faults: the case of Ethio-telecom

Zelalem Worku

October 2013

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

Applying data mining techniques for predicting telecommunication service faults: the case of Ethio-telecom

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

By
Zelalem Worku
October, 2013

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

Applying data mining techniques for predicting telecommunication service faults: the case of Ethio-telecom

By

Zelalem Worku

October 2013

Name and signature of Member of the Examining Board

Name	Title	Signature Date
Zelalem Regassa	Chairperson	
Dereje Tefera (PhD)	Advisor	
Million Meshesha	Examiner	

Declaration

I declare that the thesis is my original work and has not been presented for a degree in any other university.

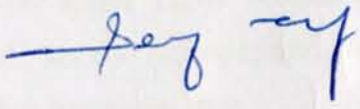
Zelalem Worku

Date October, 2013

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

Dereje Teferi (Phd)

Advisor



Acknowledgements

First and for most, I would like to thank the almighty God for letting me have the chance for this educational program and help me to finish it successfully.

I am deeply grateful to my advisor, Dr. Dereje Teferi for guiding me throughout the whole journey of this study.

I wish to express my gratitude to Dr. Million Meshesha for encouraging me to work on this research area and also for giving me his supportive ideas while preparing my research proposal.

In addition, I would like to thank all the domain experts, who have shared the tedious process of clarifying ideas and results presented in this thesis.

This research has been carried out on a data obtained from ethio - telecom, therefore I should thank the company for its cooperation and willingness.

The joy that my new born little girl Beamilak Samuel brought has been the most important facilitator for successfully finishing my study. I love you so much my little princess! You are my happiness!

Last but not least, I wish to thank all my family especially my father Worku Mamo, my mother Terefe Agonafir, my elder sister Tigist Worku and my younger brother Dr. Anteneh Worku who lit the spark, set the example and who encouraged and supported me all the way with this dream that has come true.

October, 2013

Zelalem Worku

List of Acronyms

ADSL: asynchronies digital subscriber line
BB-ADSL: Broadband-ADSL
CDMA: Code Division Multiple Access
CN: Core Network
CNR: Core Network Resource
CPE: Customer Premises Equipment
CS: Customer Service
EAAZ: East Addis Ababa Zone
EPON: Ethernet Passive Optical Network
ER & Jijiga: East Region & Jijiga
FAN: Fixed Access Network
GSM: Global Service for Mobile communication
ISDN: Integrated Service Digital Network
NAAZ: North Addis Ababa Zone
NER & Semera: North East Region
NER: North East Region
NNOC: National Network Operation Center
NR: North Region
NWR: North West Region
O&M: Operation & Maintenance
P&E: Power & Environment
PSTN: Public Switch Telephone Network
SAAZ: South Addis Ababa Zone
SER: South East Region
SR: South Region
SWAAZ: South West Addis Ababa Zone
SWR: South West Region
TMC: Technology Management Center
TT: trouble ticket
VDSL: Very-high-bit-rate DSL
WAAZ: West Addis Ababa Zone
WR: West Region

Table of Contents

List of tables:

List of Figures

ABSTRACT

Faults are inevitable in telecommunication services, therefore predicting them ahead of time is crucial to make the systems more robust and the operation more reliable. Faults in telecommunication services have direct impact on its availability and maintenance costs, so their quick elimination, prevention and removal of causes that generated them, is of special interest.

This study is aimed at applying data mining techniques to support prediction of broad band network service faults at Ethio-telecom. The subject of this study is from Ethio-telecom's Z-Smart Trouble Ticket system, which contains customer's service fault report information and remarks given by experts about the actual fault reasons after the problems solved.

In the data mining process, the first step was collecting the target data from the above mentioned system at Ethio-telecom. Then various types of preprocessing tasks were performed on the collected data so that to make the data ready for the planned data mining tasks. On the model building phase, C4.5 variant of decision tree and Naïve bayes of Bayesian network algorithms were applied for building the classifiers and accuracy results obtained using J48 and Naïve bayes was 74.06% and 69% respectively. Due to the data set imbalance observed on the class variables, SMOTE minority over sampling technique with J48 algorithm was applied and it improves the classifier accuracy to 77.90%.

The results from this study were encouraging, which strengthened the belief that applying data mining techniques could in fact support network service faults prediction activity at Ethio telecom. In the future, using a balanced data set, and incorporating more attributes and also by testing with various classification algorithms better classifier accuracy could be obtained.

Key Words: Data Mining, Classification, Prediction, Telecommunication, Network, Faults.

CHAPTER ONE

INTRODUCTION

1.1. Background

Currently most companies are collecting and storing data in large databases, and most of the owners of these data recognized the potential value of these data as a useful information source for business decisions. Today knowledge discovery and data mining are answering the dramatically increasing demand for better decision support (Jackson, 2002).

The telecommunications industry is one of the industries that generates and stores a tremendous amount of data. These data include: -

I. **Call detail data**, which describes the calls that traverse the telecommunication networks,
II. **Network data**, which describes the state of the hardware and software components of the network devices, and
III. **Customer data**, which describes the telecommunication customers.

The amount of data is so great that manual analysis of the data is difficult, if not impossible.

The need to handle such large volumes of data led to the development of knowledge-based expert systems. These automated systems performed important functions such as identifying fraudulent phone calls and identifying network faults. The problem with this approach is that it is time consuming to obtain the knowledge from human experts and, in many cases; the experts do not have the required knowledge. The advent of data mining technology promised solutions to these problems and for this reason the telecommunications industry was an early adopter of data mining technology (Weiss, 2005).

Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

In general, data mining tasks can be classified into two categories description & prediction(Siraj & Abdoulha, 2007):

- **Description**: finding human-interpretable patterns, associations or correlations describing the data.
- **Prediction**: constructing one or more sets of data models performing inference on the available set of data, and attempting to predict the behavior of new data sets.

Telecommunications network operation is a promising target for data mining applications because the network operation business consists of several activities, like network operation and management, customer care and billing, marketing, business management and so on (H¨at¨onen, 2009). Therefore, the data collected from all such kinds of operations is clearly a huge amount of data on which data mining can be applied.

This research applied data mining techniques on a data obtained from Z-Smart trouble ticket information system at ethio-telecom. The target data was a data related with broadband network service fault report information which was registered and stored by ethio-telecom for the purpose of following the progress of fault handling process. Based on the information gained from the domain experts, as of July 2012 the number of broadband network service subscribers at ethio-telecom was 30,372.

Broadband network service fault has given special consideration because the population demand for the broadband services has not stopped rising in the last couple of years.

Not only that the number of users is getting higher, but they extensively use the newly offered services(Mondéjar et al., 2008).

The possible application of data mining techniques for supporting ethio-telecom network operators in the prediction of telecommunication service faults especially broad band network faults is the theme of this study.

The current ethio-telecom is the sole telecommunication company in Ethiopia which was established in 1894 as Ethiopian Telecommunication Agency and since then it has been providing various telecommunication services in the country (Yigzaw et al., 2010). For more than a century, it is the sole and government owned telecom operator of the country. The company got its current name 'Ethio telecom' after the transformation implemented in 2010, and France telecom took the management contract.

As being in the telecommunication industry, Ethio-telecom also collected and stored a large amount of data from the various network elements as alarm fault or directly from the customers using the technicians dedicated for this purpose, via Z-smart system. To extract useful information from these accumulated data, data mining technologies play a big role.
Since the plan was to predict the behavior of new faults based on the previously happened faults, among the various types of data mining techniques, classification/prediction was applied in this study.

1.2. Statement of the problem and its Justification

In every part of the world telecommunication companies are known for providing different kinds of services for their customers. These companies provide their services based on the service agreement they made with their clients.

In the same scenario ethio-telecom has service agreements with each and every customer regarding the service it is providing. On the other hand, the customer also has different obligations to fulfill in order to continue using the service. The telecom

operator is also obliged to fix service faults, when it happens. Additionally, the company will have a chance of losing its revenue if it does not solve the fault problems on time.

Therefore, to give and take the services based on the service agreements made between both parties, they are expected to do their level best. From the ethio-telecom side the first thing which has to be taken care of is to ensure a sustainable service by preventing interruptions that can be achieved through minimizing the fault happening rate or by trying to prevent the fault from happening.

In telecommunication services, a variety of network failures are possible with typical events resulting in a failure being accidental cable cuts, hardware or network elements malfunctions, software errors, natural disasters, human error, and malicious attack (both hardware and software)(Medhi & Tipper, 1997).

Faults in telecommunication services have a direct impact on the availability and maintenance costs of the services. Therefore, predicting faults ahead of time to minimize fault happening rates or if possible eliminating faults from happening by working on the causes that generate them is of special interest; this is very useful to make the systems more robust and the operation more reliable.

Currently, telecom companies are providing various services such as telephone (wire line and wireless), internet (dialup & broad band), Mobile (pre-paid and post paid) and other value added services. Among these services, broadband services are becoming backbones of the economy of a country (Katz, 2011). If we consider the financial system like banks, customs and other government and private businesses, they are heavily dependent on broadband services.

Towards addressing the above mentioned issues, different research works have been done worldwide which are related with telecommunication network fault prediction. Based on the literatures reviewed by the researcher, most of the data mining research works on fault prediction are done based on network device alarm data generated by different network devices.

Faults in a telecommunication network are reported to management centers in the form of alarms. An alarm is a message emitted by a network element, typically when a problem is encountered. Unfortunately, a network element has a very narrow view to the network, and can therefore only report the symptoms of the fault from its limited viewpoint. On the other hand, one fault can result in a number of different alarms from several network elements (Klemettinen et al., 1999).
Since network alarms can be generated as a result of unrelated problems, this might have a negative impact on the result of the model developed based on alarm data (Weiss, 2002).

Currently, Ethio-telecom is using a Z-smart Trouble Ticket information system for collecting and storing various fault alarms from devices and fault reports from clients in to the system database. These fault information are coming either through the call center or by directly calling to dedicated technicians. Faults Alarm are collected automatically but other faults are registered manually by technicians. So far, broadband service faults are among the various service faults registered/created in Z-smart TT information system when customers report faults and the information will be archived when the fault is solved.

The aforementioned system in ethio-telecom registers different types of information related with the fault reports; the company is currently using those registered data only for tracing the fault solving process. Moreover, as confirmed by domain experts, these broad band fault records have not been used for fault prediction, by the company.

But the researcher believe that, the stored data by the system database can be processed and analyzed so that it can be used in predicting faults before they are happening in order to provide sustainable service. In other words, the findings of the research enable the company to take proactive measures and categorize the faults based on their dominance or severity. Possibility to analyze spatial distribution of faults and to predict places where future errors may arise can significantly help Ethio telecom operators who are in charge to detect and repair such problems. The data can also be used for future planning and design of telecommunication networks as it can help to identify

problematic areas and use additional measures for protection of telecommunication cables and equipment.
In general, doing a research on these broad band service fault related records benefits Ethio telecom, customers and the country.

Even if there are a number of local research works done on telecommunication, as per the researcher's knowledge there are no local researches done on predicting broad band service faults.

Therefore, this study aims to develop a predictive model for fault prediction by using the fault related data from Ethio telecom, specifically the data related with broadband network connection faults. An effort has been made to find a pattern which can help to predict future faults in the same area and enables the company to take proactive measure.

1.3. Research Questions

To this end, the research attempts to answer the following research questions.

- Which attributes are more useful for predicting broad band network faults at Ethio telecom.
- Which data mining algorithm can be more suitable to predict broadband network service faults?
- To what extent does the new predictive model give the desired result on the test data set?

1.4. Objective of the Study

Below, the general and specific objectives of this specific study are described as follows:

1.4.1. General Objective

The general objective of the study is to design a predictive model for predicting faults in telecommunication services for ethio-telecom by applying data mining approaches.

1.4.2. Specific Objective

The specific objectives of the planned study are: -

- To review different literatures which are related with the research topic, and also with all the sub tasks, which are going to be accomplished throughout the course of the study.
- To collect the necessary data from Ethio telecom and preprocessing the data so as to make it ready for applying the different techniques which are going to be used.
- To apply the selected data mining techniques on the preprocessed data in order to create predictive model.
- To evaluate the results obtained from applying the new model on test sets and identifying the best approach.

1.5. Research Methodology

The methodology adopted for this study followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) process model and this is because CRISP-DM is considered as the most and broadly adopted data mining process model (Chapman et al., 2000). This model is consisting of six phases intended as a cyclical process. These are Business Understanding; Data Understanding; Data Preparation; Modeling; Evaluation; and Deployment.

For the purpose of accomplishing this study the major tasks which were performed include literature review, data collection and preprocessing, applying data mining techniques, algorithms and finally result evaluation. The approach to each of the tasks is described below.

1.5.1. Literature review

The first research methodology which was used for the success of the study is reviewing different types of literatures from Addis Ababa University libraries and from the internet. The literatures reviewed include different books useful for the study and previously conducted research papers in the area.

1.5.2. Business Understanding

Business understanding for the study was gained through techniques such as interviews, discussions and document observations.

1.5.3. Data Collection and Preprocessing

The target data set for this particular study was collected from Z-Smart Trouble Ticket information system data base at ethio-telecom, and it is a data which is registered and collected in relation with Broadband network service fault reports.

Since the raw data was highly susceptible to noise, missing values, and inconsistency, in order to help improve the quality of the data and, consequently, of the mining results raw data was pre-processed by applying different preprocessing tasks such as data cleaning, data Integration, data Transformation, and data reduction.

1.5.4. Applying data mining techniques

After successfully completing the data preprocessing task, the next step was applying the selected data mining techniques and algorithms on the selected mining tool Weka version 3.7.4. The data mining techniques selected based on related literatures were decision tree with J48 algorithm and Bayesian network with Naïve bayes algorithm. For splitting the data set into training and test sets, Cross validation, and percentage split test options are used.

1.5.5. Evaluation Techniques

The accuracy of the classifier was tested based on the test data managed by Weka, using detail accuracy by class and also based on the analysis made on the confusion matrix results.

1.6. Scope and limitation of the study

The scope of the study was limited to applying data mining techniques for designing a predictive model for predicting broadband network faults in ethio-telecom services. Since the company provides different kind of services like GSM, CDMA, broadband internet and data, ISDN, PSTN and others, it is possible to find different types of network fault report data registered and stored at ethio-telecom for the mentioned services. But this research mainly focus on or limited to broadband internet and data faults, specifically on trouble tickets created for the fault by dedicated technicians using Z-smart systems. The data which was used for this particular study was the data related to broadband network connection faults obtained from ethio-telecom. This was because broadband networking is crucial for the country's economy as well as most widely used networking technology and believed to get special consideration to attain sustainable service provision. In addition, the time was not enough to consider all types of network service faults in this particular study.

There was a limitation for this study which should be mentioned here. The data used for this particular study was expected to be obtained from the data base but due to some sort of internal procedures at ethio-telecom it was not possible to get it from the data base rather it was obtained through the interface of Z-Smart Trouble Ticket information system. Therefore, the study was limited to only some attributes which were accessed through the information system.

1.7. Application and Significance of the Study

Broadband internet access, often shortened to just "broadband", is high speed Internet access that provides download speeds equal to or faster than 256 kbit/s. The standard broadband technologies in most areas are Digital Subscriber Line (DSL) and cable

modems. Newer technologies in use include VDSL (Very High Speed DSL) and pushing optical fiber connections closer to the subscriber in both telephone and cable plants.

Fiber-optic communication has played a crucial role in enabling broadband internet access by making transmission of information over larger distances much more cost-effective than copper wire technology.
The population demand for the broadband services has not stopped rising in the last couple of years. Not only that the number of users is getting higher, but they extensively use the newly offered services (Mondéjar, et al., 2008).

In this thesis the researcher studied the application of data mining techniques for predicting broadband network faults in the telecommunications network operations.
The result of this study can be a great help for ethio-telecom since the company can predict broad band service faults before they are going to happen and based on the prediction made, the company can take different measures so as to minimize the service interruption by: -

- Preventing faults from happening and/or
- By maintaining the faults within a reasonable time.

So that the company can have a good relationship with its customers and also can collect the maximum revenue that it should get from its services. In addition the identified algorithms in this study that best suited for the prediction of faults can be used as an input for future studies in the area.

1.8. Thesis Organization

As an introduction Chapter one consists of background, statement of the problem, objective of the study, methodology, scope and significance of the study. Chapter two describes the various concepts and approaches of data mining and its relation with other fields. It also describes the application areas of data mining in general and in telecommunication industry in particular.
Chapter three discusses selected data mining techniques for this study with the specific algorithms applied.

Chapter four explains the data preparation and preprocessing tasks applied throughout the study.

Based on the selected data mining techniques and algorithms, in chapter Five the different experimentation tasks and the evaluation performed with the results obtained are described.

Finally in chapter six, it has been tried to summarize the whole story of the thesis by way of conclusion and future research directions are recommended.

CHAPTER TWO

LITERATURE REVIEW

The growth of huge databases is a result of the progress in digital data acquisition and storage technology. This has been observed in almost all walks of life, from the ordinary (such as supermarket transaction data, credit card usage records, telephone call details, and government statistics) to the more unusual (such as images of astronomical bodies, molecular databases, and medical records). As a result of having such huge databases, an interest has grown towards tapping these data, so that to extract information which might be useful to the database owner. The discipline concerned with this task has become known as data mining (Hand et al., 2001).

2.1 Data mining overview

"Data mining is the nontrivial process of extracting valid, previously unknown, comprehensible, and useful information from large databases and using it"(Sumathi, 2006). It is an exploratory data analysis, trying to discover useful patterns in data that are not obvious to the data user.

Data mining is a multidisciplinary field, drawing work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization. Data mining emerged during the late 1980s, made great strides during the 1990s, and continues to flourish into the new millennium. (Han et al., 2006; Sumathi, 2006).

The evolution of data mining began when business data was first stored in computers and technologies were generated to allow users to navigate through the data in real time (Pujari, 2001). Data mining takes this evolutionary process beyond retrospective data access and navigation, to prospective and proactive information delivery. The three

sufficiently mature technologies that support the evolution of data mining are: massive data collection, high performance computing and data mining algorithms. The core components of the data mining technologies have been under development for decades, and today the maturity of these techniques coupled with the high performance relational database engines and broad data integration efforts have made these techniques practically applicable.

Kantardzic (2003) noted that, data mining is one of the fastest growing fields in the computer industry. Although it was once a small interest area within computer science and statistics, data mining has quickly expanded into a field of its own. One of the greatest strengths of data mining is reflected in its wide range of methodologies and techniques that can be applied to a host of problem sets.
Data mining techniques have been applied successfully in many areas, from business, science, to sports (Sumathi, 2006).

2.2 General data mining process

"Data mining is not simply a collection of isolated tools, each completely different from the other and waiting to be matched to the problem, but rather it is an iterative process" (Kantardzic, 2003). This process needs to study the data, to examine it using some analytical techniques, to look at it again in another way and modifying it, and then go back to the beginning and apply another data analysis tool to reach to a better of different result. This can be repeated many times; each technique is used to probe slightly different aspects of data—to ask a slightly different question of the data.

Therefore, we can say that data mining is a carefully planned and considered process of deciding what will be most useful, promising, and revealing.

According to Kantardzic (2003), the general experimental procedure adapted to data-mining problems involves the following steps:

2.2.1 State the problem and formulate the hypothesis

It is necessary to have domain specific knowledge and experience in order to come up with meaningful problem statement. This is required because most of the data based modeling studies are performed in a particular application domain.

The process of acquiring such knowledge requires a close interaction between the data-mining expert and the application expert. This cooperation continues during the entire data mining process in successful data mining applications (Kantardzic, 2003).

2.2.2 Collect the data

In order to apply the estimated model successfully, it is very important, to understand how data collection affects its theoretical distribution, since such a priori knowledge can be very useful for modeling and, later, for the final interpretation of results. It is also very important to make sure that the data used for estimating a model and the data used later for testing and applying a model come from the same, unknown, sampling distribution(Kantardzic, 2003).

2.2.3 Preprocessing the data

Data pre-preprocessing is an important step in the data mining process and it includes cleaning, normalization, transformation, feature extraction and selection, etc.

If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Therefore data-preprocessing steps should not be considered completely independent from other data-mining phases. Generally, a good preprocessing method provides an optimal representation for a data-mining technique by incorporating a priori knowledge in the form of application-specific scaling and encoding (Kantardzic, 2003).

2.2.4 Estimate the model

The selection and implementation of the appropriate data-mining technique is the main task in this phase. This process is not straightforward; usually, in practice, the implementation is based on several models, and selecting the best one is an additional task (Kantardzic, 2003).

2.2.5 Interpret the model and draw conclusions

Data mining models are expected to ease the decision making process in most cases. Therefore, the models need to be interpretable in order to be useful. Note that the goals of accuracy of the model and accuracy of its interpretation are somewhat contradictory. Usually, simple models are more interpretable, but they are also less accurate. Modern data-mining methods are expected to yield highly accurate results using high dimensional models (Kantardzic, 2003).

2.3 Data mining in telecom industries

Even if telecom companies have a huge amount of information which exceeds our capacity to analyze it, data-Mining techniques also have different features that make them suitable for analyzing this great amount of data available (Costea, 2006).
The data generated by telecom industries are broadly grouped into three types as Customer data (Demography), Network data, and Bill data (V. Umayaparvathi, 2012). Data mining techniques are applied in telecom database for various purposes. Each uses different type of telecom data depending on the purpose.

Currently, telecommunication companies are using data mining to improve their marketing efforts, identify fraud, and better manage their telecommunication networks. However, 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 are still a challenge to these companies (M.weiss, 2009).

M.weiss (2009) also described that, the popularity of data mining in the telecommunications industry can be viewed as an extension of the use of expert systems in the telecommunications industry. These systems were developed to address the complexity associated with maintaining a huge network infrastructure and the need to maximize network reliability while minimizing labor costs. Because it is both difficult and time consuming to elicit the requisite domain knowledge from experts, developing expert systems become expensive. Therefore, data mining can be viewed as a means of automatically generating some of this knowledge directly from the data.

2.4 Telecommunication Network Faults

Telecommunication networks often contain thousands of components and a problem with one component may quickly propagate through the network. This makes the process of managing the performance of telecommunication networks an extremely complex task (Weiss, 2002). Identifying and isolating faults in these networks is one important aspect of managing network performance. In fact, in order to maintain the availability of these networks, it is critically important to identify a fault before it results in a total failure.

Variety of network failures are possible with typical events resulting in a failure being accidental cable cuts, hardware malfunctions, software errors, natural disasters , human error (e.g., incorrect maintenance), and malicious attack both on hardware and software (Medhi & Tipper, 1997).

A fault is a malfunction that has occurred either in the hardware or software on the network. (Sterritt et al., 2000).

(Hudyma & Fels, 2004), categorize network failures into seven types these are: -

2.4.1 Hardware Problems

Variations in the quality of equipment, the quality of network planning and design, complexity of the implementation, the interaction and interoperability of components at the time of actual deployment of the networks, can be resulted in hard ware problems (Hudyma & Fels, 2004).

2.4.2 Software Problems

Faulty device drivers, subtle differences in protocol implementation and handling, and operating system faults and anomalies can be mentioned as causes of network software failures (Hudyma & Fels, 2004).

2.4.3 Operator Errors

"An operator error that affects the network reliability can arise from people's interaction with networking equipment, physical cables and connectors as well as from events by other IT devices result from user actions" (Hudyma & Fels, 2004).

2.4.4 Mass Storage Devices

"Although the failure of these devices is not by itself considered to be a network failure, there has been a rapid growth in the deployment of Storage Area Networks (SAN) where large arrays of mass storage devices are directly connected to a network through high capacity channels" (Hudyma & Fels, 2004).

2.4.5 Network Problems

"Hardware and Software problems that are directly related to the Network are included in this category" (Hudyma & Fels, 2004).

2.4.6 Denial of Service Attack

Denial of Service attacks has been a major source of network failures since 2000. Example of the impact of Denial of Service attacks is the Code Red virus and a more recent variation, Slammer worm, disrupted millions of computers by unleashing a well coordinated Distributed Denial of Service Attack. These attacks resulted in a significant loss of corporate revenues worldwide (Hudyma & Fels, 2004).

2.4.7 Disaster Scenarios

The final category of failure considered is that of disaster scenarios which occur from a wide range of circumstances, many of them environmental and some synthetic. Environmental disasters include floods, earthquakes, hurricanes, long term power outages, tornadoes and fires. Synthesized disasters can include: theft, vandalism, arson, war and acts of terrorism (Hudyma & Fels, 2004).

Currently, Quality of Service and customer satisfaction with profit maximization is the vision of any service providing company. If a company in the telecom industry have applications and databases, the logical next question is how to use this existing data to

predict locations of future faults (with acceptable margins of probability) and prevent or minimize network downtime (Medved et al., 2000).

2.5 Data Mining and Knowledge discovery from databases (KDD)

Converting data into knowledge in traditional methods is based on manual analysis and interpretation. "In any field the classical approach to data analysis relies fundamentally on one or more analysts becoming intimately familiar with the data and serving as an interface between the data and the users and products" (Fayyad, 1996).

As data volume grow significantly, manual data analysis will become impossible. Currently, computers have enabled us to gather more data than we can digest. Therefore, it is normal to look for computational techniques to help us discover meaningful patterns and structures from the huge volumes of data.

According Sumathi (2006), "an enormous proliferation of databases in almost every area of human endeavor has created a great demand for new, powerful tools for turning data into useful, task-oriented knowledge". To address this need, researchers have been exploring ideas and methods developed in machine learning, pattern recognition, statistical data analysis, data visualization, neural nets, etc. With their hard work, the researchers came up with a new research area, frequently called data mining and knowledge discovery.

The task of finding useful patterns in data has been historically given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing (Fayyad, 1996).

The term data mining has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities. It has also gained popularity in the database field.

The term knowledge discovery in databases or KDD, for short, was coined in 1989 to refer to the broad process of finding knowledge in data, and to emphasize the "high-level" application of particular DM (Ana Azevedo, 2008).

KDD has evolved, and continues to evolve, from the intersection of research in fields such as databases, machine learning, pattern recognition, statistics, artificial intelligence and reasoning with uncertainty, knowledge acquisition for expert systems, data visualization, machine discovery, scientific discovery, information retrieval, and high-performance computing (Fayyad et al., 1996).

KDD focuses on the overall process of discovering useful knowledge from data including how the data are stored and accessed, how algorithms can be scaled to massive data sets and still run efficiently, how results can be interpreted and visualized, and how the overall man-machine interaction can usefully be modeled and supported.

Data mining is a particular step in the KDD process and it is the application of specific algorithms for extracting patterns from data. Currently data mining is regarded as the key element of the KDD process (Fayyad et al., 1996).

Data mining involves fitting models to or determining patterns from observed data. The fitted models play the role of inferred knowledge. Deciding whether or not the models reflect useful knowledge is a part of the overall interactive KDD process for which subjective human judgment is usually required (Fayyad, et al., 1996).

These days, the term Data Mining is becoming more popular than KDD (Han, et al., 2006).

2.6 Data mining process models

The common process models describing the data mining process are the KDD (Knowledge Discovery in Databases) process model, SEMMA, CRISP-DM (CRoss Industry Standard Process for Data Mining) and Reinartz's framework (Welcker et al., 2012).

Due to time constraint, for the purpose of this study three of the above mentioned data mining process models (KDD, SEMMA, and CRISP-DM) will be discussed.

2.6.1 The KDD Process model

The KDD process is an organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets (Maimon & Rokach, 2005). As shown in figure 2:1 below, the KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user (Fayyad, et al., 1996). The numerous steps are summarized in nine steps as follows: -

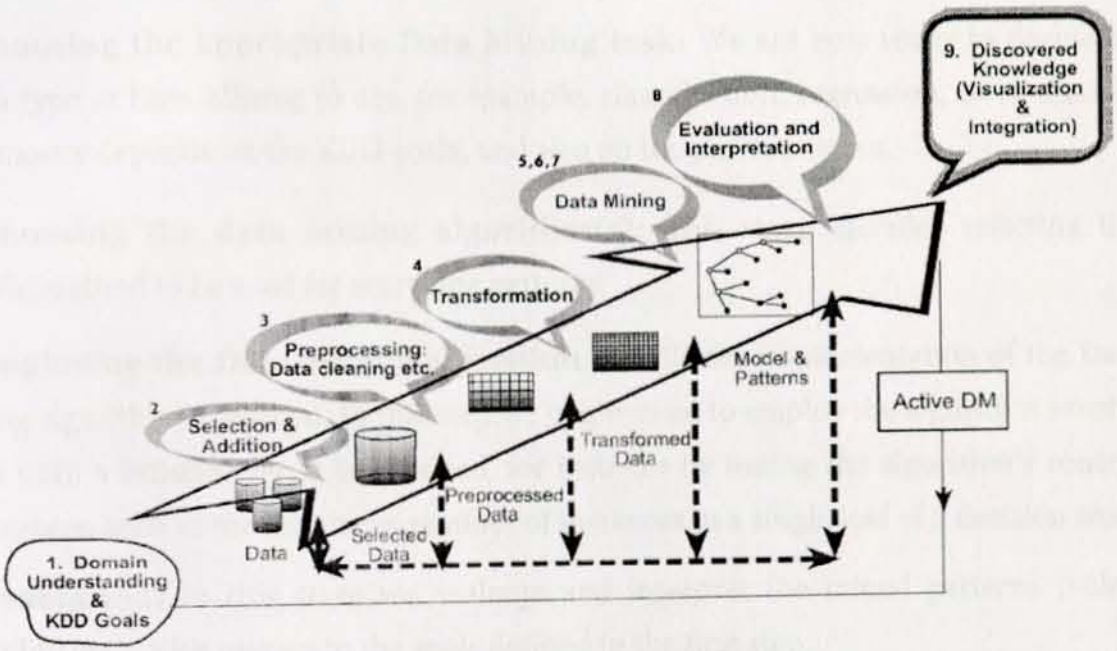


Figure 2: 1 **Process of knowledge discovery in databases (Maimon & Rokach, 2005)**

Following is a brief description of the nine-step KDD process by (Maimon & Rokach, 2005):-

1. Developing an understanding of the application domain: The people who are in charge of a KDD project need to understand and define the goals of the end-user and the environment in which the knowledge discovery process will take place (including relevant prior knowledge).

2. Selecting and creating a data set on which discovery will be performed: This includes finding out what data is available, obtaining additional necessary data,

and then integrating all the data for the knowledge discovery into one data set, including the attributes that will be considered for the process.

3. Preprocessing and data cleansing: In this stage, data reliability is enhanced. It includes data clearing, such as handling missing values and removal of noise or outliers.

4. Data transformation: In this stage, the generation of better data for the data mining is prepared and developed.

5. Choosing the appropriate Data Mining task: We are now ready to decide on which type of Data Mining to use, for example, classification, regression, or clustering. This mostly depends on the KDD goals, and also on the previous steps.

6. Choosing the data mining algorithm(s): This stage includes selecting the specific method to be used for searching patterns

7. Employing the Data Mining algorithm: Finally the implementation of the Data Mining algorithm is reached. In this step we might need to employ the algorithm several times until a satisfied result is obtained, for instance by tuning the algorithm's control parameters, such as the minimum number of instances in a single leaf of a decision tree.

8. Evaluation: In this stage we evaluate and interpret the mined patterns (rules, reliability etc.), with respect to the goals defined in the first step.

9. Using discovered knowledge: includes incorporating the knowledge into the 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 potential conflicts with previously believed (or extracted) knowledge.

2.6.2 The SEMMA process model

The SEMMA process was developed by the SAS Institute (Ana Azevedo, 2008). The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a data mining project.
The SAS Institute considers a cycle with 5 stages for the process:

1. 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. This stage is pointed out as being optional.
2. Explore – This stage consists on the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.
3. Modify – This stage consists on the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.
4. 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.
5. Assess – This stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs.

SEMMA offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects.

2.6.3 The CRISP-DM process model

The Cross-Industry Standard Process for Data Mining (CRISP-DM) process was developed by the means of the effort of a consortium initially composed with DaimlerChryrler, SPSS and NCR (Ana Azevedo, 2008).

The life cycle of CRISP-DM consists of six phases. As shown in Figure 2:2, the sequence of the phases is not rigid. Moving back and forth between different phases is always required. The outcome of each phase determines which phase, or particular task of a phase, has to be performed next. The arrows indicate the most important and frequent dependencies between phases (Chapman, et al., 2000).

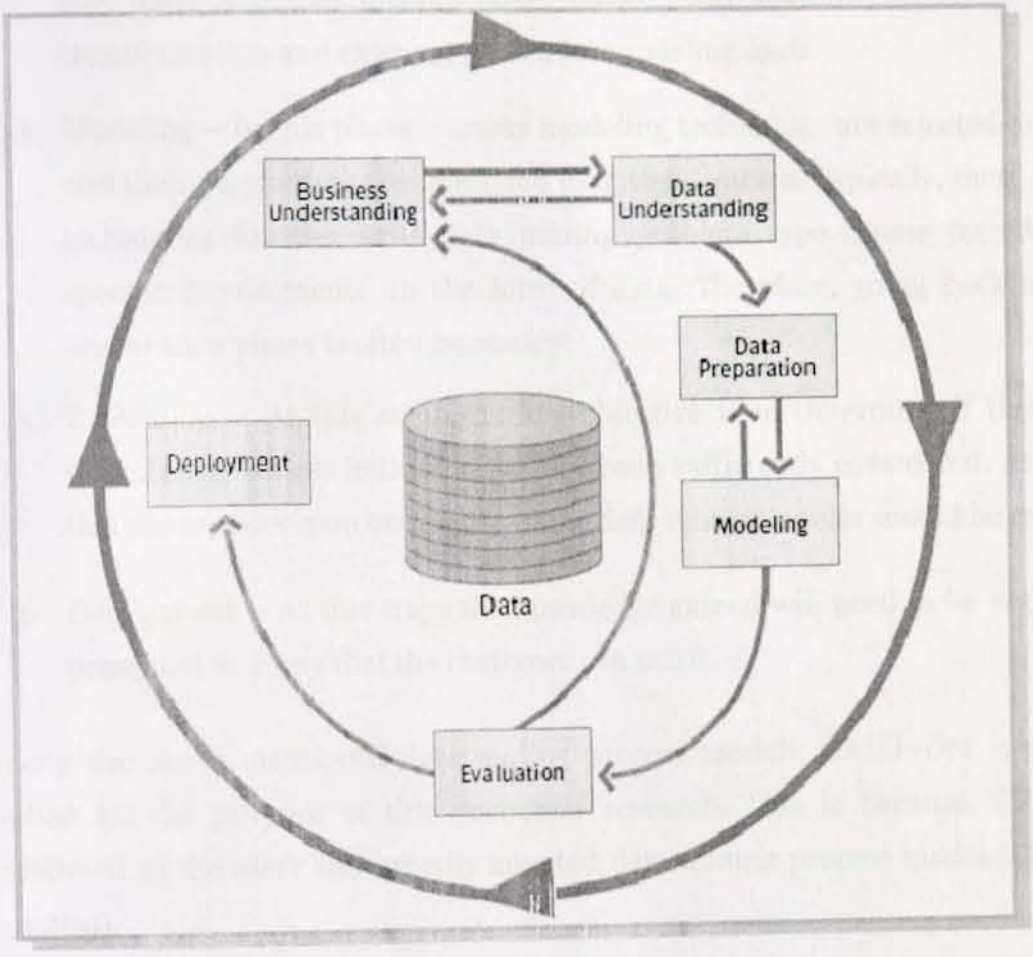


Figure 2 : 2 **CRISP-DM life cycle (Chapman, et al., 2000)**

1. Business understanding – Focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.
2. Data understanding – This phase starts with initial data collection and proceeds with activities that enable you to become familiar with the data, identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information.
3. Data preparation – The data preparation phase covers all activities needed to construct the final dataset that will be fed into the modeling tool from the initial

raw data. Activities include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools.

4. Modeling – In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary.
5. Evaluation – At this stage the key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.
6. Deployment – At this stage the knowledge gained will need to be organized and presented in a way that the customer can use it.

Among the above mentioned data mining process models, CRISP-DM is going to be applied for the purpose of this particular research. This is because, CRISP-DM is considered as the most and broadly adopted data mining process model (Chapman, et al., 2000).

2.7 Data Mining Tasks

Due to the existence of many patterns in a large database, the tasks in data mining are very diverse and distinct (Sumathi, 2006). It requires applying different methods and techniques in order to find different types of patterns. Tasks in data mining can be classified into summarization, classification, clustering, association, and trend analysis based on the patterns we are looking for. This data mining tasks are briefly discussed below.

2.7.1 Summarization

Summarization is the process of summarizing and abstracting a set of task-relevant data. As a result of this process, a general overview of the data will be obtained in a smaller set. The process can result into different abstraction level and can be viewed

from different angles. By combining the different abstraction levels, various kinds of patterns and regularities can be obtained. (Sumathi, 2006).

2.7.2 Classification

Classification derives a function or model, which determines the class of an object, based on its attributes (Sumathi, 2006). A set of objects is given as the training set. In it, every object is represented by a vector of attributes along with its class. A classification function or model is constructed by analyzing the relationship between the attributes and the classes of the objects in the training set which can be applied for classifying future objects.

Among the different types of data mining tasks, classification is going to be applied for this specific research. This because the aim of the study is to classify the data set and then based on it to predict future data sets.

2.7.3 Association

Association is the discovery of rules that reveals the associative relationships among objects. Association rules can be useful for marketing, commodity management, advertising, and so forth (Sumathi, 2006).

2.7.4 Clustering

Clustering is the process of identifying classes or groups for a set of objects whose classes are unknown (Sumathi, 2006). This is done based on some criteria defined on the attributes of the objects. Once the clusters are decided the objects are labeled with their corresponding clusters. The common features for object in a cluster are summarized to form the class description. As result of the clustering process, it will be observed that an increase in similarity among objects within the same cluster, and a decrease in similarity among object in different clusters.

2.7.5 Trend analysis

Accumulated records over time are called time series data, and such kind of data can be viewed as objects with an attribute time. The objects are snapshots of entries with values that change over time. Finding the patterns and regularities in the data evolution along the dimension of time can be fascinating (Sumathi, 2006).

2.8 Data mining techniques

Data mining adopted its techniques from many research areas, including statistics, machine learning, database systems, neural networks, rough sets, and visualization (Sumathi, 2006).

There are many up-to-date techniques for the different types of data mining tasks such as for classification (decision trees, na¨ıve Bayes classifier, k-nearest neighbor, neural networks), for clustering (k-means, hierarchical clustering, density-based clustering), for association (one-dimensional, multidimensional, multilevel association, constraint-based association), and so on.

According to S. Sumathi (2006), data mining techniques can be categorized based on their approach as follows: -

2.8.1 Statistical approaches

Many statistical tools have been used for data mining, including Bayesian network, regression analysis, cluster analysis, and correlation analysis. Usually statistical models are built from a set of training data. An optimal model, based on a defined statistical measure, is searched among the hypothesis space. Rules, patterns, and regularities are then drawn from the models (Sumathi, 2006).

- A Bayesian network is directed graph computed using the Bayesian probability theorem. It represents the causal relationships among the variables.
- Regression is the function derivation, which maps a set of attributes of objects to an output variable.
- Correlation analysis studies the correspondence of variables to each other.
- Cluster analysis finds groups from a set of objects based on distance measures.

2.8.2 Machine learning approaches

Like statistical methods, machine-learning methods search for the best model that matches the testing data. Unlike statistical methods, the searching space is a cognitive space of n attributes instead of a vector space of n dimensions. Besides that, most machine learning methods use heuristics in the search process.

The most common machine learning methods used for data mining include decision tree, inductive concept learning, and conceptual clustering (Sumathi, 2006).

- A decision tree is a classification tree, which determines an object's class by following the path from the root to a leaf node. It chooses the branches according to the attribute values of the object. Decision trees are induced from the training set. Classification rules can be extracted from the decision trees.
- Inductive concept learning derives a concise, logical description of a concept from a set of examples.
- Conceptual clustering finds groups or clusters in a set of objects based on conceptual closeness among objects.

2.8.3 Database-oriented approaches

Database-oriented methods do not search for a best model, as do the previous two methods. Instead, data modeling or database specific heuristics are used to exploit the characteristics of data in hand. The attribute-oriented induction, the iterative database scanning for frequent item sets, and the attribute focusing are representatives of the database-oriented methods (Sumathi, 2006).

2.8.4 Other approaches

Many other techniques have been adopted for data mining, including neural networks, rough sets, and visualization (Sumathi, 2006).

- A neural network is a set of interlinked nodes called *neurons*. A neuron is a simple device that computes a function of its inputs. The inputs can be outputs of other neurons or attribute values of an object. By adjusting the connection and the functional parameters of the neurons, a neural network can be trained to model the relationship between a set of input attributes and an output attribute. A neural network can be used, for example, in classification when the output attribute is the object class.
- A rough set is a set whose membership is fuzzy. A set of objects can be arranged to form a group of rough sets for use, in say, classification and clustering.
- Visual exploration is another interesting data mining technique. Data are transformed into visual objects such as dots, lines, and areas. The data is then

displayed in a two- or three-dimensional space. Users can interactively explore the interesting spots by visual examination.

These methods can be integrated or combined to deal with complicated probabilities, or provide solutions. Indeed most data mining systems employ multiple methods to deal; with different kinds of data, different data mining tasks, and different application areas (Sumathi, 2006).

2.9 Data mining and other related fields and technologies

Data mining has its origins in various disciplines, of which the two most important are statistics and machine learning (Kantardzic, 2003). For the purpose of this study the relation of data mining with machine learning and statistics are discussed. More over two technologies applied in data mining – data warehouse and OLAP - will also be explained.

2.9.1 Data mining and Machine learning

Data mining adopted its techniques from many research areas, including statistics, machine learning, database systems, neural networks, rough sets, and visualization (Sumathi, 2006).

"Machine learning is the study of computational methods for improving performance by mechanizing the acquisition of knowledge from experience" (Jackson, 2002). Machine learning aims to provide increasing levels of automation in the knowledge engineering process, replacing much time-consuming human activity with automatic techniques that improve accuracy or efficiency by discovering and exploiting regularities in training data. Although machine learning algorithms are central to the data mining process, it is important to note that the process also involves other important steps, including:

- building and maintaining the database,
- data formatting and cleansing,
- data visualization and summarization,

- the use of human expert knowledge to formulate the inputs to the learning algorithm and to evaluate the empirical regularities it discovers, and
- Determining how to deploy the results.

The most common machine learning methods used for data mining include decision tree, inductive concept learning, and conceptual clustering (Sumathi, 2006).

2.9.2 Data Mining and statistics

"Statistics is the traditional field that deals with the quantification, collection, analysis, interpretation, and drawing conclusions from data" (Benjamini & Leshno, 2005). It provides a language and framework for quantifying the uncertainty resulting when one tries to infer general patterns from a particular sample of an overall population (Fayyad, et al., 1996).

As Jackson (2002) described, the disciplines of statistics and data mining, both aim to discover structure in data. So much do their aims overlap, that some people regard data mining as a subset of statistics. But that is not a realistic assessment as data mining also makes use of ideas, tools, and methods from other areas – particularly database technology and machine learning, and is not heavily concerned with some areas in which statisticians are interested.

Statistical procedures do, however, play a major role in data mining, particularly in the processes of developing and assessing models. Most of the learning algorithms use statistical tests when constructing rules or trees and also for correcting models that are over fitted.

Data mining differs from traditional statistics on two issues: the size of the data set and the fact that the data were initially collected for purpose other than that of the DM analysis. Thus, *experimental design*, a very important topic in traditional statistics, is usually irrelevant to DM. On the other hand asymptotic analysis, sometimes criticized in statistics as being irrelevant, becomes very relevant in DM (Benjamini & Leshno, 2005).

2.9.3 Data warehousing, OLAP and Data Mining

Daily stored data into production systems at the end should serve to management. Administrative structure of the company should be able to extract useful information from large amounts of data, and use it for evaluating the results achieved, planning and making business decisions. For this purpose it is necessary to ensure a quick and easy access to data stored in complex structures of production systems.

Nowadays, organizational data can be stored in many different types of databases. One data base architecture that has recently emerged is the "data warehouse", which is a repository of multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management decision-making (Reddy, Srinivasu, Rao, & Rikkula, 2010).

Data warehouse technology includes data cleaning, data integrating, and on-line analytical processing (OLAP) that is, analysis techniques with functionalities such as summarization, consolidation and aggregation, as well as the ability to view information from different angles.

According to Jackson (2002), the data cleaning and data integration tasks in the construction of a data warehouse, can be viewed as an important preprocessing step for data mining. However, a data warehouse is not a requirement for data mining. The data to be mined can be extracted from one or more operational or transactional databases or data marts if it is not possible to have a data warehouse. It is also possible to use a logical or a physical subset of a data warehouse as a database for mining.

From a data warehouse perspective, data mining can be viewed as an advanced stage of on-line analytical processing (OLAP). However, data mining goes far beyond the narrow scope of summarization-style analytical processing of data warehouse systems by incorporating more advanced techniques for data analysis (Han, et al., 2006).
Han, et al. (2006) describe the relationship of data warehousing and OLAP with data mining as: - "The capability of OLAP to provide multiple and dynamic views of summarized data in a data warehouse sets a solid foundation for successful data mining".

2.10 Application Areas of Data mining

Many organizations use data mining to help manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers (Han, et al., 2006).

By determining characteristics of good customers (profiling), a company can target prospects with similar characteristics. By profiling customers who bought a particular product a firm can focus attention on similar customers who have not bought that product (cross selling).

Profiling also enables a company to retain customers who are at risk for leaving (reducing churn or attrition), because it is usually far less expensive to retain a customer than acquire a new one.

Examples of other industries where data mining can make a contribution include:

- Telecommunications and credit card companies are two of the leaders in applying data mining to detect fraudulent use of their services.
- Insurance companies and stock exchanges are interested in applying data mining to reduce fraud.
- Medical applications use data mining to predict the effectiveness of surgical procedures, medical tests, or medications.
- Financial firms use data mining to determine market and industry characteristics as well as to predict individual company and stock performance.
- Retailers make use of data mining to decide which products to stock in particular stores (and even how to place them within a store), as well as to assess the effectiveness of promotions and coupons.
- Pharmaceutical firms mine large databases for chemical compounds and genetic material to discover substances that might be candidates for development as agents for the treatments of disease.

2.10.1 Application of Data mining in the Telecommunication industry

The telecommunications industry was an early adopter of data mining technology and therefore many data mining applications exist (Weiss, 2005). These applications are divided into three application areas: fraud detection, marketing/customer profiling and network fault isolation.

- **Fraud Detection**

Fraud is a serious problem for telecommunication companies, leading to billions of dollars in lost revenue each year (Weiss, 2005). Fraud can be divided into two categories: subscription fraud and superimposition fraud. Subscription fraud occurs when a customer opens an account with the intention of never paying for the account charges. Superimposition fraud involves a legitimate account with some legitimate activity, but also includes some "superimposed" illegitimate activity by a person other than the account holder. Superimposition fraud poses a bigger problem for the telecommunications industry.

The most common method for identifying fraud is to build a profile of customer's calling behavior and compare recent activity against this behavior. Thus, this data mining application relies on deviation detection.

- **Marketing / Customer Profiling**

Telecommunication companies maintain a great deal of data about their customers. A serious issue with telecommunication companies is *customer churn*. Customer churn involves a customer leaving one telecommunication company for another. Customer churn is a significant problem because of the associated loss of revenue and the high cost of attracting new customers (Weiss, 2005).

Thus, it is often useful to profile customers based on their patterns of phone usage, which can be extracted from the call detail data. These customer profiles can then be used for marketing purposes, or to better understand the customer, which in turn may lead to better forecasting models.

- **Network Fault Isolation**

The other data mining application area in the telecom industry, in which the task of this specific study can be grouped into, is network fault isolation.

As Weiss (2005) described, telecommunication networks are extremely complex configurations of hardware and software. Most of the network elements are capable of at least limited self-diagnosis, and these elements may collectively generate millions of status and alarm messages each month. In order to effectively manage the network, alarms must be analyzed automatically in order to identify network faults in a timely manner.

Because of the volume of the data, and because a single fault may cause many different, seemingly unrelated, alarms to be generated, the task of network fault isolation is quite difficult. Data mining has a role to play in generating rules for identifying faults.

2.11 Related works

There are a number of studies done in the field of applying data mining techniques in telecom industry and specially for predicting telecommunication network faults.

One of the studies, presented a temporal alarm prediction scheme for fault-prediction in a Telecommunications Network based on a predictive tool known as TimeSleuth (Jaudet et al., 2005). They selected a countrywide data network of Pakistan Telecom (PTCL) as a basis for the investigation of classification algorithms to predict faults. The main problems addressed were the evaluation of alarms and development of new machine learning tools to help overcome the interoperability issues.

The motivation behind their work was to assist human operators and minimize the cost of the alarm evaluation process. During the study, they employed and adapted Time-Sleuth (decision tree based algorithm- C4.5 variant) software for analysis of alarm messages and fault-prediction. The collected data for their study had 60,000 alarms which had occurred over the previous four days. The main variable of interest for their data was separation of time between consecutive alarms. Upon their experimentation, accuracy of rules for having classification tree and its rules stayed around 80-85% and

accuracy for prediction of groups stayed around 90%. They describe that their result based on the c4.5 variant were quite accurate in terms of alarm symbol prediction.

Another study describes a project in which a temporal data mining system called Time-weaver is used to identify faulty telecommunication equipment from logs of network alarm messages (Weiss, 2002). The motivation of the study was to minimize the time taking and costly process of knowledge acquisition in the telecommunication domain. For performing their task they developed TimeWeaver data mining software package which is a genetic-based data mining system that evolves populations of prediction patterns in order to solve event prediction problems. The target dataset was formed by collecting two week worth of alarms which contain 148,886 alarms from which 1045 distinct alarms were made ready after preprocessing. Each alarm contains approximately 20 variables from which five variables were selected based on their importance for diagnosis and the available domain knowledge. The data mining task selected for their study was prediction task. The selected data mining task requires an algorithm that can identify predictive sequential and temporal patterns in the network alarm data. Because existing methods and software packages were not suitable for performing this task, the researchers develop a software package called Time-Weaver. This software package is a genetic-based data mining system that involves populations of prediction patterns in order to solve event prediction problems. Upon their experimentation, the researchers split the target dataset in to disjoint training and test sets by placing 70% into the training set and the remaining into the test set. As a result it was observed that the precision of the predictions decreases as the recall of the predictions increases. It was also observed that, the shorter the warning time the better gets the predictions.

The other study, described a formalism to model the behavior of telecommunications networks when a fault occurs and how the effects are propagated across equipment (Aghasaryan et al., 2002). The objective of their work was to ease the construction and the update of a model corresponding to the supervised network; According to researchers, their model can be used to simulate fault propagation in the network and also to process on-line diagnosis and determine primary causes of a set of observed

alarms. They also mentioned that, even if their experiment was carried out for a small network, they reasonably suppose that the approach can be extended successfully to larger, more complex networks, for the following reasons: (i) the number of classes of component will always be small, and (ii) the distributed nature of the algorithms makes them well adaptable when the size of the network grows. They also disclosed the main limitation of their approach as that the structural model is assumed to be unchanged, which is not the case in real telecommunications networks.

Another research presented a process to perform data mining on electric power fault information reports (Rayudu & Maharaj). The researcher's motivation behind their study was to device a data mining strategy to analyze power system fault information reports. Initially the researchers selected around 470 events spanning over three years. After applying preprocessing, 342 events with 30 attributes were used for training phase of the dada mining process. Three different algorithms (SSV decision tree, K nearest neighbors, and Feature Space Mapping (FSM)) used for data mining the data set.
As the researchers mentioned, no single algorithm had performed well against all outputs. All algorithms for all predictors had an average accuracy of about 66.7%. The researchers present some points that can help in improving the accuracy obtained such as reiterating the pre-processing phase and converting some important information which is in English to a machine readable language.

There are also few local research works which are conducted on the telecommunication domain such as: -"Using Data Mining to Combat Infrastructure Inefficiencies: The Case of Predicting Non-payment for Ethiopian telecommunication corporation"(Yigzaw et al., 2010), "Application of Data Mining Techinques to Customer Relationship Management (CRM): The case of Ethiopian Telecommunications Corporation" (Girma, M., 2009), and "Data Mining Application in Supporting Fraud Detection on Ethio-Mobile Services" (Gebremeskel, 2006).

As it can be seen from the above discussion, all the research works related with network fault prediction are based on network device alarm data. Which is different from the data used for this particular study, which is a data related with the broad band network

service faults information report data of Ethio-telecom – (detail explanation about the data used for this study will be presented on the coming chapters).

In addition, as per the researcher's knowledge, none of the local researches are conducted on a related data or research topic with this specific study.

Therefore, the researcher believes that, this research is unique and contributes new knowledge to the existing body of knowledge pertaining to the application of data mining techniques for fault prediction.

CHAPTER THREE

DATA MINING TECHNIQUES

In data mining, classification is one of the most important tasks. The aim of classification task is to build a classifier based on some given training data set. Then, the obtained classifier will be applied for predicting the class of new attributes. The commonly used methods for data mining classification tasks can be classified into the following groups - Decision Trees (DT's), Support Vector Machine (SVM), Genetic Algorithms (GAs) / Evolutionary Programming (EP), Fuzzy Sets, Neural Networks, and Rough Sets.

One of the classification techniques applied in this specific research is Decision trees. Decision tree is selected because its representation of acquired knowledge in tree form is intuitive and generally easy to understand by humans, the learning and classification steps of decision tree induction are simple and fast, and decision tree classifiers have in general good accuracy. Moreover, decision trees have a capability of handling missing values and compared to other sophisticated models, they can be generated very quickly. (Frank, 2000; Han, et al., 2006; Pyle, 1999).

The other data mining technique selected for this study is Bayesian Network. There are four main reasons for selecting Bayesian Networks for fault prediction system (Medved, et al., 2000). These are: -

- Bayesian networks can readily handle incomplete data sets.
- Bayesian networks allow learning about casual relationships.
- Bayesian networks in conjunction with Bayesian statistical techniques facilitate the combination of domain knowledge and data.
- Bayesian methods in conjunction with Bayesian networks and other types of models offers an efficient and principled approach for avoiding the over fitting of data.

In addition, among several research areas that are proved to be useful as a method of reasoning, network fault detection is one of the diverse numbers of application areas where Bayesian network modeling is exploited (Cayci et al.).
For this particular study, the J48 algorithm (WEKA's implementation of the C4.5 decision tree learner) and Naive-Bayes Bayesian network are applied for building the classification models from the available data set.

The researcher selected two different types of data mining techniques for the purpose of comparing experimental results.

3.1 Decision Tree

"A decision tree is a decision-modeling tool that graphically displays the classification process of a given input for a given output class labels"(Drazin & Montag, 2012).
It recursively partitions a data set of records using depth-first greedy approach or breadth-first approach until all the data items belong to a particular class. A decision tree structure is made of root, internal and leaf nodes. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values.

The tree structure is used in classifying unknown data records. At each internal node of the tree, a decision of best split is made using impurity measures. The tree leaves are made up of the class labels which the data items have been group. Decision tree classification technique is performed in two phases: tree building and tree pruning (Anyanwu & Shiva, 2009).

3.1.1 Tree building

Decision trees are produced by algorithms that identify various ways of splitting a dataset into branch-like segments. Tree building is done in a recursive top-down manner by partitioning the tree until all the data items belong to the predefined class labels. The tree will be displayed in an inverted form with the root node (the node that contains the data set to be examined) at the top. Naturally, it consists of the values of the

field that will be partitioned or examined as the decision tree grows. Because this field is the target of the analysis, it is often called the target; however, because its values can be dependent on the values of the fields that will be used to examine it, then it can also be called a dependent field or variable.

In order to retain strong relationships between inputs and the target, important inputs will be selected as a splitting criterion. In addition, the decision tree criteria will separate important branches from unimportant one. Inputs are referred to as predictors or classifiers because their values can be used to predict target values or classify target values.

The tree branches can be 2-way (binary) or multi-way (many) and are formed by splitting the target values with respect to the corresponding values in the inputs (De Ville, 2006). When a branch is identified with its associated leaves or nodes, then the members of each leaf or node are expected to be as homogenous as possible with respect to their relationship with the target. In addition, each leaf or node is maximally distinguished or differentiated from other nodes on the same branch of the decision tree.

There are different statistical measures for analyzing the association or relation of one set of value with the other. For example, measure of information gain identify, how much information about a target can be gained through knowing the corresponding information about an input. On the other hand, a measure of purity identify, how homogenous or diversified are the members of a branch of the tree. It is possible to review the partitions or classifications formed by various inputs and to either select an input based on the numerical properties of the partitioning mechanism, or to select an input based on business rules (De Ville, 2006). A general decision tree construction process is illustrated in figure 3:1 which shows the steps in building a decision tree starting from the root node towards the leaf node by applying the splitting criteria.

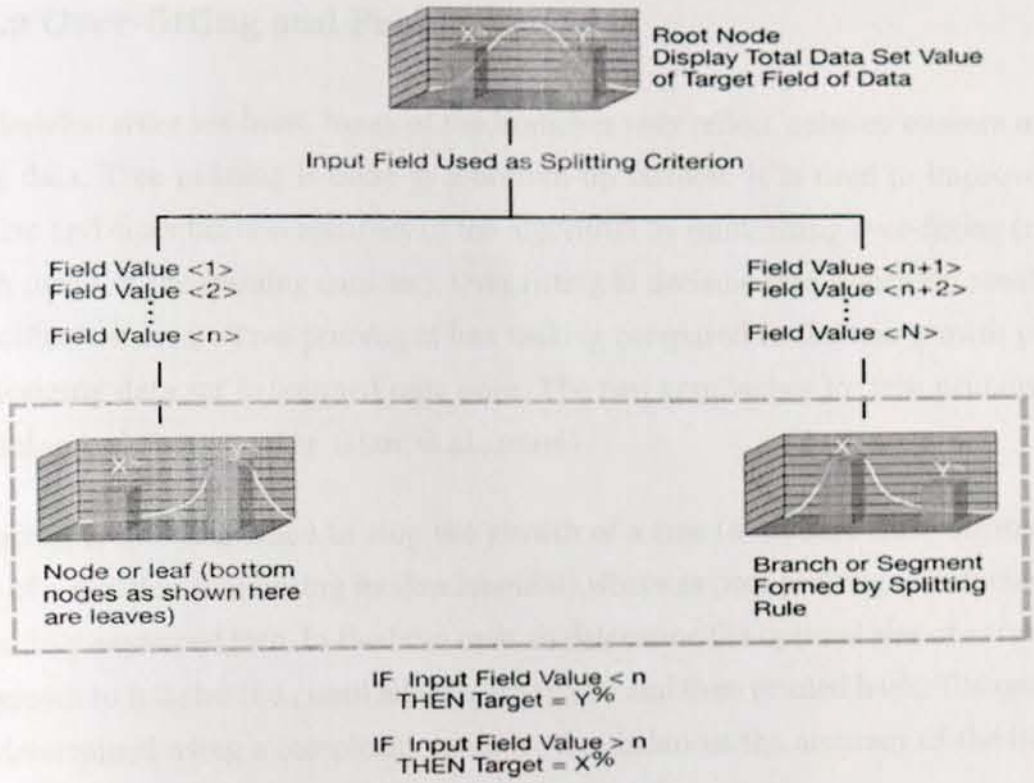


Figure 3:1 **Illustration of decision tree (De Ville, 2006)**

Decision trees are said to be grown recursively; that is, once the initial or root node is split into a branch, all subsequent nodes are also split using the same methodology. This process continues as the decision tree is grown until it either runs out of data in the descendent node, or the growth is stopped according to a stopping rule. This is called recursive partitioning growth. Various stopping rules can suggest when recursive partitioning should be stopped. It is necessary to stop at some point because deep decision trees are more complicated to understand and less useful. The validity, accuracy, and reproducibility of the decision tree can be tested through validation(De Ville, 2006).

The results, interpretation, and application of decision trees can be described, semantically, as simple IF <condition> THEN <action> rules. This way of describing relationships is very general and close to natural language, so it is readily understandable in non-scientific situations.

3.1.2 Over-fitting and Pruning

When decision trees are built, many of the branches may reflect noise or outliers in the training data. Tree pruning is done in a bottom-up fashion. It is used to improve the prediction and classification accuracy of the algorithm by minimizing over-fitting (noise or much detail in the training data set). Over fitting in decision tree algorithm results in misclassification error. Tree pruning is less tasking compared to the tree growth phase as the training data set is scanned only once. The two approaches to tree pruning are Pre-pruning and Post-pruning (Han, et al., 2006) .

Pre-pruning is deciding when to stop the growth of a tree (a method that controls the growth of a decision tree during its development) where as post pruning is reducing the size of a fully expanded tree. In the later case, to determine the optimal size of a tree, the tree is grown to full size (i.e., until all data are spent) and then pruned back. The optimal size is determined using a complexity measure that balances the accuracy of the tree as measured by cost complexity and by the size of the tree.

3.1.3 C4.5 decision tree algorithm

Among the various types of decision tree algorithms, the most well-known algorithm in the literature for building decision trees is the C4.5 (Kotsiantis, Zaharakis, & Pintelas, 2007).

C4.5 algorithm is a successor of ID3 (Iterative Dichotomiser). It adopts a greedy approach in which decision trees are constructed in a top-down recursive divide-and – conquer manner. C4.5 handles continuous and discrete attributes. Whenever a set of items (training set) is encountered, the algorithm identifies the attribute that discriminates the various instances most clearly. This is done using the standard equation of information gain(patel).

Among the possible values of this feature, if there is any value for which there is no ambiguity, i.e., for which the data instances falling within its category have the same value for the target variable, then that branch is terminated and the obtained target value is assigned to it; For all other cases, another set of attributes are looked at that

gives the highest information gain; This is continued in the same manner until either a clear decision of the value of the target variable is reached with a combination of conditions on various independent variables/attributes, or running out of attributes.

In the event of running out of attributes, or getting an ambiguous result from the available information, the branch is assigned a target value that the majority of the items under this branch possess.

3.1.4 Pros and cons of decision trees

Decision tree have an advantage over numerous techniques in that the output of a decision tree is transparent, which makes it easy for users or non technical persons to understand.

But decision tree techniques also have scalability and efficiency problems, such as substantial decrease in performance and poor use of available system resources (Danso, 2006).

3.2 Bayesian Networks

A Bayesian network is a structure that shows the conditional dependencies between domain variables and may also be used to illustrate graphically the probabilistic causal relationships among domain variables. A Bayesian network consists of a directed acyclic graph and probability tables. The nodes of the network represent the domain variables and an arc between two nodes (parent and child) indicates the existence of a causal relationship or dependency among these two nodes. Associated with each node there exist a probability table (PT). If the node has no parents, its probability table contains the prior probabilities else the conditional probabilities between the node and its parents. Although the domain variables can be continuous, they are discretized most of the time for simplicity and efficiency. Besides representing the dependencies between domain variables, a Bayesian network is used for inferencing the probability of a variable given the observations of other variables (Cayci, et al.).

3.2.1 Graphical Structure of Bayesian Networks

A Bayesian network is a graphical structure that allows us to represent and reason about an uncertain domain (Bashar, Parr, McClean, Scotney, & Nauck, 2010). For a set of variables $X = \{X1, X2, X3...,Xn\}$, a Bayesian network consists of a network structure S that encodes a set of conditional independence assertions about variables in X, and a set P of local probability distributions associated with each variable. An edge from one node to another implies a direct dependency between them, with a child and parent relationship. To quantify the strength of relationships among the random variables, a conditional probability function P is associated with each node, such that $P = \{p(X1|\Pi1), ..., p(Xn|\Pi n)\}$, where Πi is the parent set of Xi in X. If there is a link from Xi to Xj, then Xi is a parent of Xj and thus it belongs to Πj . For discrete random variables the conditional probability functions are represented as tables, called Conditional Probability Tables (CPTs). For a typical node A, with parents $B1,B2, ...,Bn$, Knowledge Discovery Using Bayesian Network Framework there is associated a CPT given by $P(A|B1,B2, ...,Bn)$. For root nodes, the CPT reduces to prior probabilities. The main principle on which BN work, is Bayes' rule: $P(H|e) = P(e|H)P(H)\ P(e)$ (1)
where $P(H)$ is the prior belief about a hypothesis, $P(e|H)$ is the likelihood that evidence e results given H, and $P(H|e)$ is the posterior belief in the light of evidence e . This implies that belief concerning a given hypothesis is updated on observing some evidence.
Applying Bayesian network (BN) techniques to classification involves two sub-tasks: BN learning (training) to get a model and BN inference to classify instances(Cheng & Greiner, 2001).

3.2.2 Naïve bayes

Naïve bayes is classification technique based on the Bayesian theorem. This technique analyses the relationship between each attribute and the class for each instance to derive a conditional probability for the relationships between the attribute values and the class.

The principle behind Naïve bayes for classification is a fairly simple process. During training, the probability of each class is computed by counting how many times it occurs in the training dataset. This is called the "prior probability" P(C=c). In addition to the prior probability, the algorithm also computes the probability for the instance x given c with the assumption that the attributes are independent. This probability becomes the product of the probabilities of each single attribute. The probabilities can then be estimated from the frequencies of the instances in the training set. Numeric attributes can have a large number (possibly infinite) of values and the probability cannot be estimated from the frequency distribution, which tend to reduce the performance of Naïve bayes (Danso, 2006).

As it is shown on figure 3:2, Naive-Bayes Bayesian network is a simple structure that has the classification node as the parent node of all other nodes (see Figure).

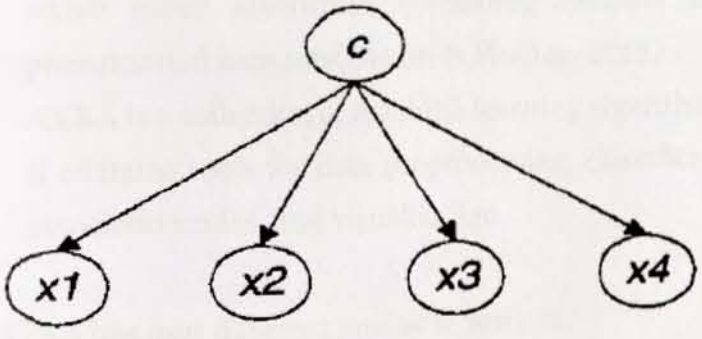


Figure 3:2 **A simple Naive Bayes structure**

Naive-Bayes has been used as an effective classifier for many years. It has two advantages over many other classifiers. First, it is easy to construct, as the structure is given a priori (and hence no structure learning procedure is required). Second, the classification process is very efficient. Both advantages are due to its assumption that all the features are independent of each other. Although this independence assumption is obviously problematic, Naive-Bayes has surprisingly outperformed many sophisticated classifiers over a large number of datasets, especially where the features are not strongly correlated (Cheng & Greiner, 1999).

Once the telecom fault data is collected, preprocessed and prepared, J48 decision tree algorithm and naïve bayes algorithm are used for tracking and creating predictive model. This model enables us to predict telecommunication service faults that may happen to Ethio telecom.

3.3 Selected tools

While selecting tools for this study, the researcher considered various aspects such as: functionality, ease of use and experience. Addressing these aspects led the researcher to conclude that Weka and MS-Excel are suitable for accomplishing the planned tasks for this particular study.

- **Weka**: - is an open-source Java application produced by the University of Waikato in New Zealand. This software bundle features an interface through which many algorithms (including decision trees) can be utilized on preformatted data sets(Drazin & Montag, 2012).
  WEKA is a collection of machine learning algorithms for Data Mining tasks.
  It contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization.

WEKA has four different modes to work in.

- Simple CLI; provides a simple command-line interface that allows direct execution of WEKA commands.
- Explorer; an environment for exploring data with WEKA.
- Experimenter; an environment for performing experiments and conduction of statistical tests between learning schemes.
- Knowledge Flow; presents a "data-flow" inspired interface to WEKA.

From the above mentioned working modes of WEKA, most of the tests made in this particular study were done by using the explorer mode.

- **Ms-Excel:**

Microsoft Excel is a spreadsheet application with different features such as calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications.

For this study Microsoft Excel was used for preprocessing and preparing the target dataset by applying different standard spreadsheet techniques and also programming with VBA was applied to handle additional tasks that were impossible with standard spreadsheet techniques. The VBA sample codes applied in the preprocessing step of this study are attached in the appendices.

CHAPTER FOUR

DATA PREPARATION & PREPROCESSING

Data preparation is a crucial step during data mining for enhancing the performance of the model created to predict telecom service fault.

4.1 Business and Data Understanding

On this section of the study the focus is to describe the knowledge gained through business and data understanding and also to explain the appropriate data mining problem definition identified with a preliminary plan design to achieve the objectives.

- **Broadband network service Fault handling at ethio-telecom**

As a sole telecom operator, ethio-telecom provides different types of telecommunication services. Broadband network service is one of the various types of services provided by ethio-telecom. The types of broad band network services are summarize below in table 4.1.

No.	Broadband Network Service Types	Description
1	BB-ADSL	For connections less than 4 mbps
2	EPON	For connections greater than 4 mbps
3	ADSL2+(PPPOE),	For shared ADSL connection (individuals)
4	BB-DDN	For data network connection (point to point)
5	BB-Fixed wireless access(FWA),	For data and internet wireless connection
6	Email-Domain Name-Web hosting	BB value added or additional services
7	BB-Dial Up	Internet connection using telephone network
8	BB-Methro	Internet and data fiber connection (direct)
9	Direct Fiber	Internet and data fiber connection (direct)
10	BB-AiroNet	Wireless connection for data + internet

Table 4:1 **Broadband technologies**

Upon providing the above mentioned types of services, the company accepts various service fault reports from its customers. For the purpose of handling such reports so as to follow the problem handling process, currently the company is using an information system called Z-Smart Trouble Ticket system.

Z-Smart Trouble Ticket is a system which has fault creation, fault follow up, fault closing or trouble ticket (TT) archiving functionalities. The company is using this system for handling TT preprocessing, dispatch, handling, confirmation and archiving tasks. It is also integrated with the automatic alarm system that is coming from the network elements. Trouble tickets are created both manually and automatically that are traced in the trouble ticket system.

By using this system, every time a customer reports a fault or service disruption, the assigned technician for registering and identifying fault types generate or create a trouble ticket for each and every fault reports. As it is indicated on table 4.2, the original data set generated from Z_Smart TT information system for this particular study, have twenty one different attributes which are related with the overall fault handling process.

Attribute Name	Data type	Description
Trouble Ticket No.	String	A unique number which identifies each fault report from customers.
Work Order No.	String	Sequential number given for a fault case till it gets the final solution.
Occurrence Date	Date	Fault happening date
Title	String	Subscriber name
BB Technologies	String	Type of service
Trouble Grade	String	The category of the subscriber
VIP Flag	Boolean	If the subscriber is among the VIP group
Warning	Date	Notification time before the TT time out.
Time-out	Date	The final time expired for finishing fault maintenance.
Current Handle Site	String	Department/section currently handling the fault/TT
Handler	String	Person or individual who handle the

Attribute Name	Data type	Description
		TT.
Check Out	String	If the TT is being handled, it will be in check out state and it is will be in check in state if the problem is solved.
Activity	String	Status of the fault handling process.
Order State	String	Whether a fault work order is being on process or completed.
State	String	Whether a sub task of the fault solving process is finished or not.
Work Order Type	String	Same as Activity attribute.
Completed	Boolean	Confirm whether problem is solved or not.
Work Order Start Date	Date	The time and date of the problem handling process started.
Completion Date	Date	The time and date of the problem solved.
Fault Type	String	The actual fault happened.
Object	Number	The service number of the faulty service.

Table 4:2 **Attribute names, data type and description**

After the fault reports are registered with all the necessary information, they will be assigned to the respective handler so that the problem can be further analyzed and get solution by the staffs in the handler sites. After solving the problem, the real cause of the problem which was the reason for the service fault to happen, will be registered in the system by field technicians.

In addition, the Trouble-tickets created in Z-Smart system and all the information registered after wards related with a specific Trouble-ticket helps for tracing the fault handling process and its progress. In addition it is also possible to know who is currently handling the case. It also tells how long the Trouble-ticket or fault delayed for the customer and with specific handler. Currently, there is no data mining technique

applied on the data registered and stored and no analysis is being made to deal with recurring fault types.

It is the researchers' believe that, after preprocessing, this data can be used for predicting the network (broad band) service faults by applying data mining techniques. To do so, among the different types of data mining tasks that can be applied on a large amount of data to extract useful knowledge, the researcher identified that classification / prediction is the one which can better suit the planned activity (predicting network faults). The researcher made this task selection, based on the business understanding gained through a discussion held with the domain experts.

4.2 Data Preparation and pre-processing

Data preparation and pre-processing, helps in obtaining quality data that leads to quality patterns by controlling incomplete, noisy and inconsistent real world data. Data preprocessing is consisting of techniques concerned with analyzing raw data such as data cleaning, data integration, data transformation, data reduction, and data discretization. Applying these techniques can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining (Han, et al., 2006).

The target data set used for this particular study was initially extracted from the Z-Smart Trouble Ticket information system, by selecting features which are related with Broadband network fault reports. Initially, around 27,900 records spanning over six months were selected.

Each event is described in terms of twenty one attributes. The attributes are of different data types, such as number, date and string. Some of the attributes of the initial data have no values in them; some records are redundant; some records contain wrong values. Since the original data was not configured for data mining usage, it was necessary to go through multiple iterations of data pre-processing to extract meaningful dataset for data mining.

Therefore most of the aforementioned preprocessing tasks are applied in the preprocessing task of this particular study.

Most of the data preprocessing tasks for this particular study were done by using Microsoft excel spreadsheets and Visual Basic for Applications (VBA) macros.

3.2.1 Data cleaning

Data cleaning is one of preprocessing tasks used to fill in missing values, smooth noisy data, identify outliers, and correct data inconsistencies. Therefore, data cleaning was applied for this particular study to fill the observed missing values and to remove noise and correct inconsistencies in the data.

Among the above mentioned 27,900 number of records, 13,299 instances were extracted as unique instances because the rest were recorded redundantly for following the progress of the fault solving process.

From the extracted 13,299 number of instances, 2191 number of instances had missing values on the class variable (fault type attribute) therefore they were deleted.

The possible 1,156 missing values from the remaining 11,108 number of instances were filled by discussing with the domain experts. Almost all of these missing values were on the attribute 'Title' which contains the name of the subscribers, and these values were filled using service numbers or the values in the 'object' column. These values in the object column uniquely identify each subscriber. By using these service numbers the domain experts identified the name of the missing subscribers.

In addition, some wrongly inserted values were also edited again with a discussion held with the domain experts.

3.2.2 Data integration and transformation

The target data set was initially extracted from the system differentiated in months and obtained on separate excel spread sheets. These data on separate spread sheets were merged and integrated into one spread sheet to make it ready for further preprocessing steps.

There were a number of attributes that needed some form of transformation. For example, the attribute "Title" contains diverse type of service subscriber names such as name of banks, insurances, embassies, hospitals, hotels, universities, government and non government organizations, internet cafes, individuals and others. Therefore, for the

purpose of managing and understanding the mining result, it was important to categorize the values to some groups. Therefore, this subscriber names were categorized into the following categories: -

- Bank or Insurance
- Business Centers
- Governmental Organizations
- Non Governmental Organizations & Embassies
- Education Centers and Others

The task of bringing the various subscriber names into the above listed categories was performed by applying excel VBA macro programming by using some common key words that can help in categorizing them. But the task of selecting a common key word or representative was a very difficult and time taking because the data was encoded by human beings and there were various types of representing same word or phrase.

For example it was observed that to represent a private limited company it was written as the following representations; plc, p.l.c, P.L.C, Plc, or using the whole phrase 'private limited company'. This type of difficulty was faced in most of the attributes that were tried to categorize.

The other attribute which was transformed was the 'BB Technologies' attributes which initially contains the name of various types of broadband network access technologies such as BB-ADSL, EPON, ADSL2+(PPPOE), BB-DDN, BB-Fixed wireless access(FWA), Email-Domain Name-Web hosting, BB-Dial Up, BB-Methro, Direct Fiber, BB-AiroNet. Therefore, these values of the attribute were generalized into three types of categories such as: -

- Wired (for copper media connections)
- Fiber and
- Wireless

The other attribute was the 'Work Order start date' which originally contains the date in date format including a time stamp like (2013-05-06 12:12:15). By discussing with the domain experts it was agreed to use only the month of this attribute values for this study

so that to make some analysis from the data mining results base on the seasons. Therefore, the month was extracted in text format by using both basic excel and VBA macro.

3.2.3 Data reduction

As previously discussed, the original data was consisting of twenty one attributes. Among those attributes there were unrelated attributes with the planned task of this study; therefore dimensionality reduction was applied on the data by removing the unrelated attributes such as [Work Order No],[Occurrence Date],[Warning], [Time-out], [Handler], [Check Out], [Activity], [Order State], [State], [Work Order Type], [Completed] and [Object] were deleted and the remaining attributes, [Title], [BB Technologies], [Trouble Grade], [VIP Flag], [Current Handle Site], [Work Order Start Date], and [Fault Type] were used for the planned task. The selected attributes are described with sample data in table 4.3 below.

No	Attribute Name	Sample Data	Description
1	Title	String	Subscriber name
2	BB Technologies	String	Type of service
3	Trouble Grade	string	The category of the subscriber
4	VIP Flag	Boolean	If the subscriber is among the VIP group
5	Current Handle Site	String	Department/section currently handling the fault/TT
6	Month	String	Date & time handler start work
7	Fault Type	String	The actual fault happened.

Table 4:3 List of selected attributes after preprocessing

Among the above selected attributes, since the aim of this study is to predict faults based on the previously occurred faults, the last attribute (Fault type) is selected as a target attribute based on a discussion held with the domain experts.

Basically this fault type attribute is an attribute which is filled by a technician who is assigned to handle the specific trouble ticket. After the technician accomplished his/ her mission of solving a specific fault, he/ she will put a remark about the real cause of the problem in a phrase or sentence form. Therefore, this attribute was labeled and given common representation by working together with the domain experts.

3.2.4 Data formatting

Finally the target data set was prepared in .CSV format so as to make it readable for the selected data mining tool (WEKA).

CHAPTER FIVE

EXPERIMENTAION

The main objective of this study is, discovering patterns for predicting network fault types within the Ethio telecom Broadband service related fault report information registered in information system data base.

Towards achieving this objective, the model building phase in the DM process of this investigation is carried out through applying the supervised classification techniques. These techniques are implemented using Weka data mining tool.

5.1 Experimentation setup

A procedure or mechanism of how to test the model's quality and validity is needed to be set before the model is actually built. In order to perform the model building process of this study, out of the total 27,956 original instances after preprocessing the original data 11,108 instances were obtained for training the classification models. In addition , from the total number of 21 attributes, after preprocessing only the 7 attributes are used for this specific study.

For validating the accuracy of the classification model, from the above mentioned number of instances Weka managed the test data set based on the test options selected for model creation.

The experimentation of this particular study was performed with the following setup: -

- J48 decision tree algorithm with 'test training set' test option
- J48 decision tree algorithm with 'cross validation -10 folds' test option
- J48 decision tree algorithm with 'percentage split-66%' test option
- J48 decision tree algorithm with various confidence factors.
- J48 decision tree algorithm with SMOTE filtering technique.
- Naïve bayes with the above mentioned three test options.

5.2 Classification Modeling

For starting the classification modeling experiments, the decision tree (in particular the J48 algorithm) and the naïve bayes methods are selected.

The training of the decision tree classification models of the experimentation is done by employing the 10-fold cross validation and the percentage split classification modes. The classification is analyzed to measure the accuracy of the classifiers in categorizing the instances into specified classes. Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications. The classification accuracy of each of these models is reported and their performance is compared in classifying new instances of records.

5.2.1 Experimentation, using J48 decision tree classification algorithms

The classification model building experimentation was done by applying J48 decision tree algorithm and naïve bayes classification algorithms based on three test options available on the WEKA data mining tool. The selected test options are 'Use training set', 'Cross-Validation (10 folds)', and 'Percentage split 66%'. These test options were tested using the default parameters and also by setting different additional parameters.

Experimentation 1: -Applying J48 decision tree algorithm with 'Use training set' test option and default parameters.

On the first experimentation, the classification model is trained by using the J48 decision tree algorithm with the 'use training set' test option of WEKA. Using the default parameters the classification model developed was having a tree with 165 numbers of leaves and 209 tree sizes. The decision tree used the total seven attributes for generating the tree.

As it can be seen from the confusion matrix presented in table 5.1 below, the developed J48 decision tree classifier classified 8227 instances (74.06%) out of the total 11,108 instances correctly and 2881 instances (25.94%) were classified incorrectly. The time taken for developing the model was 0.2 seconds.

Exp. No.	Algorithm	No. of Instance	No. of Attributes	Test option	Parameter	Time Taken	Accuracy
1	J48	11,108	7	Use training set	Default	0.2	74.06 %

Table 5: 1 Accuracy result using J48 algorithm, use training set test option

Based on the information gained from the confusion matrix, the classifier developed using the J48 decision tree algorithm with 'use training set' test option and default values, classified 3096 number of instances from the total 11,108 number of instances as 'CPE change or configuration & port' fault type.

As it can be seen from table 5.2 below, among the total instances, 2177 number of instances (70%) were correctly classified as 'CPE change or configuration & port' fault type and 917 (30%) were incorrectly classified as 'CPE change or configuration & port' fault type.

Among the total number of instances, 960 instances were classified as 'CNR or Core network problem' fault type but among these instances, 494 (51%) were classified correctly as 'CNR or Core network problem' and the rest 466 (49%) were classified incorrectly. Again from the total number of instances, 4705 instances were classified as 'Line / Fiber Problem' fault type but among these instances 4356 (93%) were classified correctly as 'Line / Fiber Problem' fault type but the rest 349 (7%) were classified as 'Line / Fiber Problem' fault type incorrectly.

The total number of instances that were classified as 'MSAG power or Transmission problem' fault type were 1844 but from this number of instances 883 (48%) were classified correctly and 961 (52%) were classified incorrectly as 'MSAG power or Transmission problem' fault type.

In addition, 332 numbers of instances were classified as 'Customer LAN or Router' type of fault but among these instances 193 (58%) were classified correctly and the rest 139 (42%) were classified incorrectly. From the 171 total number of instances classified as 'Email, web hosting & Domain name problem' fault type, 124 (73%) were classified

correctly and the rest 47 (27%) were classified incorrectly as 'Email, web hosting & Domain name problem' fault type.

Actual	Predicted						Total	Accuracy
	CPE change or configuration & port	CNR or Core network problem	Line / Fiber Problem	MSAG power or Transmission problem	Customer LAN or Router	Email, web hosting & Domain name problem		
CPE change or configuration & port	2177	166	665	82	6	0	3096	70%
CNR or Core network problem	94	494	301	69	2	0	960	51%
Line / Fiber Problem	56	70	4356	201	21	1	4705	93%
MSAG power or Transmission problem	401	35	470	883	47	8	1844	48%
Customer LAN or Router	2	4	43	90	193	0	332	58%
Email, web hosting & Domain name problem	0	3	34	8	2	124	171	73%

Table 5: 2 **confusion matrix results for experimentation 1**

From this particular experimentation, it can be observed on table 5.2 that the classifier is more confusing the 'MSAG power or Transmission problem' class with the 'Line / Fiber Problem' class.

Experimentation 2: -Applying J48 decision tree algorithm with 'Cross-Validation (10 folds)' test option and default parameters.

On this experimentation, the classification model is trained by using the J48 decision tree algorithm with the 'Cross-Validation (10 folds)' test option of WEKA. Using the default parameters the classification model developed was having a tree with 165 numbers of leaves and 209 tree sizes. The decision tree used the total seven attributes for generating the tree.

As it is shown on table 5.3 below, the developed J48 decision tree classifier correctly classified 8072 instances out of the total 11,108 instances that are 72.67 % and incorrectly classified 2863 instances that is 27.33 %. The time taken for developing the model was 0.04 seconds.

Exp. No.	Algorithm	No. of Instance	No. of Attributes	Test option	Parameter	Time Taken	Accuracy
2	J48	11,108	7	Cross Validation (10 folds)	Default p0.04	0.04	72.67 %

Table 5: 3 Accuracy result using J48 algorithm, use cross validation (10folds) test option and default parameter

Based on the information gained from the confusion matrix, the classifier developed using the J48 decision tree algorithm with 'Cross-Validation (10 folds)" test option and default values, classified 3096 number of instances from the total 11,108 number of instances as 'CPE change or configuration & port' fault type. As it can be seen from table 5.4 below, among these instances, 2162 number of instances (70%) were correctly classified as 'CPE change or configuration & port' fault type and 934 (30%) were incorrectly classified as 'CPE change or configuration & port' fault type.

Among the total number of instances, 960 instances were classified as 'CNR or Core network problem' fault type but among these instances, 464 (48%) were classified correctly as 'CNR or Core network problem' and the rest 496 (52%) were classified incorrectly. Again from the total number of instances, 4705 instances were classified as 'Line / Fiber Problem' fault type but among these instances 4258 (90%) were classified correctly as 'Line / Fiber Problem' fault type but the rest 447 (10%) were classified as 'Line / Fiber Problem' fault type incorrectly.

The total number of instances that were classified as 'MSAG power or Transmission problem' fault type were 1844 but from this number of instances 883 (48%) were classified correctly and 961 (52%) were classified incorrectly as 'MSAG power or Transmission problem' fault type.

In addition, 332 numbers of instances were classified as 'Customer LAN or Router' type of fault but among these instances 183 (55%) were classified correctly and the rest 149 (45%) were classified incorrectly. From the 171 total number of instances classified as 'Email, web hosting & Domain name problem' fault type, 122 (71%) were classified correctly and the rest 49 (29%) were classified incorrectly as 'Email, web hosting & Domain name problem' fault type.

Actual	Predicted						Total	Accuracy
	CPE change or configuration & port	CNR or Core network problem	Line / Fiber Problem	MSAG power or Transmission problem	Customer LAN or Router	Email, web hosting & Domain name problem		
CPE change or configuration & port	2162	173	655	100	6	0	3096	70%
CNR or Core network problem	114	464	304	77	1	0	960	48%
Line / Fiber Problem	97	68	4258	259	22	1	4705	90%
MSAG power or Transmission problem	412	35	456	883	49	9	1844	48%
Customer LAN or Router	6	2	42	99	183	0	332	55%
Email, web hosting & Domain name problem	3	2	34	8	2	122	171	71%

Table 5: 4 confusion matrix for experimentation 2

From this particular experimentation, it can be observed on table 5.4 that the classifier is more confusing the 'MSAG power or Transmission problem' class with the 'Line / Fiber Problem' class.

As the detail accuracy by class indicated in table 5.5, the best precision accuracy is obtained in the 'Email, web hosting & Domain name problem' class which is 0.924, followed by 'CPE change or configuration & port' class which is 0.774. Regarding recall

and F-measure, the best result is obtained by 'Line / Fiber Problem' which is 0.905 and 0.815 respectively.

Class	Precision	Recall	F-Measure
CPE change or configuration & port	0.774	0.698	0.734
CNR or Core network problem	0.624	0.545	0.483
Line / Fiber Problem	0.741	0.905	0.815
MSAG power or Transmission problem	0.619	0.54	0.479
Customer LAN or Router	0.696	0.551	0.615
Email, web hosting & Domain name problem	0.924	0.713	0.805

Table 5: 5 Detail accuracy by class

From the above information, it can be said that, precision is high for the class with less instances and recall is high for the class with more instances which is the case observed from the class 'Email, web hosting & Domain name problem' and 'Line / Fiber Problem' and .

Experimentation 3: -Applying J48 decision tree algorithm with 'Percentage split 66%' test option and default parameters.

On this experimentation, the classification model is trained by using the J48 decision tree algorithm with the 'Percentage Split' test option of WEKA. Using the default parameters the classification model developed was having a tree with 165 numbers of leaves and 209 tree sizes. The decision tree used the total seven attributes for generating the tree.

As it can be seen on table 5.6 below, the developed J48 decision tree classifier correctly classified 2745 instances out of the test set instances given by WEKA that was 72.68 % and incorrectly classified 1032 instances that was 27.32 %. The time taken for developing the model was 0.03 seconds.

Exp. No.	Algorithm	No. of Instance	No. of Attributes	Test option	Parameter	Time Taken	Accuracy
3	J48	11,108	7	Percentage split 66%	Default	0.03	72.68 %

Table 5: 6 Accuracy result using J48 algorithm, percentage split test option with default parameter.

Based on the information gained from the confusion matrix, the classifier developed using the J48 decision tree algorithm with 'Percentage Split' test option and default values, classified 1018 number of instances from the total 11,108 number of instances as 'CPE change or configuration & port' fault type. Among these instances, 750 number of instances (74%) were correctly classified as 'CPE change or configuration & port' fault type and 268 (26%) were incorrectly classified as 'CPE change or configuration & port' fault type.

Among the total number of instances, 284 instances were classified as 'CNR or Core network problem' fault type but among these instances, 139 (49%) were classified correctly as 'CNR or Core network problem' and the rest 145 (51%) were classified incorrectly. Again from the total number of instances, 1606 instances were classified as 'Line / Fiber Problem' fault type but among these instances 1432 (89%) were classified correctly as 'Line / Fiber Problem' fault type but the rest 174 (11%) were classified as 'Line / Fiber Problem' fault type incorrectly.

The total number of instances that were classified as 'MSAG power or Transmission problem' fault type were 689 but from this number of instances 314 (46%) were classified correctly and 375 (54%) were classified incorrectly as 'MSAG power or Transmission problem' fault type.

In addition, 118 numbers of instances were classified as 'Customer LAN or Router' type of fault but among these instances 66 (56%) were classified correctly and the rest 52 (44%) were classified incorrectly. From the 62 total number of instances classified as 'Email, web hosting & Domain name problem' fault type, 44 (71%) were classified

correctly and the rest 18 (29%) were classified incorrectly as 'Email, web hosting & Domain name problem' fault type.

Actual	Predicted						Total	Accuracy
	CPE change or configuration & port	CNR or Core network problem	Line / Fiber Problem	MSAG power or Transmission problem	Customer LAN or Router	Email, web hosting & Domain name problem		
CPE change or configuration & port	750	52	179	35	2	0	1018	74%
CNR or Core network problem	33	139	87	25	0	0	284	49%
Line / Fiber Problem	49	38	1432	74	9	4	1606	89%
MSAG power or Transmission problem	178	20	153	314	22	2	689	46%
Customer LAN or Router	3	0	13	36	66	0	118	56%
Email, web hosting & Domain name problem	0	2	14	1	1	44	62	71%

Table 5: 7 confusion matrix for experimentation 3

From this particular experimentation, it can be observed on table 5.7 that the classifier is more confusing the 'MSAG power or Transmission problem' class with the 'Line / Fiber Problem' class. It is observed that most of the values in the corresponding two

attribute ('BB Technology' and 'Month') values for this two fault types is similar. This can cause the classifier to confuse the fault types from one to the other.

Experimentation 4,5,6,7: using J48 decision tree algorithms with various folds of cross validation test option

The accuracy result of the additional experimentations made using J48 decision tree algorithm by changing the default parameter of cross validation test option in table 5.8 below.

Exp. No.	No. of Attributes	Algorithm	Cross validation fold values	Time	Accuracy
4	7	J48	5	0.24	72.69%
5			15	0.07	72.52%
6			20	0.04	72.69%
7			25	0.04	72.78%

Table 5: 8 summary of accuracy results for experiments made using J48 algorithm with various fold values of cross validation test option.

As it can be seen from table 5.8 above, the results obtained through applying different fold values of the cross validation test option, were not predictable because no consistent pattern of increasing or decreasing cannot be observed. But a better result is observed at the fold value of 25 which results in accuracy of 72.78%.

Experimentation 8, 9, 10, 11, 12: Applying various percentage of SMOTE filtering technique on J48 decision tree algorithm with cross validation test option

As it can be seen from figure below, through visualizing the number of instances for each of the class variables, it is observed that the classes are not approximately equally represented.

Selected attribute

Name: Fault Type
Missing: 0 (0%)

Distinct: 6

Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	CPE change or configuration & port	3096	3096.0
2	CNR or Core network problem	960	960.0
3	Line / Fiber Problem	4705	4705.0
4	MSAG power or Transmission problem	1844	1844.0
5	Customer LAN or Router	332	332.0
6	Email, web hosting & Domain name pr...	171	171.0

Class: Fault Type (Nom)

Visualize All

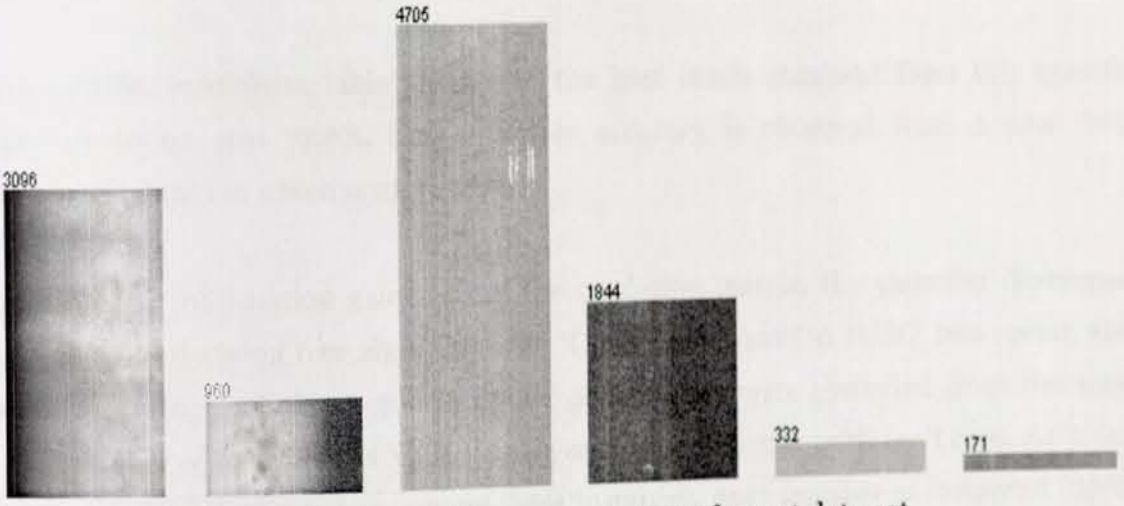


Figure 5: 1 Class variable representation of target data set.

Therefore, to address this dataset imbalance problem, on this experimentation section, Synthetic Minority Over-sampling Technique (SMOTE) was applied. This technique is an over-sampling approach in which the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement.

There are approximately 4705 instances in the majority class (Line /Fiber problem) and 171, 332, 960, 1844 and 3096 instances in the minority classes for the training set used in 10-fold cross-validation. To get a balanced number of instances for each class variables SMOTE over sampling technique is applied on the minority classes in different

percentage. As it can be seen on table 5.9 below the minority classes were over-sampled at 300% of their original size.

Exp. No.	Algorithm	Test Option	Minority classes	Applied SMOTE Percentage	New Accuracy
8	J48	10-Fold Cross Validation	Email,web hosting & Domain name problem	300%	72.46%
9			Customer LAN or Router	300%	73.18%
10			Email,web hosting & Domain name problem	300%	74.85%
11			CNR or core network problem	300%	74.60%
12			Customer LAN or Router	300%	77.9%

Table 5: 9 summary of experimentation with SMOTE filtering technique.

As it can be seen from table 5.9 above, the best result obtained from this specific experimentation was 77.9%. This classifier accuracy is obtained from a new total number of instances which is 21,533.

Based on the information gained from the confusion matrix, the classifier developed using the J48 decision tree algorithm with 'Cross-Validation (10 folds)' test option and SMOTE filtering technique, 3096 number of instances were classified from the total 21,533number of instances as 'CPE change or configuration & port' fault type. As it can be seen from table 5.10 below, among these instances, 2147 number of instances (69%) were correctly classified as 'CPE change or configuration & port' fault type and 949 (31%) were incorrectly classified as 'CPE change or configuration & port' fault type. Among the total number of instances, 3840 instances were classified as 'CNR or Core network problem' fault type but among these instances, 2963 (77%) were classified correctly as 'CNR or Core network problem' and the rest 877 (23%) were classified incorrectly. Again from the total number of instances, 4705 instances were classified as 'Line / Fiber Problem' fault type but among these instances 3589 (76%) were classified

correctly as 'Line / Fiber Problem' fault type but the rest 1116 (24%) were classified as 'Line / Fiber Problem' fault type incorrectly.

The total number of instances that were classified as 'MSAG power or Transmission problem' fault type were 1844 but from this number of instances 542 (29%) were classified correctly and 1302 (71%) were classified incorrectly as 'MSAG power or Transmission problem' fault type.

In addition, 5312 numbers of instances were classified as 'Customer LAN or Router' type of fault but among these instances 5200 (98%) were classified correctly and the rest 112 (2%) were classified incorrectly. From the 2736 total number of instances classified as 'Email, web hosting & Domain name problem' fault type, 2338 (85%) were classified correctly and the rest 398 (15%) were classified incorrectly as 'Email, web hosting & Domain name problem' fault type.

Generally, the result of the experimentation with SMOTE filtering technique shows that oversampling the minority class can improve the accuracy of the classifier.

Experimentation 13, 14, 15, 16, 17: using J48 decision tree algorithms by varying the value of confidence factor.

As it can be observed from the previous experimentations, the number of leaves and tree size is big and needs to be pruned so as to make the tree manageable. Therefore this experimentation considered various confidence factors so as to decrease the tree size. These experimentations were made using the J48 decision tree algorithm 'cross validation (10 fold)' test option and by varying the value of the confidence matrix of the classifier algorithm. The results obtained are summarized on table 5.10 below.

Exp. No.	No. of Attributes	Algorithm	Test Option	Confidence Factor	No of Leaves	Tree size	Accuracy
13				0.3	171	213	72.87

14	7	J48	Cross Validation 10 folds	0.4	190	239	72.78
15				0.02	77	99	72.29
16				0.01	73	94	72.21
17				0.001	57	74	71.99

Table 5: 10 Result summary of experimentation using J48 decision tree algorithm and different values of confidence matrix.

As it can be seen from the above table 5.10, as we increase the value of confidence matrix, the accuracy value will also increase and the number of leaves and tree size will also increase. On the other hand as the values of the confidence factor decreases, both the number of leaves and tree size will be decreased but the classifier accuracy will decrease.

Summary of best results of experimentations done using J48 decision tree algorithm

Algorithm	No. of instances	Test options	Special approach	Accuracy
J48	11,108	10 Fold Cross validation	Default	74.06%
	21,533	10 Fold Cross validation	SMOTE filtering technique (300%)	77.90%

Table 5: 11 Best accuracy result obtained using J48 algorithm, before and after applying SMOTE.

As it can be observed from table 5.11 above, from the experimentation done using the J48 decision tree algorithm with 10 fold cross validation, the best result is obtained by using SMOTE filtering technique which is 77.90% and with the accuracy obtained with the original data set was 72.14%.

The result obtained with applying SMOTE minority over sampling technique implied that if all the six class variables can get enough number of instances, a better accuracy result can be obtained.

To confirm this idea, in addition to applying SMOTE, for the sake of curiosity the researcher also took the instances of the two class variables with the large number of instances and applied J48 decision tree algorithm and 10 fold cross validation and got a classifier accuracy of 88.4%.

5.2.2 Experimentation using naïve bayes classification algorithm

Experimentation 13: Applying naïve bayes classification algorithm with 'use training set', 'cross validation (10 fold)' and 'percentage split (66%)' test options

The naïve bayes classification model building experimentation was also done by applying three of the available four test options on WEKA data mining tool. The selected test options are 'Use training set', 'Cross-Validation', and 'Percentage split'. The results are summarized in table 5:12 below.

Experimentation No.	Algorithm	No. of Attributes	Test options	Parameters	Time Taken In seconds	Accuracy
1	naïve bayes	7	Use training set	Default	0.05	69.38 %
			Cross validation	Default	0.02	69.39%
			Percentage split	Default	0.01	69.39%

Table 5: 12 summary of the experimentations made using naïve bayes classification algorithm based on the three test options.

As it can be seen from table above, the classifier accuracy obtained using all the three test options are almost equal which is approximately 69%.

Experimentation 14: Applying naïve bayes classification algorithm with SMOTE minority over sampling technique

The naïve bayes classification model building experimentation was also done by applying SOMTE minority over sampling technique and three of the available four test options on WEKA data mining tool. The selected test options are 'Use training set', 'Cross-Validation', and 'Percentage split'. The results are summarized in table 5:13 below.

Experimentation No.	Algorithm	No. of Attributes	Test options	SMOTE Filtering technique	Time Taken In seconds	Accuracy
1	naïve bayes	7	Use training set	300%	0.05	67.01 %
			Cross validation	300%	0.01	69.08%
			Percentage split	300%	0.01	66.14%

Table 5: 13 summary of the experimentations made using naïve bayes classification algorithm with SMOTE minority over sampling technique.

As it can be seen from table 5:13, the classifier accuracy obtained from the experimentation made with naïve bayes classification algorithm on a data with SMOTE minority over sampling technique, was not improved.

Comparison and Discussion of accuracy results obtained with both algorithms

As it has been discussed in the previous experimentation sections, the target data set has been tested in two classification algorithms (J48 decision tree and naïve bayes) using different WEKA test options and varying values of classifier parameters.

The results obtained showed that, the classifier accuracy obtained with J48 decision tree algorithm out performs, in all the three test options, the classifier accuracy obtained with naïve bayes classification algorithm.

In addition, applying SMOTE minority over sampling technique improved the classifier accuracy on the case of the J48 decision tree algorithm where as on same condition the classifier accuracy was decreasing for the case of naïve bayes classification algorithm. Therefore, as the above experimentation shows, from the two classification algorithms applied in this particular study, it can be said that J48 decision tree algorithm is better suitable for predicting the target data set.

5.2.3 Best rules from the experimentation made using the J48 decision tree algorithm

On this section, selected 13 best rules are presented with a discussion of the rules by grouping the best rules based on current handling site and also by the fault occurrence month. It is understood that the list is not exhaustive but only the ones believed to be presented are listed. These best rules were selected together with the domain experts.

Rule1: If Current Handle Site = FAN Addis: Line / Fiber Problem (1772.0/1.0)

Rule2: If Current Handle Site = NNOC and Month = April and BB Technologies = Wired and Trouble Grade = Critical: Customer LAN or Router (369.0/50.0)

Rule 3: If Current Handle Site = O&M Addis and Month = August and Trouble Grade = Critical and BB Technologies = Wired: CPE change or configuration & port (100.0/1.0)

Rule 4: If Current Handle Site = O&M Addis and Month = August and Trouble Grade = Critical and BB Technologies = Wireless: CNR or Core network problem (41.0/1.0)

Rule 5: If Current Handle Site = O&M Addis and Month = August and Trouble Grade = Normal: CNR or Core network problem (95.0/13.0)

Rule 6: If Current Handle Site = O&M Addis and Month = July and VIP Flag = NO: CNR or Core network problem (731.0/32.0)

Rule 7: If Current Handle Site = FAN Region and BB Technologies = Wired: Line / Fiber Problem (1208.0/25.0)

Rule 8: If Current Handle Site = NNOC and Month = April and BB Technologies = Wired and Trouble Grade = Normal: Customer LAN or Router (153.0/9.0)

Rule 9: If Current Handle Site = NNOC and Month = April and BB Technologies = Firber: Line / Fiber Problem (88.0/3.0)

Rule 10: If Current Handle Site = O&M Addis and Month = July and VIP Flag = YES

and Trouble Grade = Escalated and Title = Business Center: CNR or Core network problem (52.0/7.0)

Rule 11: If Current Handle Site = NNOC and Month = March and Title = Business Center and BB Technologies = Wired and Trouble Grade = Critical: MSAG power or Transmission problem (13.83/1.0)

Rule 12: If Current Handle Site = NNOC and Month = March and Title = Bank or Insurance and BB Technologies = Wired: MSAG power or Transmission problem (78.62/0.12)

Rule 13: If Current Handle Site = NNOC and Month = August and BB Technologies = Wireless and Trouble Grade = Critical: MSAG power or Transmission problem (114.0/3.0)

Among the above listed best rules, Rule 1 and 7 shows that, most of the faults handled by FAN addis current handle site were Line / Fiber.

The other information indicated by the generated rule is that in the rainy seasons the most frequently happening broadband network fault is 'CNR or Core network problem' this can be seen from rules 4, 5, 6 and 10. According to the domain experts, in the rainy season the central network resource transmission media, be it wireless, wired or fiber, can easily be affected by the weather. Especially, the wireless network is highly affected by the CNR since there is high interaction with it, including the DHCP server. Therefore, there is a need to give due attention for such type of faults during this period of the year.

Rule 2 & 8 indicated that most of the network fault types handled by NNOC current handle site are 'Customer LAN or Router'. This also indicates that such faults should not come to the NNOC and there is a need to take measure so that the national network operation center should spend its time on critical fault types. Such faults should be solved at first line support and this also shortens the time to fix the fault.

In general it is the domain experts' belief that the generated rules can be used by the company for supporting the fault handling process, and this can happen by integrating the rules with the Z-Smart TT system.

5.3 Model Evaluation

Evaluation in a data mining process serves two purposes, these are: - the prediction of how well the final model will work in the future (or even whether it should be used at all), and as an integral part of many learning methods, which help find the model that best represents the training data (Souza, Matwin, & Japkowicz, 2002).
The classifier's evaluation is most often based on prediction accuracy (the percentage of correct prediction divided by the total number of predictions)(Kotsiantis, et al., 2007).
In order to determine to what extent the classification model meets the business objectives, the classifiers accuracy was evaluated and based on the obtained accuracy measures comparisons were made. In addition a discussion with the domain experts was held regarding the usefulness of the classifier based on the obtained classifier performances. The domain experts confirmed that the developed classifier can be used by the company for predicting broad band network types.

5.4 Deployment

Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the user, not the data analyst, who will carry out the deployment steps. In any case, it is important to understand up front what actions will need to be carried out in order to actually make use of the created models(Wirth & Hipp, 2000).

As it was previously mentioned, the result of the study was discussed with the domain experts and it was suggested that since the existing information system (Z Smart TT) has a feature that can allow incorporating the obtained rules, it is possible to use the result of the study for the company.

Based on the results obtained in this experimentation section, the conclusion of this study and recommendations for future studies are presented on the coming chapter (chapter 6).

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1 Conclusion

In this research an attempt has been made to study the application of data mining techniques specifically classification/prediction on the data set which is related with broad band service fault customer report information obtained from ethio-telecom. The collected data was iteratively preprocessed so as to make the data ready for the data mining tool and to get the desired output from it.

Among the various types of classification algorithms J48 decision tree algorithm and naïve bayes classification algorithms were applied for this particular study. The classifier accuracy obtained by J48 decision tree algorithm was 74.06% with the preprocessed data set and after applying SMOTE minority over sampling technique the accuracy was improved to 77.90%.

It was observed that J48 decision tree algorithm is the best classification algorithm of the two algorithms used for this study while tested on unseen instances. The accuracy result of the classifier obtained by naïve bayes classification algorithm was around 69%.

The rules obtained from the decision tree indicated that on the rainy seasons the most frequently happening fault is 'CNR or Core network problem'. In addition it was observed that most of the faults are Line / Fiber related faults which are being handled by 'FAN addis current handle site'.

As it was clearly seen in the result of experimentations, the confusion matrix obtained with the target data set indicated that the dominant fault type is Line/fiber problem with 42.35 % of the total faults. The company can take this dominant fault type and take proactive action to improve its customer satisfaction. If the company manages to maintain these faults proactively, it can minimize 42.35 % percent of the fault or customer complaints. As recommended by the domain experts, this proactive measure

includes maintaining the old lines or fibers before they become the cause for service failure.

Some of the rules obtained from this research can be used in Z-smart system so as to enhance the fault handling process and improve customer satisfaction. Ethio telecom can use the rules to improve the fault assignment process so as to assign particular fault problem for the relevant handler site or technician.

This research also identified fault kinds faced by different customer segments like critical and major customers. This also helps the company what to do for specific customer groups based on the fault type they are facing and the sensitivity of the service failure or the customer. Since all service types are not equally sensitive for all customers or customer segments. For instance, internet services are critical for flower farm plc where as data services are critical for financial institutions.

The findings from the rules generated by the algorithms showed that specific customers groups like banks or insurance and business centers are affected by MSAG power or transmission problem/ fault kind. This fault kind is true for both critical and major customers of the company. As domain experts commented, MSAG power problem can simply be resolved by tapping generator power from nearby customers/mobile transmission stations. Therefore, due consideration should be given for this fault type, since those customers are important for the country in general and for the company in particular.

Generally, based on the evaluation made on the accuracy results and testing the results with the domain experts, even if the result of the classifier is promising, it was understandable that it needs an improvement by considering different factors. Some of these factors are mentioned as a recommendation below.

6.2 Recommendation

The predictive model, which is developed in this research, generates various promising patterns and rules. But the obtained classifier accuracy needs to be improved. Therefore, based on the findings of the study, the researcher made the following recommendations which can possibly help in improving the accuracy results of the predictive models.

- The target data set used for this study has six class variables. But, the instances for each of these class variables were not balanced. It was observed that this imbalance problem affected the classifier accuracy. Therefore in future studies, increasing the number of instances and having a balanced data set for each class variables can improve the classifier accuracy.

- The target data set used for this study was not obtained directly from the system data base rather it was generated through the Z-Smart TT information system. But according to the domain experts there are other additional attributes that can be obtained from the data base which are related with the target data set. Therefore by adding more attributes such as 'Area', 'Resource Type' and other selected attributes a test can be made to improve the accuracy.

- Some attributes in the original data set, applied for this study contains very diverse values. An attempt was made to bring the values into some common categories and representations. But still it needs a serious consideration because, since the attributes values were encoded by human beings, there are many factors (spelling errors, upper/lower cases, abbreviations, etc..) to be considered in picking the key word that represent the specific value.

- The target data set used for this particular study can be tested using other data mining tools to check for better classifier accuracy.

- And also the target data set used for this particular study can be tested using various types of classification algorithms.

- Integrating the best rules with the existing information system should be given due consideration.

Refferences

Aghasaryan, A., Dousson, C., Fabre, E., Pencolé, Y., & Osmani, A. (2002). *Modeling fault propagation in telecommunications networks for diagnosis purposes.* Paper presented at the XVIII World Telecommunications Congress.

Alarcón Mondéjar, M. J., Zorzano Mier, F. J., Jevtić, A., & Andina de la Fuente, D. (2008). Telecommunications Network Planning and Maintenance.

Ana Azevedo, M. F. S. (2008). *KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW* Paper presented at the IADIS European Conference Data Mining.

Anyanwu, M. N., & Shiva, S. G. (2009). Comparative analysis of serial decision tree classification algorithms. *International Journal of Computer Science and Security, 3*(3), 230-240.

Bashar, A., Parr, G., McClean, S., Scotney, B., & Nauck, D. (2010). Knowledge discovery using Bayesian network framework for intelligent telecommunication network management *Knowledge Science, Engineering and Management* (pp. 518-529): Springer.

Benjamini, Y., & Leshno, M. (2005). Statistical methods for data mining *Data Mining and Knowledge Discovery Handbook* (pp. 565-587): Springer.

Cayci, A., Eibe, S., Menasalvas, E., & Saygin, Y. Bayesian Networks to Predict Data Mining Algorithm Behavior in Ubiquitous Environments. *Mining Ubiquitous and Social Environments (MUSE 2010)*, 23.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Cheng, J., & Greiner, R. (1999). *Comparing Bayesian network classifiers.* Paper presented at the Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence.

Cheng, J., & Greiner, R. (2001). Learning bayesian belief network classifiers: Algorithms and system *Advances in Artificial Intelligence* (pp. 141-151): Springer.

Costea, A. (2006). The Analysis of the Telecommunications Sector by The Means of Data Mining Techniques. *Journal of Applied Quantitative Methods*, 144.

Danso, S. O. (2006). *An Exploration of Classification Prediction Techniques in Data Mininig: The Insurance Domain.* Masters Degree in Advanced Software Engineering, Bournemouth.

De Ville, B. (2006). *Decision trees for business intelligence and data mining: using SAS enterprise miner*: SAS Institute.

Drazin, S., & Montag, M. (2012). Decision Tree Analysis using Weka. *Machine Learning-Project II, University of Miami*, 1-3.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39*(11), 27-34.

Frank, E. (2000). *Pruning Decision Trees and Lists.* Waikato, Hamilton, New Zealand.

Gebremeskel, G. (2006). *Data Mining Application in Supporting Fraud Detection on Ethio-Mobile Service.* Ph.D.Dissertation, Addis Ababa University.

H¨at¨onen, K. (2009). *Data mining for telecommunications network log analysis Kimmo H¨at¨onen.* phd Series of Publications, University of Helsinki Finland.

Han, J., Kamber, M., & Pei, J. (2006). *Data mining: concepts and techniques*: Morgan kaufmann.

Hand, D. J., Mannila, H., & Smyth, P. (2001). Principles of data mining (adaptive computation and machine learning): The MIT Press.

Hudyma, R., & Fels, D. I. (2004). Causes of failure in IT telecommunications networks. *Proceedings of SCI, Florida*, 35-38.

Jackson, J. (2002). Data mining: a conceptual overview. *Communications of the Association for Information Systems, 8*, 267-296.

Jaudet, M., Iqbal, N., Hussain, A., & Sharif, K. (2005). *Temporal Classification for Fault-prediction in a real-world Telecommunications Network.* Paper presented at the Emerging Technologies, 2005. Proceedings of the IEEE Symposium on.

Kantardzic, M. (2003). *Data Mining: Concepts,Models, Methods, and Algorithms* (Vol. 1).

Katz, R. (2011). The impact of broadband on the economy: Research to date and policy issues. *Trends in Telecommunication reform 2010, 11*, 23-82.

Klemettinen, M., Mannila, H., & Toivonen, H. (1999). Rule discovery in telecommunication alarm data. *Journal of Network and Systems Management, 7*(4), 395-423.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

M.weiss, G. (2009). Data Mining in the Telecommunications industry

Maimon, O., & Rokach, L. (2005). Introduction to knowledge discovery in databases *Data Mining and Knowledge Discovery Handbook* (pp. 1-17): Springer.

Medhi, D., & Tipper, D. (1997). Towards fault recovery and management in communication networks. *Journal of Network and Systems Management, 5*(2), 101-103.

Medved, D., Brlas, K., & Saric, D. (2000). *Fault analysis and prediction in telecommunication access network.* Paper presented at the Electrotechnical Conference, 2000. MELECON 2000. 10th Mediterranean.

patel, j. e. PERFORMANCE EVALUATION OF DECISION TREE CLASSIFIERS FOR RANKED FEATURES OF INTRUSION DETECTION.

Pujari, A. K. (2001). *Data mining techniques*: Universities press.

Pyle, D. (1999). *Data preparation for data mining* (Vol. 1): Morgan Kaufmann.

Rayudu, R. K., & Maheswaraj, A. Data Mining Fault Information Reports for Prediction.

Reddy, G. S., Srinivasu, R., Rao, M. P. C., & Rikkula, S. R. (2010). Data Warehousing, Data Mining, OLAP and OLTP Technologies are essential elements to support decision-making process in industries. *International Journal on Computer Science and Engineering, 2*(09), 2865-2873.

Siraj, F., & AlGahtha, M. A. (2007). Mining Enrolment Data Using Predictive and Descriptive Approaches. *Knowledge-Oriented Applications in Data Mining*, 53-72.

Souza, J., Matwin, S., & Japkowicz, N. (2002). *Evaluating data mining models: a pattern language.* Paper presented at the Proceedings of the 9th Conference on Pattern Languages of Programs, USA.

Sterritt, R., Adamson, K., Shapcott, C., & Curran, E. (2000). Data mining telecommunications network data for fault management and development testing. *Proceedings of Data Mining Methods and Databases for Engineering, Finance and Other Fields, Cambridge, UK*, 299-308.

Sumathi, S. N. S. (2006). *Introduction to Data Mining and its Applications.*

Usama Fayyad, G. P.-S., Padhraic Smyth. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine Volume 17.*

V. Umayaparvathi, K. I. (2012). Applications of Data Mining Techniques in Telecom Churn Prediction. *International Journal of Computer Applications (0975 – 8887) Volume 42– No.20.*

Velickov, S., & Solomatine, D. (2000). *Predictive data mining: practical examples.* Paper presented at the Artificial Intelligence in Civil Engineering. Proceed. 2nd Joint, Workshop. March, Cottbus, Germany.

Weiss, G. M. (2002). Predicting telecommunication equipment failures from sequences of network alarms. *Handbook of Knowledge Discovery and Data Mining*, 891-896.

Weiss, G. M. (2005). Data mining in telecommunications *Data Mining and Knowledge Discovery Handbook* (pp. 1189-1201): Springer.

Welcker, L., Koch, S., & Dellmann, F. (2012). Improving classifier performance by knowledge-driven data preparation *Advances in Data Mining. Applications and Theoretical Aspects* (pp. 151-165): Springer.

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining.* Paper presented at the Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.

APPENDICES

Appendix 1: Sample VBA code applied in the preprocessing step of the study.

```
Sub Button2_Click()

Dim var1 As String

Dim usedrows As Long

Dim myarray1 As Variant

Dim myarray2 As Variant

Dim myarray3 As Variant

Dim myarray4 As Variant

Dim myarray5 As Variant

myarray1 = Array("finance", "state", "org", "organization", "bureau", "agency", "authority",
"ministry", "aa", "city", "woreda", "auth", "zone", "era", "erca", "ercs", "family", "federal",
"federation", "fira", "office", "immigration", "institute", "board", "zonal", "administration",
"corporation", "road", "water")

myarray2 = Array("usa", "unhcr", "undp", "gtz", "clinton", "relief", "foundation", "frontier", "u n
h c r", "children", "world", "african", "africa", "goal", "joy", "food", "save")

myarray3 = Array("university", "unversity", "college", "collage", "school", "preparatory",
"academy", "acadamy", "collge")

myarray4 = Array("business", "center", "trade", "trading", "engineering", "enterprise", "studio",
"plc", "p.l.c", "hotel", "company", "service", "stationary", "import", "export", "solution",
"limited", "partnership", "ayala", "resturant", "travel", "advertizing", "moenco", "brewery",
"bus", "shop", "enginering", "cocacola", "shop", "guest", "electronics", "secretarial", "secreterial",
"tech", "ltd")

myarray5 = Array("hospital", "clinic", "laboratory", "health", "medical")

usedrows = ActiveSheet.UsedRange.Rows.Count
```

```
usedrows = usedrows + 3
For i = 4 To usedrows
  var1 = Cells(i, 2)
  Lvar1 = LCase(var1)
  For x = 0 To 27
    testarray1 = InStr(Lvar1, myarray1(x))
    If testarray1 > 0 Then
      If Cells(i, 1) <> " " Then
      Cells(i, 3) = "Governmental Organization"
      End If
    End If
  Next
  For y = 0 To 16
    testarray1 = InStr(Lvar1, myarray2(y))
    If testarray1 > 0 Then
      If Cells(i, 1) <> " " Then
      Cells(i, 3) = "Non Governmental Organization"
      End If
    End If
  Next
   For x = 0 To 5
    testarray1 = InStr(Lvar1, myarray3(x))
    If testarray1 > 0 Then
```

```
    If Cells(i, 1) <> " " Then
    Cells(i, 3) = "Education Center"
    End If
  End If
Next
For y = 0 To 34
  testarray1 = InStr(Lvar1, myarray4(y))
  If testarray1 > 0 Then
    If Cells(i, 1) <> " " Then
    Cells(i, 3) = "Business Center"
    End If
  End If
Next
For y = 0 To 4
  testarray1 = InStr(Lvar1, myarray5(y))
  If testarray1 > 0 Then
    If Cells(i, 1) <> " " Then
    Cells(i, 3) = "Health Center"
    End If
  End If
Next
var2 = InStr(Lvar1, "bank")
var3 = InStr(Lvar1, "airline")
```

```
var33 = InStr(Lvar1, "ticket")
var4 = InStr(Lvar1, "internate")
var5 = InStr(Lvar1, "enternate")
var6 = InStr(Lvar1, "internet")
var7 = InStr(Lvar1, "embassy")
var8 = InStr(Lvar1, "insurance")
var9 = InStr(Lvar1, "cbb")
'var10 = InStr(Lvar1, "labratory")
var11 = InStr(Lvar1, "cbe")
var12 = InStr(Lvar1, "eepco")
'var13 = InStr(Lvar1, "university")
var14 = InStr(Lvar1, "embacy")
var15 = InStr(Lvar1, "cafe")
If var2 > 0 Or var8 > 0 Or var9 > 0 Or var11 > 0 Then
  If Cells(i, 1) <> "" Then
  Cells(i, 3) = "Bank or Insurance"
  End If
ElseIf var3 > 0 Or var33 > 0 Then
  If Cells(i, 1) <> "" Then
  Cells(i, 3) = "Air Line Company"
  End If
ElseIf var4 > 0 Then
  If Cells(i, 1) <> "" Then
```

```
    Cells(i, 3) = "Hotel"
    End If

  ElseIf var4 > 0 Or var5 > 0 Or var6 > 0 Or var15 Then
    If Cells(i, 1) <> " " Then
    Cells(i, 3) = "Internet Cafe"
    End If
  ElseIf var7 > 0 Or var14 Then
    If Cells(i, 1) <> " " Then
    Cells(i, 3) = "Embassy"
    End If

  ElseIf var12 > 0 Then
    If Cells(i, 1) <> " " Then
    Cells(i, 3) = "EEPCo"
    End If
  'ElseIf testarray1 = 0 Or testarray2 = 0 Then
    'If Cells(i, 1) <> " " Then
    'Cells(i, 3) = "others"
    'End If
  End If
Next
End Sub
```

```
Sub Button3_Click()
usedrows = Sheet1.UsedRange.Rows.Count
usedrows = usedrows +3
Dim var1 As Variant
For i = 2 To usedrows
var1 = Cells(i, 9)
If Cells(i, 9) <> " " Then
Cells(i, 10).Value = Left(var1, 10)
End If
Next
End Sub
Sub Button5_Click()
Dim var1 As String
Dim usedrows As Long
Dim myarray1 As Variant
Dim myarray2 As Variant
Dim myarray3 As Variant
Dim myarray4 As Variant
Dim myarray5 As Variant
myarray1 = Array("FAN-CAAZ", "FAN-NAAZ", "FAN-SWAAZ", "FAN-SAAZ", "FAN- EAAZ",
"FAN-WAAZ", "FAN- EAAZ", "FAN - EAAZ", "FAN- WAAZ")
myarray2 = Array("CS B&CC", "CS Broadband Agent", "CS Enterprise Agent", "CS Network &
TT", "CS VIP Team")
```

```
myarray3 = Array("O&M ER", "O&M NER", "O&M NR", "O&M NWR", "O&M SER", "O&M
SWR", "O&M SR", "O&M WR")

myarray4 = Array("FAN NWR", "FAN SER", "FAN ER", "FAN SR&SSWR", "FAN SWR", "FAN
WR", "FAN NER", "FAN NR", "FAN NWR")

usedrows = ActiveSheet.UsedRange.Rows.Count

usedrows = usedrows + 3

For i = 4 To usedrows

  var1 = Cells(i, 6)

  For x = 0 To 5

    testarray1 = InStr(var1, myarray1(x))

    If testarray1 > 0 Then

      If Cells(i, 1) <> " " Then

      Cells(i, 7) = "FAN-Addis"

      End If

    End If

  Next

  For x = 0 To 4

    testarray1 = InStr(var1, myarray2(x))

    If testarray1 > 0 Then

      If Cells(i, 1) <> " " Then

      Cells(i, 7) = "Customer Service"

      End If

    End If
```

```
Next
For x = 0 To 7
  testarray1 = InStr(var1, myarray3(x))
  If testarray1 > 0 Then
    If Cells(i, 1) <> " " Then
    Cells(i, 7) = "O&M Regional"
    End If
  End If
Next
For x = 0 To 8
  testarray1 = InStr(var1, myarray4(x))
  If testarray1 > 0 Then
    If Cells(i, 1) <> " " Then
    Cells(i, 7) = "FAN Regional"
    End If
  End If
Next
Var = InStr(var1, "O&M SR")
If Var > 0 Then
  If Cells(i, 1) <> " " Then
  Cells(i, 7) = "O&M Regional"
  End If
End If
```

```
Var0 = InStr(var1, "IS")
If Var0 > 0 Then
  If Cells(i, 1) <> " " Then
  Cells(i, 7) = "IS"
  End If
End If

'var2 = InStr(var1, "FAN - ")
'var3 = InStr(var1, "FAN-")
'If var2 > 0 Or var3 > 0 Then
  ' If Cells(i, 1) <> " " Then
   'Cells(i, 7) = "FAN-Regions"
   'End If
'End If
var3 = InStr(var1, "O&M Addis Zonal")
If var3 > 0 Then
  If Cells(i, 1) <> " " Then
  Cells(i, 7) = "O&M Addis Zonal"
  End If
End If
var4 = InStr(var1, "O&M Addis Regional")
If var4 > 0 Then
```

```
    If Cells(i, 1) <> " " Then
    Cells(i, 7) = "O&M Addis Regional"
    End If
  End If

  var5 = InStr(var1, "NNOC")
  If var5 > 0 Then
    If Cells(i, 1) <> " " Then
    Cells(i, 7) = "NNOC"
    End If
  End If
Next
End Sub
```