ADDIS ABABA UNIVERSITY

MASTERS THESIS

Performance Comparison of Classifiers for Prospective Buyers Identification in ethio telecom Mobile Cross-Selling Market

Author: Getachew Gemechu
Supervisor: Ephrem Teshale (PhD)

A Thesis Submitted in Partial Fulfilment of the Requirements for the degree of Master of Science in

(Telecommunication Engineering)

Addis Ababa institute of Technology
School of Electrical and Computer Engineering

February 22, 2020

Addis Ababa University

Addis Ababa institute of Technology

School of Electrical and Computer Engineering

This is to certify that the thesis prepared by **Getachew Gemechu**, entitled *Performance Comparison of Classifiers for Prospective Buyers Identification in ethio telecom Mobile Cross-Selling Market* and submitted in partial fulfilment of the requirements for the degree of Master of Science complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Examiner 1 _____ Signature _____ Date _____

Examiner 2 _____ Signature _____ Date _____

Supervisor Ephrem Teshale (PhD) Signature _____ Date _____

_____

Dean, School of Electrical and Computer Engineering

Declaration of Authorship

I, Getachew Gemechu, declare that this thesis titled, "Performance Comparison of Classifiers for Prospective Buyers Identification in ethio telecom Mobile Cross-Selling Market" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at Addis Ababa University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

_____

Date:

_____

Abstract

Direct marketing is a form of communicating an offer directly to a targeted group of customers through a variety of media. It plays a major role in customer retention and service provisioning tasks. Retaining customers by providing products and services that meet their need is one of the main objectives of customer relationship management. Identifying prospective customers for direct marketing enables a company to reach specific audiences which will more effectively respond to promotions. Moreover, direct marketing helps businesses to optimize their marketing budget, keeps current customers loyal to them, and makes businesses capable of measuring the result obtained from promotions.

Ethio telecom promotes service packages to its customers through SMS and mass communication channels. However, promotions should target customers based on the specific services they use, and customer over-touching should be reduced especially during SMS advertisement. In the current practice, no scientific methodology is implemented to estimate the potential respondents to cross-selling market promotion. Promotions are communicated to both potential buyers and non-buyers without distinguishing the two groups. Direct marketing approaches help the company to effectively allocate resources and give services based on the interests of customers.

The aim of this thesis is to identify prospective customers in ethio telecom mobile value-added service market. To achieve this goal, five classifiers namely Naive Bayes, Neural network, SVM, K-nearest neighbour, and Decision tree (J48) tested with customers service usage historical data. In this process, 900,000 customers' actual CDRs from ethio telecom were gathered and raw data aggregated with the aim of representing users' behaviour. The representation was based on users' responses towards service fee and time preference to use services. Sixteen feature variables and one predictor variable are constructed from the raw CDR collected. Data cleaning and class balancing done, and the selected classifiers tested for their accuracy in identifying prospective buyers of service packages.

The right customers for direct marketing are identified and ways to minimize customer over-touching during the promotion of low-price packages are shown. So, beyond selecting a classifier with high accuracy, the study aims at maximizing correctly classified instances. Accordingly, except Naïve Bayes classifier, the other four classifiers resulted in better performance. with Neural network classifier more than 90% of voice and 93% of SMS packages potential buyers are identified. whereas Decision tree (J48) has scored the best result with data package buyers dataset by identifying more than 94% of potential buyers.

Keywords: **Cross-selling, Classification techniques, SVM, Neural Network, KNN, Decision Tree(J48), Mobile value-added service market**

Acknowledgements

First of all, I would like to thank the almighty God, the lord and holy saviour Jesus Christ and his mother for keeping me safe and giving me the continuous health and courage to go through this journey. Next, I heartily acknowledge Dr Ephrem Teshale, my advisor, for continuing to encourage me through the long number of months writing and rewriting these chapters. His modest guidance and professional style will remain with me as I continue my career.

I also thank my committee members, Dr Yalemzewd Negash and Dr Murad Ridwan, for their valuable recommendations pertaining to this study and assistance in my professional development. Gratitude is extended to ethio telecom for full funding to pursue this research.

To ethio telecom staffs for the data, data, and extra data I would like to thank you. To my best friend Tesfa who helped me as my editor, though a small word of thanks is not enough for valuable comments he gave me, I do thank you from the bottom of my heart. To my best friend, Daniel Kebede, thank you for support in materials. To my wife Weynshet, and my daughter Ephrata, who sacrificed their precious time of togetherness, your encouragement is greatly appreciated and your love is the greatest gift of all.

Getachew Gemechu
January 2020

Contents

List of Figures

List of Tables

List of Abbreviations

ABT	**A**nalytic **B**ase **T**able
ADT	**Alternating D**ecision **T**ree
ANN	**A**rtificial **N**eural **N**etwork
ARPU	**A**verage **R**evenue **P**er **U**ser
AUC	**A**rea **U**nder **C**urve
BFT	**B**readth-**F**irst **T**ree
BMS	**B**ulk **M**essaging **S**ervice
CART	**C**lassification **A**nd **R**egression **T**ree
CBS	**C**onvergent **B**illing **S**ystem
CI	**C**lass **I**mbalance
CRBT	**C**olorful **R**ing **B**ack **T**one
CRM	**C**ustomer **R**elationship **M**anagement
CVR	**C**lassification **V**ia **R**egression
DT	**D**ecision **T**ree
EDA	**E**xploratory **D**ata **A**nalysis
FN	**F**alse **N**egative
FP	**F**alse **P**ositive
GA	**G**enetic **A**lgorithm
ID3	**I**terative **D**ichotomiser **3**
KNN	**K N**earest **N**eighbour
LB	**L**ogit **B**oost
LR	**L**ogistic **R**egression
MCC	**M**ulti **C**lass **C**lassifier
MMS	**M**ultimedia **M**essaging **S**ervice
MSISDN	**M**obile **S**tation **I**nternational **S**ubscriber **D**irectory **N**umber
MVAS	**M**essage **V**alue **A**dded **S**ervice
NB	**N**aive **B**ayes
NBM	**N**aive **B**ayes **M**ultinomial

OTAPS	**O**ver the **A**ir **P**rovisioning **S**ervice
RAF	**R**andom **F**orest
ROC	**R**eceiver **O**perating **C**haracteristics
SDP	**S**ervice **D**elivery **P**latform
SMS	**S**hort **M**essage **S**ervice
SQL	**S**equential **Q**uery **L**anguage
SVM	**S**upport **V**ector **M**achine
TN	**T**rue **N**egative
TP	**T**rue **P**ositive
USSD	**U**nstructured **S**upplementary **S**ervice **D**ata
VAS	**V**alue **A**dded **S**ervice
VMS	**V**oice **M**ail **S**ervice
WMS	**W**elcome **M**essage **S**ervice

Chapter 1

INTRODUCTION

In this chapter, an introduction of the thesis is provided so that the reader gets an overall view of the work. Starting with background information, the chapter discusses the research problem, objectives of the research, and research questions. Finally, the importance of the study and its outline are also presented.

1.1 Background of the study

Prospective customers are potential customers to product or service market. In the context of cross-selling market, prospective customers are those that have bought a certain type of products and/or services but with good probability to buy additional add-in products or a superior version of the already acquired products. Business objectives of getting, developing, and retaining customers supported by marketing applications like customer acquisition, cross-selling, up-selling, deep-selling, and customer retention[1]. One of these customer development techniques is Cross-selling which is a sales technique used to get a customer to spend more by purchasing a product related to what already bought. Marketing applications and campaigns also supported by classification modelling to identify prospective buyers in the cross-selling markets. [2].

In a competitive market, customer retention activity is much cheaper than customer acquisition activity. Customers can be retained with some attractive offers and discounts [3]. One of these customer retention activities is cross-selling. It is the strategy of selling other products to a customer who has already purchased (or signalled their intention to purchase) a product from the vendor [2]. In addition to providing a variety of services, in a telecom service market, the targets of cross-selling action

are three optional services (low-priced local call, inexpensive late call, and cheap calls inside the network) that allow the customer to lower connection cost. However, the effectiveness of this activity is based on persuading customers to purchase add-in products, which highly depends on promotion activities.

Communicating customers is a simpler task for a telecom service company. Contact centre, SMS (short messaging service) channels, and customer emails collected during customer acquisition period make this task simpler. However, there should be great care during contacting customers. Even though SMS communication incurs minor cost to a telecom company, it results with too many SMS communication during promoting a large number of offers. Such SMSs annoy customers and as the result customers quickly learn to ignore them. This kind of customers reaction minimize the limited opportunity in the cross-selling market. So, campaigns should be carefully targeted to maximize the probability to meet their target [2].

Targeted marketing is becoming not only a management practice but also the only efficient way to maintain a good relationship with customers. Due to the existing relationship, mutual trust, and the familiarity of the current business process, focusing on existing customers will bring more business value. There is also the possibility of identifying common attributes of successful cases among the existing customers and targeting efficiently and cost-effectively [4].

Marketers use channels like mail, the Internet, e-mail, telemarketing (phone), and other direct channels for direct marketing campaigns to communicate a message to their customers, in order to prevent churn (attrition) and to drive customer acquisition and purchase of add-on products.

Cross-selling, deep-selling or up-selling campaigns are implemented to sell additional products, more of the same product, or alternative but more profitable products to existing customers. When not refined, these campaigns, although potentially effective, can also lead to a huge waste of resources and to bombarding and annoying customers with unsolicited communications. Data mining models and classification (propensity) models, in particular, can support the development of targeted marketing campaigns. They analyse customer characteristics and recognize the profiles of the target customers. New cases with similar profiles are then identified, assigned a high propensity score, and included in the target lists [1].

1.2 Literature review

In cross-selling market or VAS (value added service) market, prospect identification has been studied by different researchers. The researches [2] [5] [6] [7] [8] [9] focus to predict respondents of promotions in the cross-selling market using data mining techniques. Different data mining models have been tested for accuracy of predicting respondents to cross-selling promotion.

The researcher in [2] used two cross-selling approaches: based on classifiers for prediction of potential respondents and based on Bayesian networks to predict respondents of the cross-selling market with visualization of interesting association rules. This paper compared accuracy of Naive Bayesian classifier, Decision trees (J4.8), Boosted decision trees (AdaBoostM1), and Support vector machines. It also modelled Bayesian Network that predicts the respondents of promotion. Based on the comparison of the two approaches classifier-based approach succeeded in better accuracy than Bayesian network-based approach. However, Bayesian network-based approach gave better insight into which customers are more likely to buy a certain service.

Researchers in [6] have also proposed a model to facilitate cross-selling. In this model, the accuracy of three algorithms Logistic Regression (LR), Artificial Neural Network (ANN) and Decision Tree (DT) optimized by Genetic Algorithm (GA) tested for the prediction of respondent. They suggested a model that combines the results of each classifier by weighted-averaging method, although conventional combination methods just combine the results by applying simple techniques such as averaging and majority voting. Their proposed model uses two thresholds, (i.e. lower threshold between low probability prospects and grey zone unpredictable and upper threshold between grey zone and highly probable prospects), for interpretation and integration of the continuous scores into final decision more effectively.

The research in [7] have studied the prediction of likings, success factors and generation of business intelligence, to understand telecom customers and their preferences. The prediction accuracy of ANN and SVM with three kernel functions (Polynomial, Radial basis function and Sigmoid) have been studied. The result was SVM-RBF Kernel is a very powerful kernel technique for the classification of telecom customers when compared with different kernel methods. Additionally, ANN

also tested over a number of repetitions.

The study in [4] compares fourteen algorithms (NaïveBayesMultinomial (NBM), ClassificationViaRegression (CVR), LADTree (LADT), ThreshholdSelector(TS), Bagging, DecisionTable (DT), BFTree (BFT), ADTree (ADT), LogitBoost (LB), NaïveBayes (NB), RotationForest (ROF), RandomForest (RAF), J48, MultiClassClassifier (MCC)) in the interest of finding the most efficient and accurate data-mining algorithm in targeted marketing. The Comparison was performed in terms of processing cost, accuracy rate, Type I and Type II error. Among the fourteen algorithms compared NaïveBayes Multinomial scored the lowest type II error without feature selection. When feature selection was done accuracy has been improved by 10% on average, type II error decreased by 40% on the average. By using two different class imbalance ratio for positive and negative (Yes/No) class instances in combination with and without feature selection the research concluded that Naïve Bayes Multinomial the ideal algorithm considering all factors (class-imbalance, and feature selection).

Cross-selling prospect identification studies done in telecommunication and banking service sectors and which are considered more relevant to this thesis are summarized with Table 1.1. The table shows some of the research titles algorithms used high accuracy algorithms found by each research, data used and research consideration for class imbalance.

TABLE 1.1: Summarization of some cross-selling researches in banking and telecom sector.

	Research Title	*sector*	*Algorithms used*	*High Acc. Algorithm*	*Data*	*CI consideration*	*Ref.*
1	Cross-selling models for telecommunication services	Telecom	BN, DT(j48), AdaBoostM1, SVM	SVM	Synthetically generated	Not explained	[2]
2	Facilitating cross-selling in a mobile telecom market to develop customer classification model based on hybrid data mining techniques	Telecom	LR, DT, ANN, GA	Combination of the four i.e. optimization of LR, DT, & ANN with GA	Korean telecom data	Not explained	[6]
3	A Computational Intelligence based Approach to Telecom Customer Classification for Value Added Services	Telecom	ANN & SVM (Polynomial, RBF, and Sigmoid)	SVM - RBF	Telecom companies (undisclosed)	Not explained	[7]
4	Conducting Efficient and Cost-effective Targeted Marketing using Data Mining Techniques	Bank	NBM, CVR, LADT, TS, Bagging, DT, BFTree, ADTree, Logit Boost, NB, ROR, RAF, J48, MCC	NBM	Bank direct marketing	1:5 and 1:20	[4]
5	A Classification Based Model to Assess Customer Behavior in Banking Sector	Bank	ANN, DT, KNN	ANN	Banking data	Not explained	[10]

1.3 Problem definition

The core service market is becoming saturated globally. Communication service providers are using value-added services to get more revenue by adding extra features in the mobile phone network. Provision of services that are customized to users should be supported with market targetting techniques in order to get a better achievement in the sales of these value-added services. In support of this, scholars are studying market targetting with data mining techniques. The aim of these target customer identification study is to minimize customer attrition and to enhance the effectiveness of promotions. As described earlier promotions need to be directed carefully in order to protect customers from being fed-up with frequent promotional messages.

Ethio telecom provides voice, data, and SMS core communication services and additionally two categories of mobile value-added services (VAS) in the cross-selling market. The first category is value-added services which are used for direct communication of company to service users and among service users each other. The second category is platforms for other services, security, and business support applications.

Ethio Gebeta is one of the USSD based package selling service which can be used with any mobile phone(feature/smart). Ethio Gebeta users purchase voice, SMS, and data packages by calling the short-code number "*999#". The service adopts the advantages of USSD service that is no store and forward and menu-based communication with the user. Ethio telecom releases promotions continually since the start of provision of value-added services.

Promotions have been disseminated through channels TV & Radio, bulk short messages, social media, and brochures. These channels have their own promotional expenses. Monetary expenses are not the only costs rather customers response to too many promotional communications like bulk SMSs should also be considered. Improper use of promotional bulk SMSs results in customer over-touching. Hence, it forces customers to learn ignoring promotional contacts. Over all, customers perception to promotional communications will be distorted and the business misses target of promotion [2].

During promoting services in the cross-selling market, selecting potential respondents is one of the solution strategies to decrease customer over-touching. Potential respondents of cross-selling market promotion have been identified using data mining in earlier researches. But smaller number of instances and unbalanced class in a dataset are the two major challenges in data mining [11]. In earlier researches some used smaller number of instances. For example, [6] have used 3,080 instances of Korean telecom customers data and [7] have used dataset with 6,000 instances by dividing into equal half for training and testing. Additionally, [2] studied with statistically generated CDR data. So far, consideration of class imbalance has not been explained in all these researches.

The data used in this research is ethio telecom's users CDR (call detail record). This assures the feasible representation of behaviour of telecom users. The other advantage is that a greater number of instances are used in this research than the related researches. Class balancing is considered in order to minimize the effect of prediction to the majority classes.

1.4 Research questions

This research answers the following key research question.

1. Which of the data features have more information about the interaction between package service market and its users?
2. What type of classification model predicts prospective respondents of the cross-selling market with better accuracy?

1.5 Research objectives

1.5.1 General objective

The general objective of this study is to select effective classification technique for identification of potential respondents in the cross-selling market promotion based on user's historical usage data.

1.5.2 Specific Objectives

- To identify data features more useful for classification of cross-selling market promotion respondents.
- To propose a classification technique with better accuracy for cross-selling promotion respondent identification.

1.6 Research contribution

Upon the completion of this thesis:

- Provision of value-added service to customers who are subjected to use them will be achieved. This reduces ineffective promotional communications with customers that are not interested with these extra featured services. As a result, high client satisfaction will be achieved and the company will be reputable to attract and hold customers.
- Campaigns and promotion will be provided based on the preference of customers in order not to be rejected [12]. Marketing communication with customers which does not consider customers interest will be exasperating communication.
- Investigation of customers usage data for valuable marketing knowledge will be realized. Large amount of customers usage data (CDR) archived or purged in few months without further investigation for market valuable information and strengthening customer relationship management practices.
- Provides an opportunity of testing classifiers performance with ethio telecom users data. Mostly data mining studies use synthetically generated or publicly available data sets. These data sets are not always applicable to represent users behaviour towards services provided by specific telecom service provision companies.

1.7 Research methodology

Data to be used in this research will be collected from Convergent billing system as raw CDR. One million MSISDN (mobile station international subscriber directory

number) numbers will be randomly sampled from the collected CDR with Oracle simple random selection function (rand()). In the sampling process one thousand instances from 92 available prefix are selected randomly. Additionally, target user's MSISDN number which is phone number of users who bought service packages will be collected from package buyers report. Both these data will be uploaded to Oracle server prepared for the purpose of this study. In the next step raw CDR will be aggregated into a format that represent users behavioural information in using the three services. After collection and aggregation of usage data, target user's numbers will be marked by using package users' information available in the database.

In addition to cleaning the data from outliers and extreme values, class instances will be balanced with SMOTE (synthetic minority oversampling technique). After this step a clean and balanced data will be partitioned into training and testing datasets. The training dataset will be supplied to the five selected classifiers and the performance of these classifiers will be checked with testing instances. This process is depicted in the Figure below.

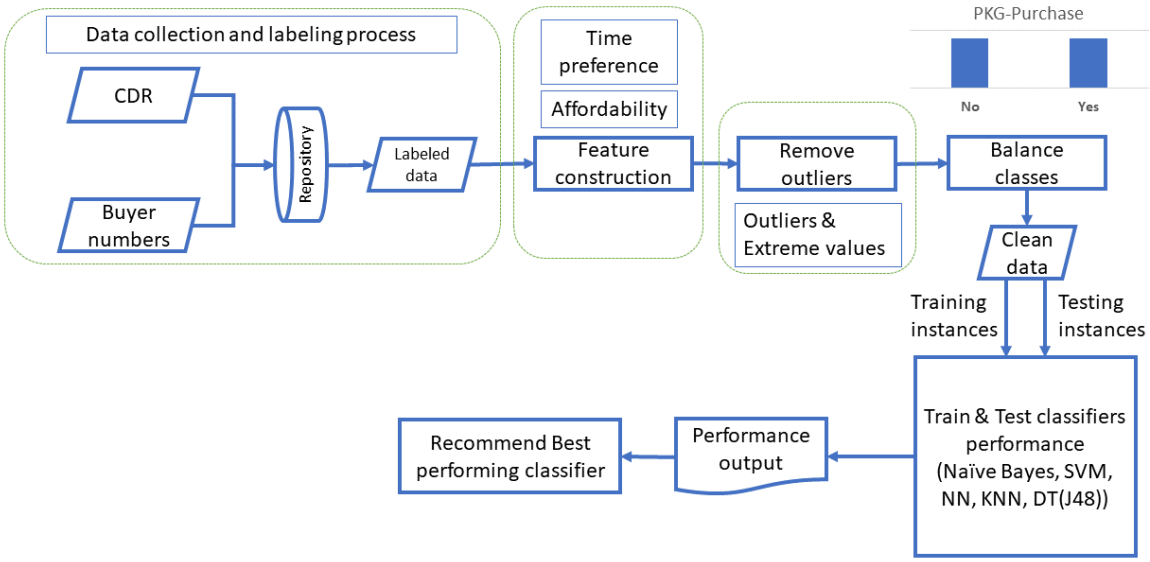

FIGURE 1.1: Diagrammatic representation of research methodology.

1.8 Outline of the thesis

Six chapters are included in this thesis. The first chapter provides the background of this study, research problem definition, the objective of this research and research question with a brief discussion about the importance of the study. Related studies are reviewed in the second chapter which discusses the importance of cross-selling market promotional costs in this market, behavioural representation of users from their usage data and prior studies are presented. In the third chapter technical review about data mining processes in a classification-study is discussed by briefly introducing the data mining process steps and core concept of five selected algorithms from five categories of classification techniques. The fourth chapter is the one which discusses the methodology of this research. It discusses about data collection, data aggregation technique and feature construction method used, data cleaning and experiment setup. Results of experiments will be discussed in chapter five and final analysis and conclusion will be presented in chapter six.

Chapter 2

CROSS-SELLING MARKET IMPORTANCE AND CHALLENGES

This chapter, discusses about importance of cross-selling market its promotional cost considerations, and techniques of users behaviour representation from their usage record in earlier studies, in cross-selling market, and the .

2.1 Classification of cross-selling respondents with data mining

Value Added Service (VAS) is a digital service that adds extra features in the mobile phone network, like online games, image and ring tones download, email, voucher and electronic transactions etc. According to [6] the traditional voice communication service is becoming widespread and operators are experiencing difficulties in attracting more customers. Most of the studies agree that VAS market has high and flexible ARPU (Average Revenue Per User) returns [12][7] [10].

The goal of data mining-based study is to design a system which predicts, the customers that are likely to accept when services that lower connection cost for a given customer are suggested.

Data mining models are machine learning algorithms that learn the relationship between a set of descriptive features and a target feature based on historical examples, or instances. In the process of searching for consistent models, training useful prediction models is not the only target because of two reasons. The first is noise in the dataset forces prediction models to make incorrect predictions even though they are

consistent with noisy data, and the second is the limit in number of the possible set of sample instances in the domain. These two reasons make machine learning an ill-posed problem [11].

Data mining researches have been done to predict market prospects in banking and telecom sectors [2] [5] [6] [7] [8] [9]. All these researches in cross-selling market prediction area were searching for best model that captures the association between feature and class variables. However, the nature of classification research in these modelling was not a simple task. One of the challenges arises from the size of the dataset used for the process. Increasing the size of the dataset may not avoid the challenge. When the dataset size increases noise also increases. Training with noisy data makes prediction inconsistent. Training sets mostly represent a small sample of possible set of raw variable combinations in the domain. These two reasons make machine learning problem an ill-posed problem. A unique solution cannot be determined using the only information that is available.

The challenges of data mining arise from the problem briefly stated above. Formulation of a solution for data mining problem needs different considerations like careful selection of data mining algorithms and consideration of class-imbalance.[13].

When samples of one class outnumber the other, some of the accuracy assessment methods became sensitive to class imbalance. Especially, methods that use data from both columns of confusion table are highly sensitive to imbalanced data. Some metrics such as accuracy and precision use both columns of the confusion matrix and change in data distribution change these metrics without affecting classifier performance. There are also metrics like Geometric Mean (GM) and Youden's index (YI) which will not be affected with imbalanced data while using both columns of the confusion matrix [14].

2.2 The cost of promotion in cross-selling marketing

Several marketing communications channels as contact centre, SMS and phone calls exist in the case of telecommunication service market. A large number of offers sent, assuming that some of those channels (especially SMS messages) incur almost no cost. However, the reality is quite different. The reason for that is the negative reaction of customers to too many offers. Since too many SMS are simply annoying, the

users quickly learn to ignore them. It follows that the amount of cross-selling opportunities is in fact quite limited, and the campaigns have to be carefully targeted such that the probability of a "hit" is maximized [2].

Cross-selling managers face challenging tasks of improving the low customer response rate and avoiding "over-touching" the customers. In cross-selling recommendations offering the right product to the right customer at the right time is very important [5].

2.3 Importance of cross-selling market

According to [2] cross-selling offers several advantages. High revenue from the extra products sold, an increase in the reliance of the customer on vendor and minimization of churn are some of the advantages which can be mentioned. Companies may also classify certain customers as former customer [15] if they are not responsive to promotions and campaigns. The reason for such classification is these customers are leading the company to losses because they have no value in return. Customers that delay or default on payments, or participates in fraud, including unfair abuse of the company's policies, the company may choose to stop approaching them for further sales and classify them as former customers, which is the last stage in the customer life cycle. As a result, former customers are rarely targeted for CRM campaigns [15].

With a special focus on cellular network operators, some of the specific aspects of cross-selling in the telecommunication industry are high volatility of telecom markets and very low level of customer loyalty as well as profitable offers for new customers from most service providers. However, customer loyalty is at a very low level in this sector, due to anti-monopoly actions taken by governments, as well as profitable offers for new customers from most service providers. Cross selling is thus very important for cellular operator since the more services a user has subscribed, the closer he/she bound to the company, and it is hard for him/her to change to another provider. The paper in [6] has stated the mobile operators are turning from traditional voice communication to mobile value-added services (VAS), which are new services to generate more ARPU (average revenue per user) and recommends cross-selling as critical way for mobile telecom operators to expand their revenues and profits.

2.4 Behavioural representation of telecom service users using CDR data

Users behaviour can be extracted from their usage history data in CDR. User behaviours like mobility, affordability to make a call, social interaction and others can be represented from historical usage data. The paper in [16] used Smart Home usage data (data produced by internet of things devices in home) to mine users' behaviour. As the result of this prediction to enhance intelligence of Smart Home systems. The paper in [17] studied data usage, mobility pattern and application usage behaviour of 2G and 3G mobile users using mobile traffic data collected at core metropolitan network. Researchers in [18] studied users' mobility behaviour using aggregated calling profiles of mobile phone users and subdivided users into three categories; resident, commuters, and visitors. User communication behaviour based on the incoming/outgoing call holding time has also been modelled in [19]and communication behaviour indicators like affordability, social role and calling habit of users are modelled using incoming and outgoing call frequency. All these studies tell us that usage related data collected can reflect user's behaviour whenever analysed with data mining techniques.

Chapter 3

CLASSIFICATION TECHNIQUES

Classification techniques are the methods used by certain types of classifiers (classification algorithms) to perform their task. Technical review of the underlying theory about these types of classification techniques and the five selected classifiers from each classification techniques is discussed in this chapter. Furthermore, the theoretical aspects concerning data preprocessing steps in data mining and classification accuracy evaluation concepts are also parts of this chapter.

3.1 Data collection and preprocessing

Understanding the business problem and formulating hypothesis, collection of raw data, preprocessing data, testing accuracy of algorithms, and interpretation of results are data mining process steps [20]. Classification techniques automatically learn a model of the relationship between descriptive features and target features from instances of past data [11].

During the first phase in any analytics project, prediction model building is not the goal of predictive data analytics projects, rather new customer cultivation, products sales, and process enhancement are the goals. Therefore, fully understanding the business (or organizational) problem that is being addressed, is the primary goal of data analyst and designing a data analytics solution for it is the next. The primary task is to understand different data sources available within an organization and various kinds of data that are available from these sources. However, this should be after the predictive data analytics method which addresses a business problem has been decided.

There are two possibilities in data generation or collection process which are designed experiment (when data generation process controlled by an expert) and observational approach (when there is no possibility to influence data generation process) [20].

Data aggregation is the next step after collection of the necessary data. There are raw and descriptive types of data features. Raw features are features that come directly from raw data source whereas derived data features are constructed from raw features with aggregation techniques like sum, average, minimum, maximum, etc. Since derived features are constructed from one or more data row, they are also more descriptive than raw features [13]. Users behaviour representation was the central focus during aggregation of usage records [21].

Undergoing Exploratory Data Analysis (EDA) plays a major role to improve the effectiveness of data mining tasks [22][20]. EDA helps to visually summarize the data and to find patterns and data anomalies [23].

After getting insight about the data there are methodologies to handle missing and erroneous entries in the data. Ignoring the tuple for a classification task, fulfilling the missing values manually, filling with global constants, using a measure of central tendency, filling with similar class-based attribute mean or median, and using most probable values are among those methodologies [22]. However, estimating missing values creates additional errors which affect results of data mining algorithms. The general methodology of handling missing and erroneous records is domain specific [24].

Building predictive data analytics models require specifying data, organizing this data in a specific kind of structure known as analytics base table (ABT). This includes all the activities required to convert the dissimilar data sources that are available in an organization into a well-formed ABT from which machine learning models can be induced. This part of data mining process is known as data preparation.

The Modelling process is when the machine learning work occurs, different machine learning algorithms are used to build a range of prediction models from which the best model will be selected for deployment. Five categories of classifiers used one from each classification technique (Instance-based, Statistical, SVM, Perceptron, and logic-based).

Finally, selecting models that are fully evaluated and verified to fit the organizational purpose is important. All the evaluation tasks required to show that a prediction model will be able to make accurate predictions after being deployed are covered during evaluation process and that it does not suffer from over-fitting or under-fitting.

3.2 Classification techniques

In data mining, classification is a data analysis task that categorizes a certain event to a given class based on recorded activity historical data. In this task, a model or classifier is constructed to predict class or categorical labels such as "safe" or "risky" for loan application data "yes" or "no" for market response data. It is a two-step process classification model construction (learning step) and using the constructed model to predict class labels for a given data (classification step). In the learning step, a classifier is built describing a predetermined set of data classes or concepts. In this step, a classification algorithm builds the classifier by analysing or learning from a training set made up of database tuples and their associated class labels. Each tuple belongs to a predefined class called class label attribute. The class label attribute is discrete-valued and unordered. Training tuples are formed from individual tuples by randomly sampling them from database under analysis. Since each tuple is provided with a class label, (classifier is told to which class each training tuple belongs) such type of learning process is called supervised learning [22].

There are various groups of classification models. Based on the technique they operate on, the most well-known are tree-based classifiers, rule-based classifiers, probabilistic models, instance-based classifiers, support vector machines, and neural networks [24] [25].

3.2.1 Probabilistic models

Bayes classifier (generative classifier) and Logistic regression (discriminative classifier) are two of the most used probabilistic classifier models. It is assumed that the data points within a class are generated from a certain probability distribution such as the Bernoulli or multinomial distribution in Bayes classifier. A naive Bayes assumption of class-conditioned Feature independence is sometimes used to make the modelling simpler. While in Logistic regression the target variable is assumed to

be derived from a Bernoulli distribution whose mean is defined by a parametrized logit function on the feature variables. Thus, the probability distribution of the class variable is a parametrized function of the feature variables. This is in contrast to the Bayes model which assumes a specific generative model of the feature distribution of each class [24].

3.2.1.1 Naive Bayes classifier

Bayes classifier is based on the Bayes theorem for conditional probabilities. This theorem quantifies the conditional probability of a random variable (class variable) given known observations about the value of another set of random variables (feature variables). In some cases, a conditional event usually corresponds to a combination of constraints on different feature variables, rather than a single feature variable. This makes a direct estimation of posterior probability much more difficult [22].

Naive Bayesian classifier predicts tuple X belongs to the class C_i which is one of the m classes some members of class C that is C_i if and only if

$$P(C_i|X) > P(C_j|X) \qquad for 1 <= j <= m, j \neq i \tag{3.1}$$

The objective is to maximize $P(C_i|X)$. The class C_i for which $P(C_i|X)$ is maximized is called the maximum posteriori hypothesis [22].

In rare class learning, probabilistic classifiers predict the labels of individual test instances. A set of test instances are provided to the learner, and it is desired to rank these test instances by their propensity to belong to a particularly important class C [24].

However, by using Bayes' theorem

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{(P(X)P(C_i|X))} \tag{3.2}$$

As $P(X)$ is constant for all classes, only $P(X|C_i)P(C_i)$ needs to be maximized. if the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely and $P(X|C_i)$ could be maximized otherwise, maximization should be done on $P(X|C_i)P(C_i)$.

The probability of test instances $(a_1...a_k)$ belonging to a class can be estimated within a constant of proportionality as follows:

$$P\left(C = c|x_1 = a_1, ...x_k = a_k\right) \propto (C = c) \prod_{j=1}^{k} P\left(x_j = a_j|C = c\right) \tag{3.3}$$

The relevance of this constant of proportionality, which is the inverse of the generative probability of specific test instance, is not while comparing scores across different classes but while comparing scores across different test instances. the constant of proportionality can be estimated easily using normalization [24].

Therefore, the Bayes probability can be estimated as follows: the class label c is assumed to be an integer drawn from the range {1 ... k} for a *k*-class

$$P\left(C = c|x_1 = \ a_1, \ \ldots \ x_k = \ a_k\right) = \frac{P\left(\ C = c\right) \ \prod_{j=1}^{k} P(x\_j = \ a\_j \ |C = c)}{\sum_{c=1}^{k} P\left(C = c\right) \prod_{j=1}^{k} P(x_j = \ a_j|C = c)} \tag{3.4}$$

When attributes in the datasets are many, the naive assumption of class-conditional independence is made to reduce computation. Referring attribute value A_k for tuple X with X_k, if A_k is categorical then $P(X|C_i)$ is the number of tuples of class C_i in D. and if A_k is continuous-valued, then we need to do somewhat more work, a continuous-valued attribute is typically assumed to have a Gaussian distribution with a mean μ, standard deviation σ and the calculation will be as shown in the below equation 3.5

$$g\left(x, \mu, \sigma\right) = \ \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.5}$$

So that

$$P(X_k|C_i) = g(X_k, \mu C_i, \sigma C_i) \tag{3.6}$$

Bayesian classification exhibits high accuracy and speed when applied to large databases. It has also comparable performance with decision tree and selected neural networks. Class conditional independence simplifies Naive Bayesian classifiers computation by assuming that the effect of a feature variable value on a given class is independent of the other feature variables values. The Bayesian classifier works based on Bayes' theorem [22].

Bernoulli model, multinomial model or Gaussian model with numeric data and parametric form of the conditional feature distribution are suitable for the use of Bayes model. If the process of discretization is implemented it can also be used with numeric datasets [24].

Implementation of Bayes model is possible using more general multivariate estimation methods. Such methods may be computationally more expensive and inaccurate with increasing dimensionality especially with limited training data. So, using theoretically more accurate assumptions, it is impossible to get signification practical accuracy [24].

3.2.2 Support vector machines

Support vector machines naturally defined for binary classification of numeric data. Since the class labels are drawn from {-1,1}, SVMs use separating hyperplanes as decision boundary between two classes and the optimization problem of determining these hyperplanes is set up with the notion of margin. A maximum margin hyperplane is one that cleanly separates the two classes, and for which a large region (or margin) exists on each side of the boundary with no training data points in it [24].

When the data is linearly separable there is an infinite number of possible ways of constructing a linear separating hyperplane between the classes. Performance of two classifiers may vary because of placement of a test instance in a noisy and uncertain boundary region between classes which is not easily generalizable from the

available training data. The most robust and correct classification can be achieved when minimum perpendicular distance to training points from both classes is as large as possible. It is possible to construct parallel hyperplanes with respect to separating hyperplane. These constructed hyperplanes touch the training data of either side and have no data points between them. The training data points on these hyperplanes are referred as support vectors and the distance between the two hyperplanes as the margin [22].

The maximum margin hyperplane can be set up with a non-linear programming optimization formulation that maximizes the margin by expressing it as a function of the coefficients of separating hyperplane. Optimal hyperplane can be determined by solving this optimization problem.

The separating hyperplane is of the form

$$\bar{W}.\bar{X} + b = 0 \tag{3.7}$$

Where $(\bar{X}_i)$ is a d dimensional row vector corresponding to i^{th} data point in the training set D which is denoted by $(\bar{X}, y_1)...(\bar{X}_n, y_n)$ for n data points and $y_i \in \{-1, 1\}$ is the binary class variable of the i^{th} data point. The row vector representing the normal direction to the hyperplane in d dimensional space is $\bar{w} = (w_1...w_d)$, and b is a scalar which is also known as bias. The vector $\bar{w}$ regulates the orientation of hyperplane from the origin. The $(d + 1)$ coefficients corresponding to $\bar{w}$ and b need to be learned from the training data to maximize the margin of separation between the two classes. Since it is assumed that the classes are linearly separable, such hyperplane can also be assumed to exist. All data points $\bar{x_i}$ with $y_i = +1$ will lie on one side of the hyperplane satisfying $\bar{W}.\bar{X} + b \leq 0$.

$$\bar{W} \,.\, \bar{X}_i + b \;\geq 0 \;\forall_i : y_i = +1 \tag{3.8}$$

$$\bar{W} \,.\, \bar{X}_i + b \;\geq -1 \;\forall_i : y_i = -1 \tag{3.9}$$

$\bar{W}$. $X + b = 0$ which can be assumed as a constraint is located in the centre of the two-margin defining the hyperplanes. The distance between hyperplanes touching the support vectors can be expressed by introducing parameter c.

$$\bar{W} \,.\, \bar{X} + b = +c \tag{3.10}$$

$$\bar{W} \,.\, \bar{X} + b = -c \tag{3.11}$$

For appropriate scaling of $\bar{W}$ and b, the value of c can be set to 1. So, the hyperplanes can be in the form of margin constraints as

$$\bar{W} \,.\, \bar{X} + b = \ +1 \tag{3.12}$$

$$\bar{W} \,.\, \bar{X} + b = \ -1 \tag{3.13}$$

In the decision boundary between these two hyperplanes, and all training data points for each class are mapped to one of the two extreme regions. This can be expressed as point-wise constraints on the training data points as follows:

$$\bar{W} \,.\, \bar{X}_i + b \ \geq +1 \ \forall_i : y_i = +1 \tag{3.14}$$

$$\bar{W} \,.\, \bar{X}_i + b \ \geq -1 \ \forall_i : y_i = +1 \tag{3.15}$$

The goal is to maximize the distance between the two hyperplanes for positive and negative instances denoted as margin. This distance is normalized difference between the constant terms, where normalization factor $L_2 - norm \ \ \|\bar{W}\| = \sqrt{\sum_{i=1}^{l} w_i^2}$ of the coefficients, because the difference between the constant terms is $\frac{2}{\|\bar{w}\|}$.

The high computational complexity of Support vector machines arises from larger optimization problem due to each training data point as a constraint. Such constrained non-linear programming problems are solved using a method known as La Grange relaxation.

Combining 3.14 and 3.15 we get

$$y_i \left(\bar{W} \, . \, \bar{X}_i + b \right) \geq +1 \ \forall_i \tag{3.16}$$

Using Lagrangian formulation 3.16 can be solved by using Karush-Kuhn-Tucker (KKT) conditions [22].

For small data less than 2000 tuples an optimization software package for solving constrained convex quadratic problems can be used to find the support vectors and maximum margin hyperplanes (MMH). But for larger data special and more efficient algorithms for training SVMs can be used.

Based on the Lagrangian formulation mentioned before the (MMH) can be rewritten as the decision boundary

$$d\left(X^T \right) = \sum_{i=1}^{l} y_i \alpha_i X_i X^T + b_0 \tag{3.17}$$

Where y_i is the class label of the support vector X_i ; X^T is a test tuple; $y\alpha_i$ and b_0 are numeric parameters that were determined by the optimization for SVM algorithm. And l is the number of support vectors when a test tuple X^T plug into 3.17 and checked the sign of the result, positive tells X^T falls on above MMH where the prediction result will be X^T belongs to $+1$ whereas if the sign is negative X^T belongs to class -1.

The complexity of the learned classifier is characterized by the number of support vector rather than the dimensionality of the data. So, SVMs tend to be less prone to over-fitting than some other methods [22].

If all other training tuples were removed and training were repeated, the same separating hyperplane would be found. Moreover, the number of support vectors found can be used to compute an (upper) bound on the expected error rate of SVM classifier, which is independent of the data dimensionality. An SVM with a small number of support vectors can have good generalization ever when the dimensionality of the data is high [22].

The real-world data is not linearly separable and in this case, training data points can violate the margin constraints at the expense of penalty. The tow margin hyperplanes separate out "most" of the training data points but not all of them. The level of violation for each margin constraint by training data point $(\bar{X}_i)$ is denoted by a slack variable $\xi_i \geq 0$. The new set of soft constraints on the separating hyperplanes may be expressed as follows [24].

$$\bar{W} \,.\, \bar{X}_i + b \;\geq +1 - \;\xi_i \;\; \forall_i : y_i = +1 \tag{3.18}$$

$$\bar{W} \,.\, \bar{X}_i + b \;\geq -1 - \;\xi_i \;\; \forall_i : y_i = +1 \tag{3.19}$$

$$\xi_i \;\geq 0 \; \forall_i$$

The slack variable interpreted as the distance of training data points from the separating hyperplanes when they lie on the "wrong" side of separating hyperplanes. The values of the slack variables are 0 when they lie on the correct side of the separating hyperplanes violations of positive ξ_i values are penalized by $c.\xi_i^r$, where c and r are user-defined parameters regulating the level of softness in the model. Small values of c result relaxed margins, whereas large values minimize data errors

and result in narrow margins. Sufficiently large c disallows any training data errors in separable classes, and it is the same as defaulting to the hard version of the problem. Mostly r is chosen 1 referred as hinge loss.

The objective function for soft-margin SVMs with hinge loss is defined as

$$O = \frac{\|\bar{W}\|^2}{2} + c\sum_{i=1}^{n}\xi_i \tag{3.20}$$

This is also a convex quadratic optimization problem that can be solved using Lagrangian methods.

The other method to solve SVM formulation is using kernels. When kernels are used knowledge about feature values is not important. But defining pairwise dot product of similarity function in the d'-dimensional transformed representation $\Phi\bar{X}$ with the use of kernel function $K\left(\bar{X}_i\ ,\ \bar{X}_j\right)$is important.

$$K\left(\bar{X}_i, \bar{X}_j\right) = \Phi\left(\bar{X}_i\right)\ .\Phi\left(\bar{X}_j\right) \tag{3.21}$$

In this case, all computations are performed in the original space and actual transformation $\Phi\left(\ .\ \right)$ does not need to be known if the kernel similarity function $K\left(\ .\ ,\ .\ \right)$ is known. Arbitrary non-linear decision boundaries can be approximated using carefully chosen kernel-based similarity [24]. The following are some of the admissible kernel functions:

(3.22) **Polynomial kernel of degree h:** $K\left(\bar{X}_i, \bar{X}_j\right) = \left(X_i.X_j + 1\right)^h$

(3.23) **Gaussian radial basis function Kernel:** $K\left(\bar{X}_i, \bar{X}_j\right) = e^{\frac{-\|X_i - X_j\|^2}{2\sigma^2}}$

(3.24) **Sigmoid kernel:** $K\left(\bar{X}_i, \bar{X}_j\right) = \tanh\left(k\bar{X}_i \,.\, \bar{X}_j - \sigma\right)$

There are no predefined rules in selecting kernels which will result in the most accurate SVM. Kernel chosen does not generally make larger difference in resulting accuracy [22]. However, different kernels have different flexibility. Optimal values of kernel parameters depend not only on the shape of the decision boundary but also on the size of the training dataset. Parameter tuning is important in kernel methods. With proper tuning, many kernel functions can model complex decision boundaries [24].

3.2.3 Neural network classifiers

Neural networks are models that simulate the human nervous system. In the same analogy as biological networks, artificial neural networks are referred to as neurons. These neurons are units of computation that receive input from some other neurons, make computation on these inputs and feed them into yet other neurons. The computation function at a neuron is defined by the weight on the input connection to that neuron. The computation function can be learned by changing these weights appropriately. Training data is used as external stimulus in artificial neural networks. Weights are incrementally modified whenever incorrect predictions are made. The architecture is the key to effectiveness of neural networks. Based on architecture from simple single-layer neural network to complex multi-layer networks exist [24].

An Artificial neural network is modelled with three basic elements [20]:

1. A set of connecting links from different inputs X_i (or synapses), each of which is characterized by a weight or strength w_{ki}. The first index refers to the neuron in question and the second index refers to the input of the synapse to which the weight refers. In general, the weights of an artificial neuron may lie in a range that includes negative as well as positive values.

2. An adder for summing the input signals X_i weighted by the respective synaptic strengths w_{ki}. The operation described here constitutes a linear combiner.

3. An activation function f for limiting the amplitude of the output y_k of a neuron.

It includes also an externally applied bias b_k which has the effect to increase or decrease the net input of the activation function depending on whether it is positive or negative.

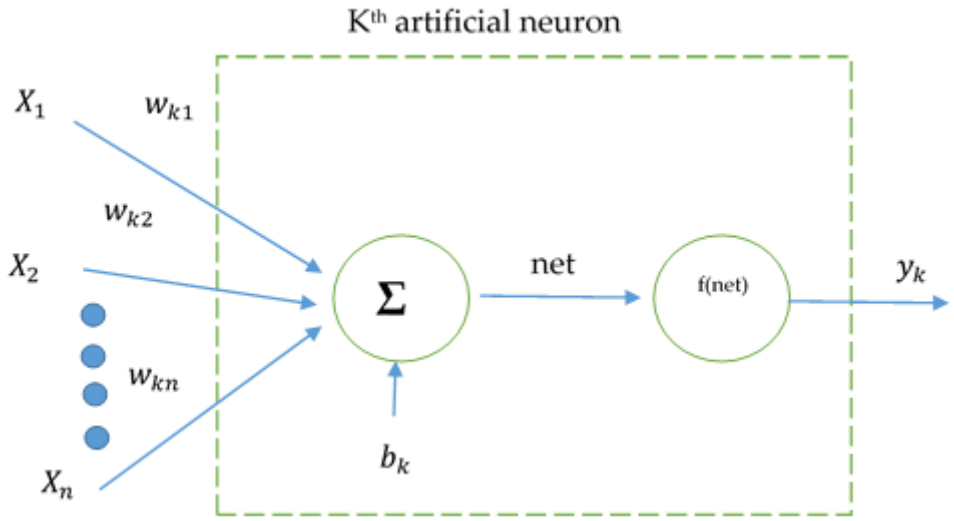

FIGURE 3.1: Model of an Artificial neural network [20]

There are several inputs $X_i, i = 1, ..., m$ and each input are multiplied by the corresponding weights w_{ki} where k is the index of the neuron the weighted sum of products $X_i w_{ki}$ for $i = 1, ..., m$ usually denoted as net.

$$net_k = X_1 w_{k1} + X_2 w_{k2} + \ldots + X_m w_{km} + b_k \tag{3.25}$$

Setting $w_{k0} = b_k$ and the default input $X_0 = 1$ the new summarization will be

$$net_k = X_0 w_{k0} + X_1 w_{k1} + X_2 w_{k2} + \ldots + X_m w_{km} = \sum_{i=0}^{m} X_i w_{ki} \tag{3.26}$$

This sum can be expressed as a scalar product of two m-dimensional vectors:

(3.27) $$net_k = X \,.\, W$$

Where X =$X_0, X_1, X_2, ..., X_m$; W = $w_{k0}, w_{k1}, w_{k2}, ..., w_{km}$

Artificial neuron computes the output y_k as a certain function of net_k value

(3.28) $$y_k = f(\ net_k)$$

The function f is called an activation function and it has various forms. Some commonly used activation functions are as shown in Table 3.1.

TABLE 3.1: Common activation functions [20]

Activation Function	**Input/output Relation**
Hard limit	$f(x) = \begin{cases} 1, & if\ net \geq 0 \\ 0, & if\ net < 0 \end{cases}$
Symmetrical hard limit	$f(x) = \begin{cases} 1, & if\ net \geq 0 \\ -1, & if\ net < 0 \end{cases}$
Linear	$y = net$
Saturating linear	$f(x) = \begin{cases} 1, & if\ net \geq 0 \\ net & if\ 0 \leq net \leq 1 \\ 0, & if\ net < 0 \end{cases}$
Symmetric saturating linear	$f(x) = \begin{cases} 1, & if\ net \geq 0 \\ net & if -1 \leq net \leq 1 \\ 0, & if\ net < 0 \end{cases}$
Log sigmoid	$y = \frac{1}{1+e^{-net}}$
Hyperbolic tangent sigmoid	$y = \frac{e^{net}-e^{-net}}{e^{net}-e^{-net}}$

3.2.4 Instance-based classifiers

Instance-based classifiers which are named also lazy learners work based on distance measures. Different distance measures are used to classify a given tuple. Unlike other classifiers, lazy learners wait until the last minute before doing any model construction to classify a given test tuple. These categories of classifiers do less work when a training tuple is presented and more work when making a classification or numeric prediction. Due to their requirement of efficient storage techniques and they better suite with implementation on parallel hardware and are computationally expensive. However, they naturally support incremental learning and are capable of modelling complex decision space having hyper-polygonal shapes [22].

One of these lazy learners is K-nearest-neighbour classifier which was first described in the early 1950s. It is based on learning by analogy which is a method of comparing test tuples with training tuples. It stores training tuples in an n-dimensional space and searches this space for closest Neighbors for the unknown tuple. The k training tuples are the k closest tuples or k nearest Neighbors for unknown tuple [22]. Distance is calculated based on distance metrics like Euclidean distance which is not more effective in terms of sensitivity [22], unsupervised Mahalanobis Metric, and Nearest Neighbors with Linear Discriminant Analysis [24].

The K parameter choice is commonly based on domain knowledge about the classification problem at hand. Mostly odd K values which from 1 up to 100 are chosen. The decision for value of k is done by testing iteratively and checking the result [20].

3.2.5 Tree-based classifiers

Decision tree classification is one of tree-based technique. In this type of technique, a set of hierarchical decisions on the feature variables are used to model the classification process. Split criterion, which is a condition used to make a decision on feature variables, is used at a particular node of the tree. Nodes in a tree are logical representations of a subset of the data spaced defined by split criteria in the nodes above it [24].

Because of their good accuracy, capability to handle multi-dimensional data, and appropriateness for exploratory knowledge discovery decision tree are popularly used. In addition to their no requirement to domain knowledge or parameter setting, their representation of acquired knowledge in tree form is intuitive and easy

to assimilate by humans. Learning and classification steps in decision tree are also simple and fast [22].

Decision tree induction is the learned from class-labelled training tuples using algorithms like C4.5, ID3(Iterative Dichotomiser), and CART (Classification and Regression Trees) [24]. ID3 algorithm was developed by J. Ross Quinlan in the late 1970s and early 1980s and expanded by E. B. Hunt, J. Marin and P. T. Stone. C4.5 is a successor of ID3 which was later developed by Quinlan. CART was developed by a group of statisticians (L. Breiman, J. Freidman, R. Olshen, and C. Stone). These algorithms adopt a greedy (i.e., non-backtracking) approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner. In this approach, the training set is recursively partitioned into smaller subsets as the tree is being built [22].

The separation of different classes among the children nodes is maximized with split criteria. Split criterion depends on the nature of the underlying attribute: only one type of split is possible for binary attributes, in the case of categorical attribute for r different values of attribute there are r-way split and $2r - 1$ binary splitting is also possible but it may not be possible when the value of r is large. This can be simplified by converting categorical variables to binary data with the use of binarization. For numeric attributes with small number r of ordered values, it is possible to create an *r-way* split for each distinct value but in the case of continuous numeric attributes, the split is performed by using a binary condition such as X less than or equal to a, for attribute value x and a constant a. Attribute selection methods like information gain, gain ratio, Gini index, and error rate can be used to select splitting attribute [24].

Information gain which is used by ID3 algorithm is based on information-theory of Claude Shannon. It minimizes the expected number of tests needed to classify a given tuple and guaranties that a simple tree is found. Entropy is closely related to information gain at a conceptual level. Still, there is no difference between using either of the two for a split [24] [22]. However, entropy measure can be used with both ID3 and C4.5 and the lower values are more desirable.

The split criterion based on error rate simply uses one minus the fraction of instances in a set of data points S belonging to the dominant class. For an r-way split of set S into $S_1, ..., S_r$, the overall error rate of the split may be quantified as the

weighted average of the error rates of the individual sets S_i where the weight of S_i is its magnitude. the split with the lowest error rate is selected from the alternatives [24].

The noise which results from decision tree growth to the very end until every leaf node contains only instances belonging to a particular class is contributed by the lower-level nodes contains a smaller number of data points. Generally, simpler models (shallow decision trees) are preferable to more complex models (deep decision trees) if they produce the same error on the training data.

Stopping the growth of the tree early is one possibility of reducing the level of overfitting. However, it is difficult to mark the correct stopping point. There are many different criteria for such a decision. One strategy is to explicitly penalize model complexity with the use of the minimum description length principle (MDL). In this approach, the cost of a tree is defined by a weighted sum of its error and its complexity (e.g., the number of the nodes). Information-theoretic principles are used to measure three complexity. Therefore, the tree is constructed to optimize the cost rather than only the error. However, the main problem with this approach is that the cost function is itself a heuristic that does not work consistently well across different datasets. holding out 20%of training data is more intuitive strategy. Pruning impact is tested on the holdout set and if it improves the classification accuracy, the hold is pruned. leaf nodes are iteratively pruned until it is no longer possible to improve the accuracy. Such an approach reduces the amount of training data for building the tree. the impact of pruning generally outweighs the impact of training-data loss in the tree-building phase [24].

3.3 Classifier performance assessment methods

Training more than one classifier and choosing the best one over another based-on performance on a given dataset is denoted as model selection. Holdout and random sampling, cross-validation, and bootstrap methods are some of the common techniques used for assessing classifier accuracy. In holdout method data randomly sampled into two independent sets (training and test). Two third of the data allocated for training set and one third for test set. The other variation of holdout method, random sampling repeats holdout method K times and take the average accuracies from each iteration. In k-fold cross-validation the initial data

randomly partitioned into k approximately equal size, mutually exclusive subsets or "folds" and training and testing performed K times. In the first iteration of the cross-validation method datasets other than the first portion used to train the first model and this process repeated K times leaving the second, third portions until k^{th} portion. Bootstrap method samples the given training tuples uniformly with replacement [22].

Classifiers' accuracy assessed based on information from a confusion matrix. Confusion matrix is built on four building blocks: true positives (TP) positive tuples that were correctly labelled by classifier, true negative (TN) negative tuples which were correctly labelled by classifier, false positive (FP) negative tuples which were incorrectly labelled by classifier as positive, and false-negative (FN) positive tuples which were incorrectly labelled by classifier as negative [22]. Metrics such as sensitivity, specificity, and accuracy are scalar performance measures that are obtained from the confusion matrix. Using these measures has problems such as sensitivity to imbalanced data (when one class in the dataset outperform the other) and ignoring the performance of some classes. Assessment metrics which use values from both the right and left columns of actual class column in Table 3.2 are unreliable because of their sensitivity to errors that arise due to class imbalance [14].

TABLE 3.2: Confusion matrix for a binary class classifier

		Actual Class	
		Positive	**Negative**
Predicted Class	**True(T)**	True Positive (TP)	False Positive (FP)
	False(F)	False Negative (FN)	True Negative (TN)

As the data distribution between the actual positive and negative classes change the values of classifier accuracy measuring metrics that use both columns get changed though the classifier performance does not change. So, these metrics do not differentiate between the number of corrected labels from different classes. However,

there are metrics like Geometric Mean and Youden's Index which use both columns but not sensitive to class imbalance [14].

The other types of comparison metrics are receiver operating characteristic (ROC curve) and area under ROC (AUC) are commonly used graphical metrics. ROC is obtained by plotting true positive rate on y-axis and false positive rate on the x-axis. Its objective is to select the best threshold value by varying threshold to improve the performance of the classifier. For the best classifier the plot of this curve will be near the upper left coordinates. the area under ROC (AUC) is the arithmetic mean of true positive rate to true negative rate [26]. AUC metric is always bounded between zero and one and always greater than 0.5 [14].

There are data level and algorithm level techniques to enhance classifiers' accuracy. Algorithm level methods use ensemble methods. Data level methods consider the class distribution. Most of the real-world data set are imbalanced (members of one class outnumber the other). Classification error caused by over-fitting the majority class in the dataset. So, class imbalance is one factor that negatively affects the performance of classifiers. There are sampling techniques like (minority oversampling or majority under-sampling) to balance classes in a dataset. However, oversampling creates synthetic data that resembles minority class instances without considering the distribution of majority classes. Majority under-sampling also removes important information from the dataset. Many scholars have studied how the impact of class imbalance, over/under-sampling techniques on classifiers' accuracy. Moreover, from application-driven nature of machine learning and domain-specificness of class imbalance problems, the importance of realization of exploring ideographic solutions and understanding more knowledge on the domain was underlined in [27].

Chapter 4

EXPERIMENT SETUP

There are different methodological approaches in the process of achieving research objectives and answering the research questions. So, starting with the data collection techniques, the methodological approaches followed by this research are discussed in this chapter. The chapter further continues to discuss the underlying concepts in the feature construction step and data preprocessing measures undertaken. Outlier detection and removal procedures which are parts of data preprocessing in data mining researches are explained in this chapter. Furthermore, the experiment setup followed is also explained.

4.1 Data collection

Data collection is the first and most generic step in a data mining project [28]. Users' profile represented based on information from their usage record [17], [18], [27]. Call detail records (CDR) are the source of information which represent the number of minutes/volume of calls by call type, source/destination network, and another call-related information. This CDR information represents the complete service usage profile of users [1]. There are data mining studies which have used CDR for the classification of the cross-selling market [6] [2] [4].

The CDR data used for this study was generated by CBS (Convergent Billing System) during customers service usage time. As discussed in section 3.1 designed experiment and observational approach are possibilities in data collection process [20]. Since CDR record is generated automatically during customers service usage, there is no possibility to influence the data rather than collecting. One of the columns in the CDR data is MSISDN, a 10-digit subscriber identification number

with a format like 251-9xx-xxxxxx. Samples are selected using this MSISDN number by ignoring the first three digits (country code for Ethiopia) and taking the next three digits as prefixes. About 92 distinct types of prefixes taken for sampling. Samples selected based on these 3-digit prefixes and for the 3 types of services (voice, data, and SMS) (Figure 4.1).

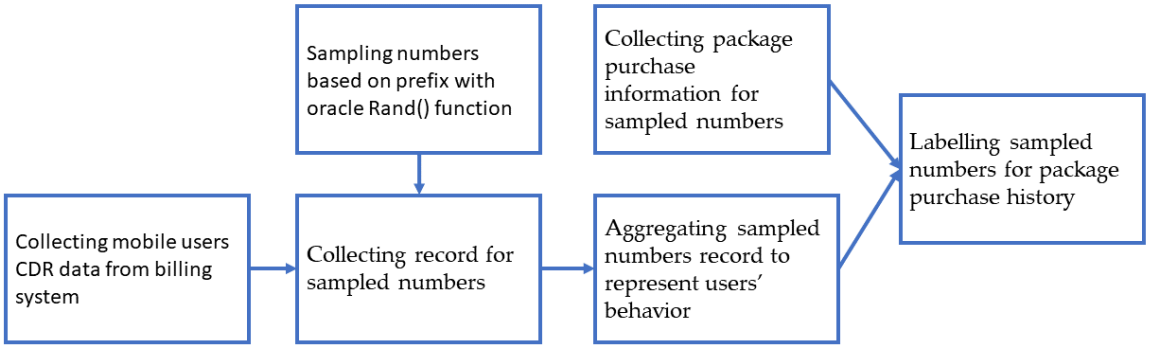

FIGURE 4.1: Conceptual Diagram for data collection, aggregation and labelling process

For convenience in the processing of large Data, sampling, collection and aggregation of data was done in three steps.

- The first step was identifying MSISDN numbers using a stratified random sampling technique for the available 92 prefixes. Using oracle random function, which uses the pseudo-random function of seed 1, about 10,000 MSISDN numbers from each prefix were selected and stored into SAMPLED_NUMBERS database table.
- The next step was selecting records for the month of June 2019, based on the SAMPLED_NUMBERS table and populating to local repository.

- The last step was an aggregation of monthly records and labelling aggregated data using already collected target numbers information (MSISDN number, date of purchase and type of package purchased).

The number of distinct records sampled numbers for a one-month period and three types of services are as shown in Table 4.1 below.

TABLE 4.1: Number of distinct records for the three datasets

Data set for service type	**Number of instances**
Voice	967,745
Data	958,427
SMS	986,706

The data used for this research was collected from two sources that are package buyers' Information from report and customers usage-data form CDR generated by CBS (Convergent Billing System). Package buyers report consists of 8 columns (MSISDN, Account type, Customer type, Mobile type, package purchase date, package volume, total package purchase, and total revenue) and usage record CDR contains 33 columns (Appendix A).

CDR records and package buyers Data were uploaded into local data repository and further analysed with SQL query to map and label package user's usage record with the type of packages purchased, and date of purchase information's.

4.2 Data aggregation and feature construction

Voice, data, and SMS usage CDR data used in this study has been collected from CBS (Convergent Billing System). This CDR contains about 33 columns. Based on their purpose on the CDR, these columns can be grouped as Traffic generator and receiver identity, traffic source and destination (location), traffic start and end time, data upload and download volume, billing Information, and service charging amount.

Except for the billing information category of CDR column the rest contains service usage volume, spatial, and temporal information of service users. Raw features in

this record were not applicable to use for the data mining process [13]. Data features which are grouped like traffic generator and receiver identity, billing information, and traffic source and destination location were not selected for use in this data mining process. Others like call start and end time, data upload and download volume, and SMS usage need to be aggregated to use them in the data mining process.

Usage record collected from CBS system for the duration of one month aggregated for each user. In this aggregation process, 30 data columns are generated from the 8 columns of raw CDR. Usage record contains usage-data for 3 types of services (voice, data, and SMS). Since different price schemes are applied for these services, different features derived from the raw features with consideration of user's reaction to service fee and convenience of service usage time. During construction of features voice service usage patterns during peak and off-peak hours, calls made and received, and the amount of fee paid for the service was used. Data service usage patterns are also aggregated based on upload and download volume, payment for the service, and users' time-period preference (during morning, afternoon, evening, or after midnight) were used. In SMS usage the number of sent and received SMS, receiving and reply pattern to bulk SMS were used.

As discussed in section 3.1 data aggregation done in consideration of representing user's behaviour in voice, data, and SMS usage pattern [21]. Since voice service price plan is different for peak and off-peak hours, users' voice service usage behaviour represented in consideration with user's tendency (usage frequency and duration) towards peak and off-peak hours.

Data service price plan is flat. So, users were not expected to select a time period for the sake of minimizing their expense rather they select convenient time due to their business need and ease of use (network performance on users' location). As a result, time preference for data service defers for each user. With this consideration data aggregation done by dividing a day into four time periods (00:00 - 06:59, 07:00 - 11:59, 12:00 - 17:59, 18:00 - 23:59) and aggregation of users download-volume and upload-volume done in these time periods. SMS usage is aggregated by counting the number of SMSs sent, received, bulk SMS responded and received. As shown in figure 4.2 a total of 31 features constructed using data collected from the two sources. Among these 31 features, 30 of them are feature (predictor) variables and one class variable which contains users' purchase response for one of the three types

of service packages. The detail of constructed features is depicted in Appendix A.2 in detail.

TABLE 4.2: Constructed data features for each service

Voice usage-dataset features	**Data usage-dataset features**	**SMS usage-dataset features**
Total service usage fee	Upload traffic volume	Sent SMS count
Outgoing call duration	Upload traffic fee	Received SMS count
Outgoing call frequency	Download traffic volume	Bulk SMS response count
Incoming call duration	Download traffic fee	Bulk SMS received count
Incoming call frequency	Morning time upload volume	
Off-peak usage frequency	Morning time download volume	
Off-peak usage duration	Afternoon time upload volume	
Off-peak usage fee	Afternoon time download volume	
Peak hour usage frequency	Evening time upload volume	
Peak hour usage duration	Evening time download volume	
Peak hour usage fee	After midnight download volume	
International call usage frequency	After midnight upload volume	
International call usage duration		
International call usage fee		

After aggregating a total number of rows in a raw data, 967,745 for voice package buyers dataset, 873,793 for data package buyers dataset, and 801,145 rows for SMS package buyers data obtained. Each row in these datasets represents service usage behaviour and response for service package advertisements. The total number of

mobile package buyers for the month of June 2019, was 2,407,956. By mapping the package buyers service number to sampled numbers usage record using SQL, usage record of data package buyers flagged (positive class or buyer class) and those which lacks this flag marked as package non-buyers (negative class). The number of rows, positive class, and negative class in each dataset was as shown in Table 4.3.

TABLE 4.3: Structure of the three datasets used in the experiments

Data set	Total number of rows	Number of positive class instances	Number of negative class instances
Voice package buyers	967,745	321,599	646,146
Data package buyers	873,793	212,056	661,743
SMS package buyers	801,145	95,775	705,370

Based on the three types of service packages, buyers of these packages categorized into three. The classification problem considered as three binary-class classification problems. Rows in each dataset represent users of services and columns represent customers' usage patterns for the specified column entity which are called explanatory variables or features. The last column of the three datasets represents user's belongingness to either "buyers" or "non-buyers" class. In this column, each of the rows is marked "Yes" if the customer has bought a package for the specified service or "No" if the user does not involve in buying these packages.

Usage record of package buyers for SMS, data, and voice packages marked accordingly. The column "PKG_HIST" in each row of the dataset, has a value "Yes" for a user which have bought voice package in the month and "No" otherwise. In similar fashion labelling done for Data and SMS buyers' datasets (Figure 4.2).

ICF	OFPF	PF	OFPFEE	PFEE	OFPCD	PCD	OGD	UTR	DTR	BND	BNU	AFNU	AFND	BMND	BMNU	AMNU	AMND	SENDFR	PKG_HIS
1	14	259	0	1103.05	0.24	0.63	0.93	21.99	1015.71	182.85	1015.71	96.38	8655.93	11985.39	60.34	25849.44	1711.71	1	Yes
1	14	259	0	1103.05	0.24	0.63	0.93	21.99	1015.71	182.85	1015.71	96.38	8655.93	11985.39	60.34	25849.44	1711.71	1	Yes
1	14	259	0	1103.05	0.24	0.63	0.93	21.99	1015.71	182.85	1015.71	96.38	8655.93	11985.39	60.34	25849.44	1711.71	1	Yes
14	14	102	0	2451.34	0.48	2.1	2.84	130.8	4922.89	1799.57	4922.89	704.18	47452.6	47614.56	481.98	622600.3	37229.28	2	Yes
8	12	240	0	7133.22	0.85	1.95	3.11	35.06	2351.99	86.51	2351.99	1.66	813.57	20337.78	31.17	15491.88	6220.03	152	No
7	16	109	0	734.52	0.35	0.78	1.21	29.87	2480.42	198.91	2480.42	16.85	3208.44	15579.79	143.52	7602.45	3227.11	63	Yes
7	16	109	0	734.52	0.35	0.78	1.21	29.87	2480.42	198.91	2480.42	16.85	3208.44	15579.79	143.52	7602.45	3227.11	63	Yes
	13	76	0	0	0.58	2.43	3.2	0.18	0	0.37	0							1	No
9	11	122	0	2273.01	0.27	2.3	2.9	0	0	0	0							33	No
2	11	70	0	0	0.43	1.4	2.17	0.23	100.1	0	100.1							2	No
	[illegible]	123		0	0.29		[illegible]	[illegible]	0	0.01								2	No

FIGURE 4.2: Partial view of aggregated data set

4.3 Data cleaning and class balancing

Data cleaning is one of the important processes in data mining. Erroneous values in the dataset adversely affect the performance of classifiers. These erroneous values appear to exist due to various reasons. As described in section 3.1 of this document there are techniques to handle data errors.

After aggregation of usage CDR data, the resulting dataset has null values. Different methodologies to fulfil missing values (null entries) can be used. Since features used in this research are constructed by aggregating usage records, the resulting dataset has null values and outliers which results from customers usage pattern. These outliers, especially null values were not the results of error in data collection process. Rather each null value represents user's behaviour in each row. Whenever a null value appears in a specific row and column it reflects absence of usage for the specific service. Outliers as extreme value are also results of customers' usage pattern. Generally, the dataset is free of errors caused by external factors like data entry and failure of raw data source. Exploratory data analysis was done on the three datasets before and after data cleaning. Based on results of exploratory data analysis before data cleaning, 14 of the extracted features have null(zero) values for more than 94% in each column. With this fact, 14 features removed from each of the three datasets. The correlation analysis after data cleaning shows us the predictor variable (PKG_HIST) has a non-linear correlation with the input variables (Figure 4.3). This is also true for Voice and SMS packages buyers datasets (Figure 4.4, and Figure 4.5).

Classifiers accuracy highly affected by extreme and null values. So, it is important

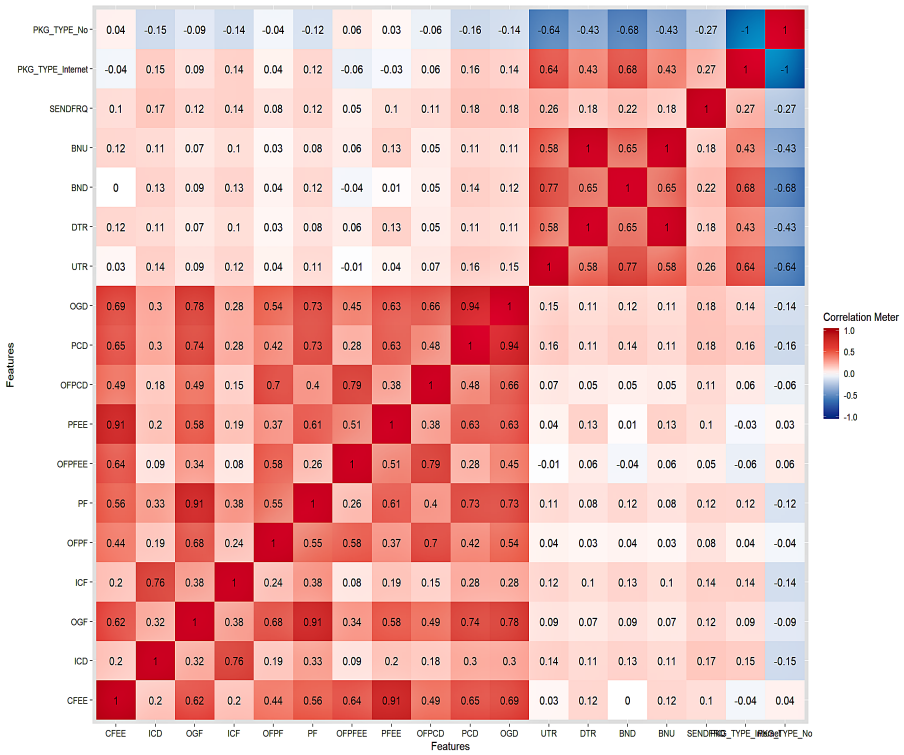

FIGURE 4.3: Correlation analysis for data package buyers' dataset. the first two upper rows ("PKG_HIST_Yes" and "PKG_HIST_No") shows correlation with colour heat map the brighter the colour the lower the correlation the darker the colour the higher the correlation

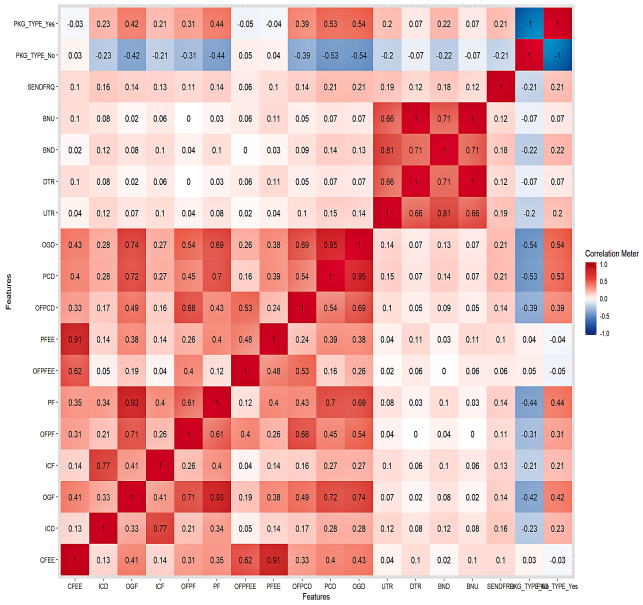

FIGURE 4.4: Correlation analysis for voice package buyers' dataset.

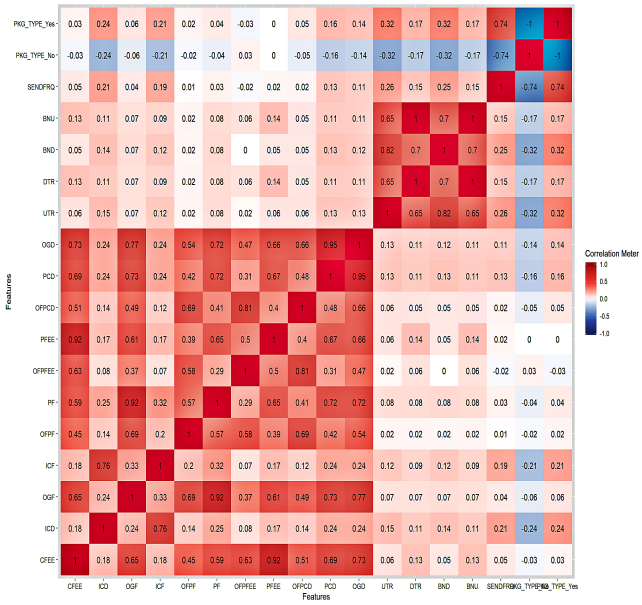

FIGURE 4.5: Correlation analysis for SMS package buyers' dataset.

to clean the datasets from these values. Replacing null values with value 0, removing outliers and extreme values were the data cleaning steps undertaken in this experiment.

Based on interquartile range, outliers and extreme values were marked and removed from the data set. Values between the first and third quartile kept and the rest that lies out of these quartiles marked and removed as outliers [22]. This reduced the number of rows in the resulting dataset. The number of positive-class and negative-class instance were highly unbalanced. For data sets with very high class-imbalance, negative-class instances were re-sampled keeping the positive-class instances untouched. The resampling technique used was simple random sampling without substitution. The reason for consideration of re-sampling was to minimize the computation time, especially during class balancing.

4.4 Classification experiment setup

This experiment used version 3.8.3 of Waikato Environment for Knowledge Analysis, which is developed by the University of Waikato, New Zealand. Using this general public license software, the five algorithms which were selected from each of classification techniques Naïve Bayes, Single layer Perceptron(Neural network), SVM with polynomial kernel, KNN (K-nearest neighbour) and Decision Tree (J48) with C4.5 algorithm are tested with ethio telecom dataset.

In Naive Bayes classifier, numeric precision values are chosen during the training phase and were not updatable [29].

Perceptron with single hidden layer was used. The number of input is equal to the number of data features (16 input features) and the number of output was 2 for buyer and non-buyer. It uses sigmoid function because of the categorical type class variables (Yes, and No for each service package buyers) in the three datasets.

Support vector classifier with polynomial kernel was also used. This classifier implements John Platt's sequential minimal optimization algorithm for training a support vector classifier. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. In that case, the coefficients in the output are based on the normalized data, not the original data.

The number of nearest neighbours selected based on testing for odd-numbered K values ranging from 1 up to 19 [20]. The best accuracy was observed with K values 11 for voice package, and 7 for data and SMS package buyers datasets. Further increment of K values results in decrease of accuracy of the classifier. Percentage accuracy for different K values of the three datasets is shown in Table 4.4 below.

TABLE 4.4: Percentage correct values for each nearest neighbour variable (k) of KNN classifier for the three datasets.

	K	1	3	5	7	9	11	13	15	17	19
Dataset	Voice	80.42	83.11	83.66	83.92	83.95	84.05	83.72	83.66	83.56	83.48
	Data	75.13	75.91	76.12	76.45	76.34	76.22	76.23	76.12	75.96	75.77
	SMS	89.48	90.24	90.53	90.6	90.57	90.45	90.45	90.46	90.44	90.4

From tree-based classification technique Decision tree (j48) used with C4.5 pruning algorithm. Tree induction is learned from training tuples by recursively partitioning training set into smaller subsets [22].

A clean and balanced dataset is supplied to these five classifiers. In these dataset class representation of each service package buyers' row (positive classes) was marked as "Yes" and each of non-buyers (negative classes) row marked as "No" in the dataset column that contains package purchase history (PKG_HIST).

In the reviewed researches [4] and [10] have used dataset with 45,211 instances from UCI (University of California, Irvine) machine learning repository of Portuguese bank marketing data. Furthermore, [6] have used 3,080 instances of Korean telecom customers data and in [7] statistically generated dataset with 6,000 instances of which 3,000 instances used for training and 3,000 instances for testing. The consideration of the class imbalance has not been explained in all researches except [4] which focuses on the effect of the ratio of negative and positive classes on classifier performance. Most of the instances used in these researches are small.

On the contrary, the datasets used in this research have enough number of instances, each dataset split into two (training 66% and test 34%). Training set supplied to the five algorithms and the resulting model was evaluated with the remaining test set.

Chapter 5

RESULT ANALYSIS

In this chapter, data preprocessing and performance results of classifiers are explained with four performance measures: Accuracy, precision, recall and area under ROC. The effect of data prepossessing on the number of instances of each dataset shown and the number of instances in the confusion matrix are also explained. Performance analysis was presented for the three types of datasets (package buyers for voice, data, and SMS services).

5.1 Data preprocessing results

In data preprocessing, instances that lie in the second and third interquartile range have been preserved and the rest has been removed as outliers and extreme values. These resulted a 58.79% in voice package buyers dataset,72.20% in data package buyers dataset, and 84.14% in SMS package buyers dataset decrease in number of instances in the final clean and class-balanced dataset. The number of positive and negative instances in each dataset before class-balancing were highly unbalanced. This was caused due to small number of package users (2,407,956) for the month of June 2019 when compare to the total number of mobile service users (more than 41 million [30]).

The number of negative class instances (package non-buyer instances) outnumber the positive-class instances (package buyer instances). Positive class instances are minority in number. Random down-sampling was done on the three datasets to decrease the number of negative class instances. The main reason for down-sampling was not only to decrease the number of synthetic instances created to balance classes but also to decrease the need for high computational capacity. After

down-sampling the negative class instances, additional positive class instances created using class balancing technique called synthetic minority oversampling technique (SMOTE) on each of the three datasets and class-balanced dataset obtained. The number of instances in each of the three datasets initially and after cleaning and re-sampling shown in Table 5.1. The number of instances in the sixth column of Table 5.1 contains equal half for positive and negative instances.

TABLE 5.1: Number of instances in the three datasets before and after data cleaning.

Dataset	Total number of instances		Number of instances cleaned dataset		Number of instances after cleaning and balancing
	Before cleaning	After cleaning	Positive	Negative	
Voice	967,745	664,454	164,308	500,146	398,773
Data	873,799	607,259	83,108	524,151	242,905
SMS	801,145	480,687	95,775	384,912	127,090

5.2 Classifier performance analysis results

Performance of the five classifiers has been tested with datasets of the three services (Voice, Data, and SMS). In these three datasets users' usage behaviour represented for the three services as discussed in section 4.2. Considering metrics other than accuracy (precision, recall, and AUC) is important to understand whether there is the effect of predicting only to the majority class or not. During the comparison of those performance measures (accuracy, precision, recall and AUC) Wekas' paired t-test corrected with a two-tailed 95% confidence interval was used.

5.2.1 Result analysis for voice package buyers' dataset

The dataset used in this experiment was the CDR aggregated for the three types of services and target groups in this dataset are customers who bought a voice package. Using this dataset, performance measures accuracy, precision, recall and area under ROC curve for the five selected algorithms is presented in Table 5.2. Decision Tree (J48) algorithms scored high performance in the three selected performance measures accuracy, precision, and recall. However Neural network classifier has a

higher performance than Decision Tree in area under ROC curve measure. The three classifiers other than Naive Bayes are almost equal in all performance measures.

TABLE 5.2: Performance of classifiers on voice package dataset

Classifier	Accuracy	Precision	Recall	AUC
Naïve Bayes	0.74	0.75	0.74	0.84
SVM	0.9	0.91	0.9	0.9
NN	0.91	0.91	0.91	0.95
KNN	0.9	0.9	0.9	0.94
J48	0.92	0.92	0.92	0.93

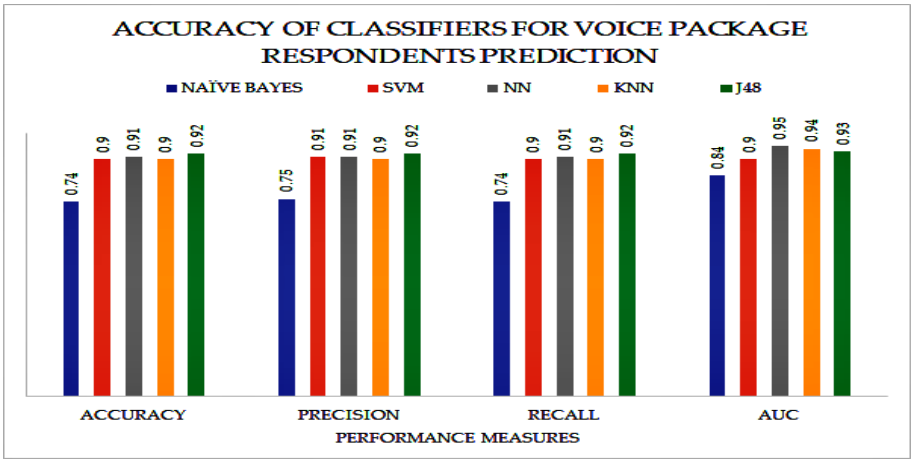

FIGURE 5.1: Accuracy of classifiers with four measures for SMS package buyers dataset. Vertical axis accuracy in percentage and horizontal axis shows accuracy measures for the five classifiers.

The other comparison measure is confusion matrix values which are presented in Table 5.3. The result of classifiers in terms of the number of instances which are classified as true-positive(positive instances classified as positive), true-negative(negative instances classified as negative), and the errors of classifiers in terms of false-positive and false-negative. False-positive and false-negative are instances which are wrongly classified as positive while belonged to negative classes and vice versa.

Support Vector Machine classifier classified the highest number of true-positive instances than the other four classifiers. Naive Bayes classifier has a comparatively higher number of false-positive and false-negative (incorrectly classified) instances (Figure 5.2).

TABLE 5.3: Confusion matrix of classifiers

	Naïve Bayes	**SVM**	**NN**	**KNN**	**J48**
TP	55767	65315.4	63370.3	61619.5	62582.3
TN	44544.2	56974	60387	60349.6	61770.3
FP	23264.2	10834.4	7421.4	7458.8	6038.1
FN	12007.4	2459	4404.1	6154.9	5192.1

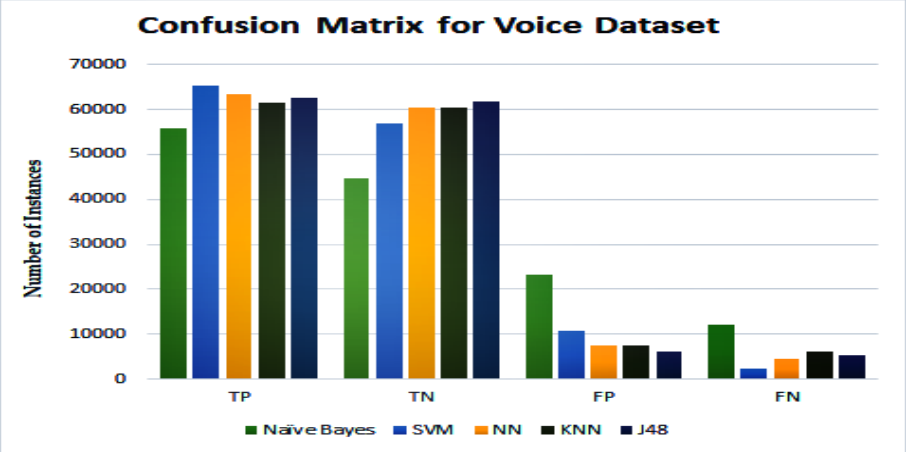

FIGURE 5.2: Confusion matrix of the five classifiers with voice package buyers dataset. Vertical axis shows number of instances and horizontal axis shows confusion matrix elements.

5.2.2 Result analysis for data package buyers' dataset

Data package buyers' usage record was used for analysing performance of five classifiers. In this experiment also the four classifiers (SVM, Neural Network, KNN and Decision tree) other than Naive Bayes have exhibited good performance. However, the performance of Naive Bayes classifier was not too bad. Accuracy of Classifiers

other than Naive Bayes scored higher than 94%. SVM and KNN classifiers scored equal accuracy whereas Neural Network was the highest with 96% and Decision tree was the next with 95%. For the three measures precision, recall and AUC Neural Network was with the highest performance. The least performances with all performance measures were recorded by Naive Bayes classifier (5.4).

TABLE 5.4: Performance of classifiers on data package dataset

Classifier	Accuracy	Precision	Recall	AUC
Naïve Bayes	0.83	0.84	0.83	0.92
SVM	0.94	0.95	0.94	0.94
NN	0.96	0.96	0.96	0.98
KNN	0.94	0.91	0.91	0.96
J48	0.95	0.95	0.95	0.97

These differences in performance of classifiers so far explained using results shown in Table 5.4 is graphically represented in Figure 5.3. From the confusion matrix in Table 5.5, SVM classifier identified the highest number of true-positive instances. The least number of false-positive and false-negative (incorrectly grouped instances) identified by Neural Network classifier. Furthermore, the highest number of errors exhibited by Naive Bayes classifier in this dataset.

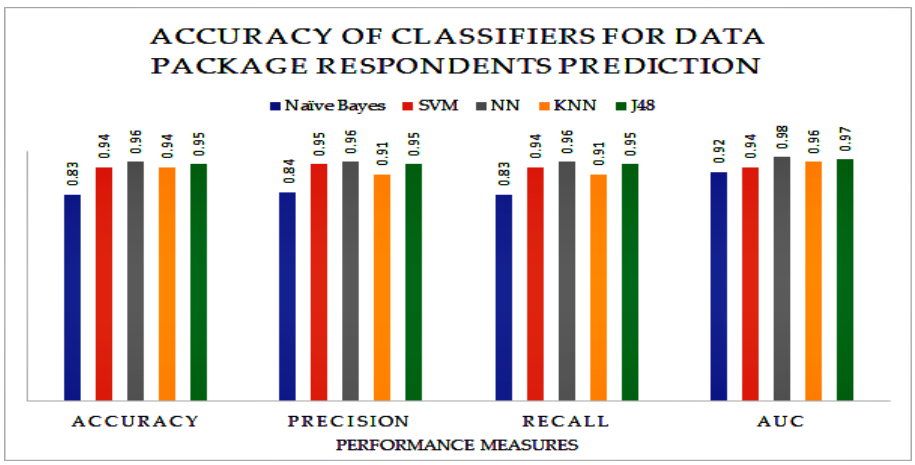

FIGURE 5.3: Accuracy of classifiers with four measures for data package buyers dataset. Vertical axis accuracy in percentage and horizontal axis shows accuracy measures for the five classifiers.

TABLE 5.5: Number of instances in the confusion matrix of the five classifiers.

	Naïve Bayes	**SVM**	**NN**	**KNN**	**J48**
TP	21811	24085	23862	22827	23702
TN	19438	22723	23850	22388	23564
FP	5436	2151	1024	2486	1310
FN	2868	594	817	1852	977

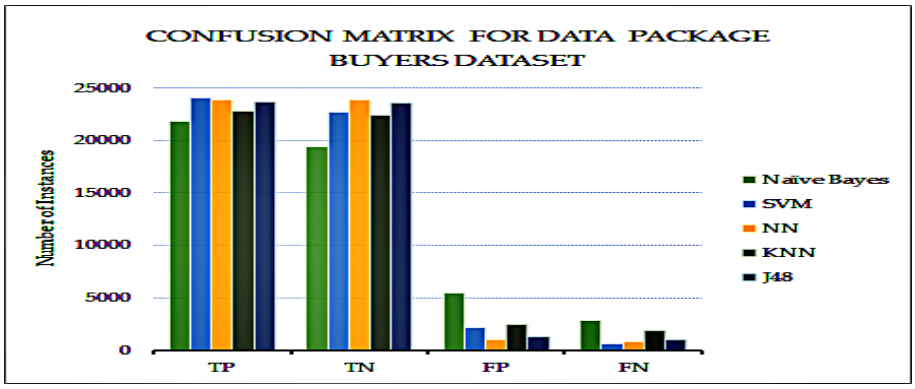

FIGURE 5.4: Confusion matrix of the five classifiers with data package buyers dataset. Vertical axis shows number of instances and horizontal axis shows confusion matrix elements.

5.2.3 Result analysis for SMS package buyers' dataset

In a similar fashion as voice and data package buyers' dataset, SMS package buyers' dataset also labelled for service users who bought SMS packages. Five classifiers (Naive Bayes, SVM, Neural Network, KNN and Decision Tree) also tested with this dataset for their performance in four performance measures. Accuracy of Neural Network and Decision Tree (j48) classifiers was the highest i.e. 95%. The second-best accuracy (94%) was obtained by SVM classifier whereas KNN and Naive Bayes are the least accordingly (Table 5.6). Neural Network classifier scored-highest performances also in performance metrics precision, recall, and AUC. In AUC metric Neural Network scored the highest performance which is 99% and this is clearly visible in Figure 5.5. The least performance in accuracy precision and recall were scored by Naive Bayes classifier. However, Naive Bayes has better AUC measure than SVM.

TABLE 5.6: Performance measure of classifiers based on the four measures accuracy, precision, recall and AUC using SMS package buyers dataset.

Classifiers	Accuracy	Precision	Recall	AUC
Naïve Bayes	0.88	0.88	0.88	0.95
SVM	0.94	0.94	0.94	0.94
NN	0.95	0.95	0.95	0.99
KNN	0.93	0.93	0.93	0.96
J48	0.95	0.95	0.95	0.97

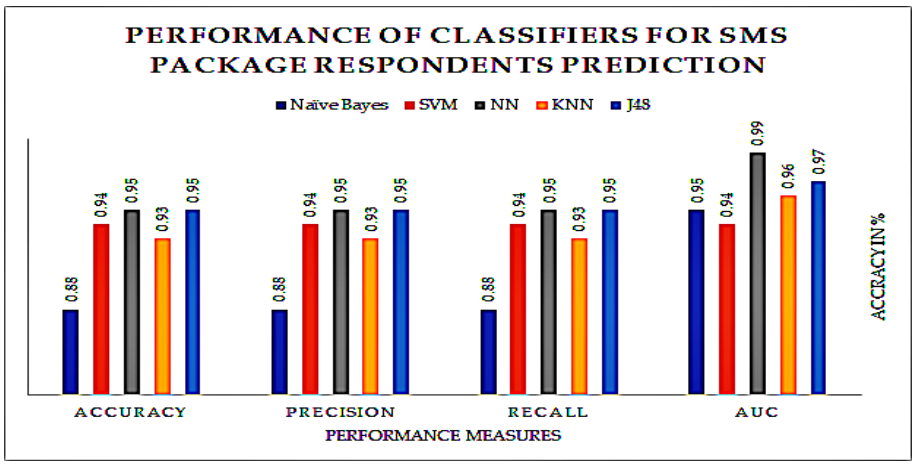

FIGURE 5.5: Accuracy of classifiers with four measures for SMS package buyers dataset. Vertical axis accuracy in percentage and horizontal axis shows accuracy measures for the five classifiers.

Based on confusion matrix for the five classifiers, Neural Network has the highest number of correctly classified instance as well as the lowest number of incorrectly classified instances. SVM and Decision Tree classifiers are the next in classifying instance correctly. Naive Bayes classifier has the highest number of incorrectly classified instances with this dataset also (Table 5.7 and Figure 5.6).

TABLE 5.7: Confusion matrix for SMS package buyers dataset

	Naïve Bayes	**SVM**	**NN**	**KNN**	**J48**
TP	18770.6	20585.3	20278.1	20152.2	20258.3
TN	19241.8	20187.8	20661.6	19919.8	20607.5
FP	2383.2	1437.2	963.4	1705.2	1017.5
FN	2815.2	1000.5	1307.7	1433.6	1327.5

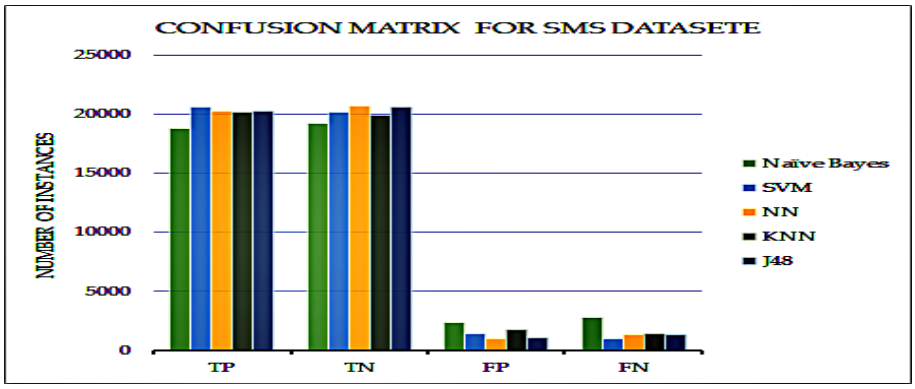

FIGURE 5.6: Confusion matrix of the five classifiers with SMS package buyers dataset. Vertical axis shows number of instances and horizontal axis shows confusion matrix elements.

5.3 Summary of results

As discussed in section 3.1 raw data should be converted to a more informative data using aggregation techniques. In the process of extracting users behaviour from usage records, raw data has been changed to more informative format. Among the six groups of CDR columns traffic start and end time, data upload and download volume, and service charging amount column groups have been used to construct predictor variables. This process mainly focused on users reaction to service charge and parts of the day (morning, afternoon, evening, and late-night). So, CDR columns under these groups contain more useful information in the process of classification of the three service packages buyer customers.

Information in voice package buyers dataset mainly extracted by grouping parts of the day to peak and non-peak hours, and by examining the usage duration and frequency in these time periods. Data usage behaviour was also extracted from raw

data by dividing time of the day into four parts and querying users' download and upload volume in these parts of the day. Generally, the overall information about users interaction with the services can be obtained from call start and end time, call duration, call fee, upload traffic, and download traffic columns by systematic aggregation of this raw data columns.

Performance comparisons with selected accuracy measures and confusion matrix instances have been discussed so far in this chapter. These performances haven tested with three datasets voice, data and SMS package buyers. Since each dataset has its own structural differences with the number of instances for predictor and target class, the performance of classifiers has also exhibited different characteristics. The correlation between predictor classes and the target class is mostly non-linear as shown in section 4.3. The four classifiers other than Naive Bayes have comparatively similar performance. These four classifiers (SVM, NN, KNN, and DT) with dataset of non-linear correlation performed better. Especially, SVM classifier performance was enhanced due to the polynomial kernel as discussed in section 3.2 of this document. In the next subsections performance of classifiers is ranked and depicted graphically. Ranking figures: Figure 5.7, Figure 5.9, and Figure 5.11 show classifiers rank in the vertical axis and performance measures used for ranking in the horizontal axis. And figures Figure 5.8, Figure 5.10, and Figure 5.12 show rank of classifiers based on number of instances in the confusion matrix in the vertical axis and classifiers in the horizontal axis.

According to literature, the Naive Bayesian classification technique is more suitable for discrete datasets or categorical features. Even though the three datasets used in this research are continuous and the degree of association (bivariate correlation between predictor variables and the class variable) is not strong, classifiers which are more suitable for linearly correlated datasets like SVM has increased performance due to the kernel trick which transforms the non linear data space to linear one.

Among these four groups of classifiers, Neural network is the best classifier for classification of voice package buyers and SMS package buyers. Since it is more capable of identifying more true-positive instances and also have better performance in terms of performance metrics such as accuracy, precision, recall, and AUC with these datasets.

In the case of Data package buyers classification, J48 have better performance with the dataset used in this experiment.

5.3.1 Result with voice package buyers' dataset

With voice package buyers' dataset classifiers performances ranked. In this ranking, Decision Tree (J48) and Neural Network are the first and second respectively in three performance measures. Support Vector Machine and K-nearest neighbour are the third performers with accuracy measure but performance of SVM is equal with Neural network in precision. K-nearest neighbour is the fourth performer in precision. While Naive Bayes have least performance with all measures in the group. How ever, there is no highly exaggerated performance difference between the four classifiers (Decision Tree, SVM, NN, and KNN) except for AUC measure which ranges between 90% and 95% (Figure 5.1). Naive Bayes classifier is the least performer in all the four measures (accuracy, precision, recall, and AUC) among the classifiers compared with this dataset. It is the first ranked in the two classification error instances that are false-positive and false-negative. The major aim of classification is to get the highest number correctly classified instances (true-positive and true-negative) and to minimize error (Figure 5.7).

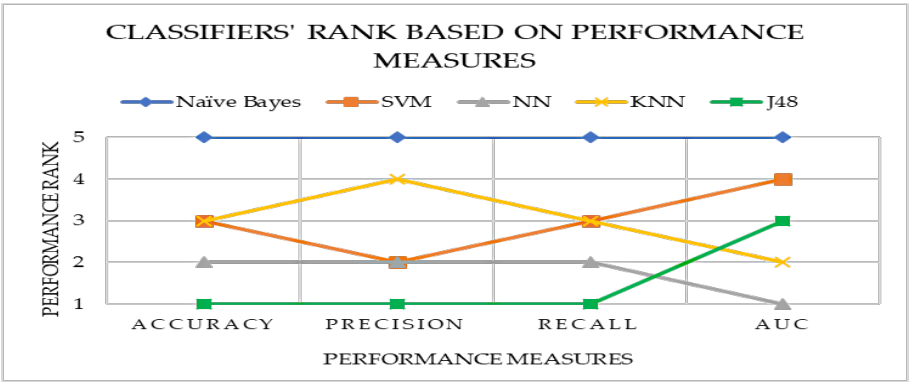

FIGURE 5.7: Ranking of classifiers based on performance measures accuracy, precision, recall and AUC for voice package buyers dataset

When classifiers are compared with the number of instances in the confusion matrix, Support vector machine ranked first among the group by classifying a high number of true-positive instances, but it is in the second rank with misclassified or false-positive instances. Even though the number of true positive instances for

Neural network classifier is less than the number for SVM, Neural network has identified a greater number of true-negative instances than SVM (Figure 5.8).

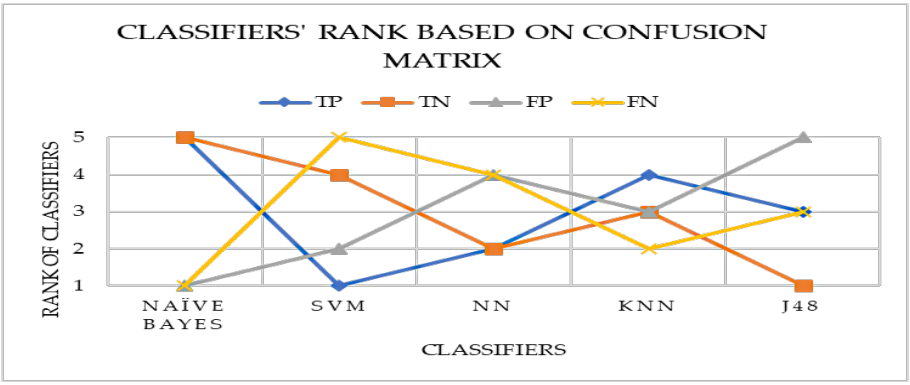

FIGURE 5.8: Ranking of classifiers based on the number of instances in confusion matrix of the five classifiers with voice package buyers dataset. Classifiers with highest the rank for both true positive and true negative have the better performance.

5.3.2 Result with data package buyers' dataset

The five categories of classifiers performance also compared with data package buyers dataset. In this comparison, Neural network classifier and Decision tree(J48) are the first and second in all the four performance measures. SVM and KNN are equally in the third rank with accuracy measure while SVM is equally the second performance score with Decision tree classifier in precision metric. In this dataset also Naive Bayes performance is the least of all classifiers tested with this dataset (Figure 5.9).

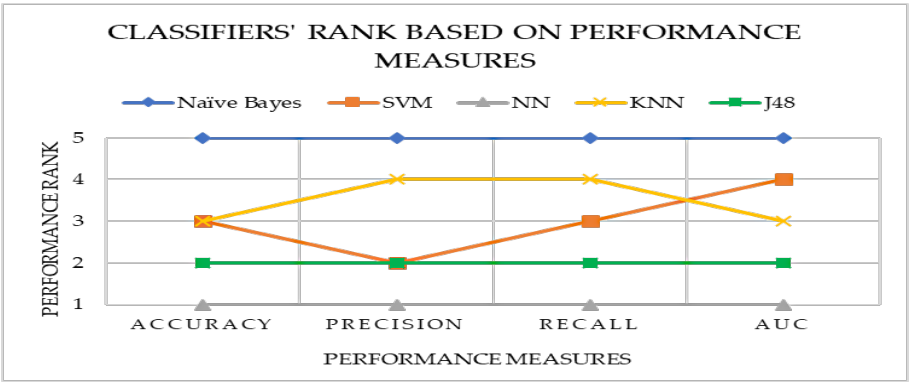

FIGURE 5.9: Ranking of classifiers based on performance measures accuracy, precision, recall and AUC

From the comparison of classifiers based on the number of instances in the confusion matrix, Naive Bayes classifier has the highest number of false-positive and false-negative instances than the others. Support vector machine has the second largest number of true-positive instances in the group. It has also the third-largest number of true-negative and false-positive also. The Neural network has the first-largest number of true-negative instances in the group and the third-largest number of true positive instances. The number of false-positive instances has also well minimized by this classifier so that it has the least number of such instances. KNN classifiers' rank in the number of true-positive and true negative instances is the fourth in the group and next to Naive Bayes it is the second least performer because of its largest number of error instances (false-positive and false-negative) among these group of classifiers and this dataset. Decision tree(J48) has the first best performance in the group by identifying the largest number of true-positive instances and also the second-largest number of true-negative instances (Figure 5.10).

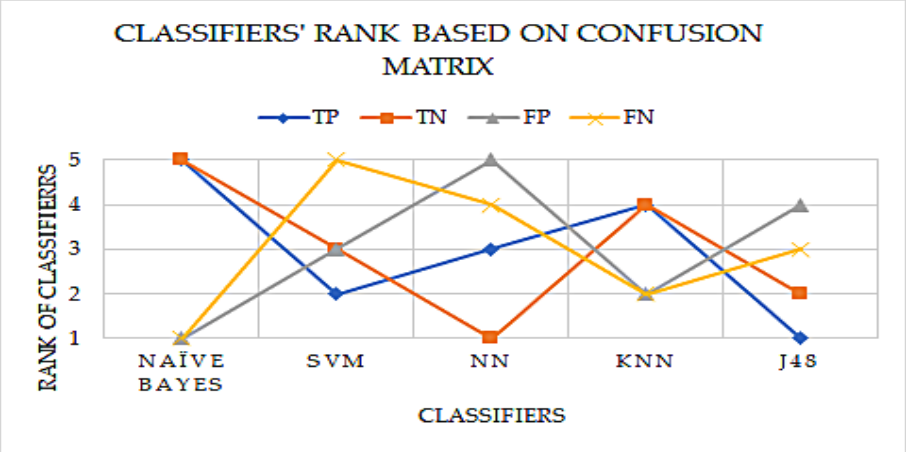

FIGURE 5.10: Ranking of classifiers based on the number of instances in confusion matrix of the five classifiers with data package buyers dataset. Classifiers with highest the rank for both true positive and true negative have the better performance.

5.3.3 Result with SMS package buyers' dataset

In SMS package buyers' dataset Decision tree(J48) and Neural network classifiers have both the first rank in the three measures accuracy, precision, and recall whereas Decision tree is the second-highest percentage rank in AUC measure. Support vector machine also has the third rank in the three measures other than AUC which is

the least performance rank in AUC measure. Similar fashion was also true for both KNN the fourth in the group and Naive Bayes the least. But these two classifiers also exhibit a better performance rank than SVM (Figure 5.11).

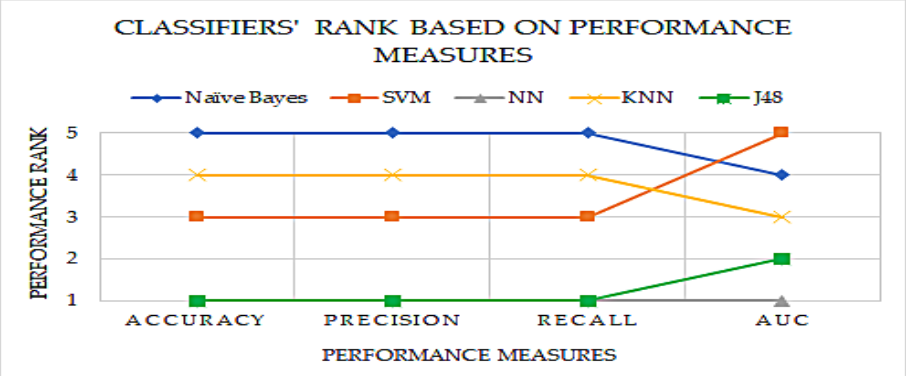

FIGURE 5.11: Ranking of classifiers based on performance measures accuracy, precision, recall and AUC for data package buyers dataset.

The second ranking for this dataset is also based on the number of instances in the confusion matrix. In this ranking, Neural network classifier has comparatively better performance than the others by having the first largest number of true-negative instances and the second in true positive instances. Support vector machine has exhibited the first largest number of true positive instances in the group, but it was also the third-largest number of true-negative as well as false-positive. KNN has the poorest performance next to Naive Bayes by exhibiting the largest number of error instances (Figure 5.12).

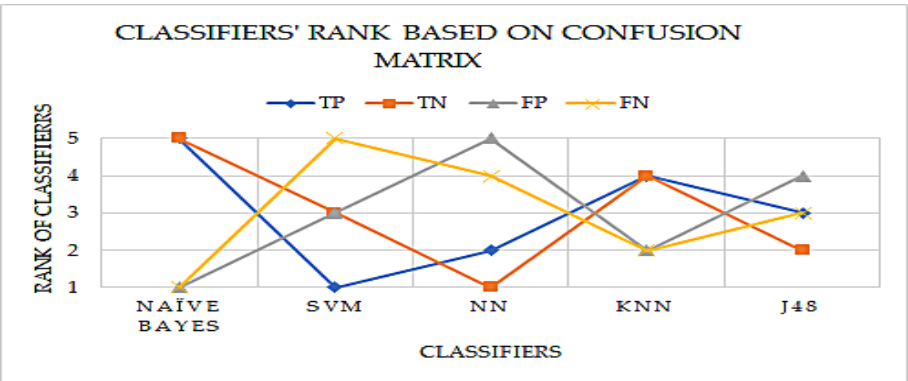

FIGURE 5.12: Ranking of classifiers based on the number of instances in confusion matrix of the five classifiers with SMS package buyers dataset. Classifiers with highest the rank for both true positive and true negative have the better performance.

Chapter 6

CONCLUSION AND RECOMMENDATION

Conclusion based on experiment results analysed in the previous chapter are provided and recommendation for future studies suggested.

6.1 Conclusion

The two major interests of this research were to identify the data features which have more information about the interaction between package service market and its users, and to select classification model which predicts prospective respondents of the cross-selling market with better accuracy.

Based on the experiments undergone the features constructed from raw CDR columns call start time, call end time, service fee, data upload volume, download volume, and SMS sent have more information about the interaction of package service users and the market. The data features used in the three datasets are constructed from these raw CDR columns by focusing in representation of user's response to service fee in different price plan, convenient data service usage time, and frequency of SMS communication. These constructed data features gave better results in the identification of potential package buyers of the three services. Upon analysis of correlation between constructed features and target feature for voice package buyer's dataset, peak call duration and outgoing call duration features exhibit linear correlation with positive target feature instances (buyers) and inverse linear correlation with negative target feature instances (non-buyers). This is due to user's interest to minimize the payment for voice service usage.

In this study, 91% of prospective voice service package buyers have been identified with Neural network classifier. However, the accuracy of SVM, and K-nearest neighbor is equal (90%), and the accuracy of Decision tree (J48) classifier is 92%. Neural network is selected as the best classifier for classification of voice service package potential buyers. Since the major aim is to have effective direct marketing by communicating the potential buyers of the service packages and avoiding unnecessary promotional contact to non-buyers, in-depth focus on measures beyond accuracy of classifiers is important. Consequently, by ranking all the classifiers based on the number of correctly and incorrectly classified instances in the confusion matrix, Neural network maximizes the correctly classified instances and minimizes the incorrectly classified. Hence, using Neural network for classification of voice package buyers gives better opportunity in the process of identifying potential buyers and avoiding contacting non-buyers, even though Decision tree have higher accuracy.

With data-package buyer identification process, 94% of buyers have been identified with SVM and KNN classifiers which are comparatively less accurate of the best performing groups. With this dataset, Neural network classifier has exhibited the best performance (96%). Further checking the correctly and incorrectly classified instances in the confusion matrix, Decision tree (J48) classifier results with the highest number of correctly identified instances. But, the accuracy of Neural network classifier is greater than others, better targeting of buyers can be achieved with Decision tree with the dataset used in this experiment.

Neural network classifier is also effective in SMS package buyers classification with both accuracy measure and higher number of correctly classified instances. The accuracy of the four classifiers except Naive Bayes is also good in SMS package buyers' dataset. The least 93% buyers have been identified with KNN classifier and 95% with Neural network and Decision tree. From the results exhibited by Naive Byes classifier, its performance is comparatively less which is 88% in the case of SMS package buyers' dataset and even less (74%) in voice package buyers' dataset. Naïve Bayes classifier is more effective with dataset that have discrete and/or categorical features. But the dataset used in this study have continuous numerical features.

Generally, it possible for ethio telecom to identify more than 90% of prospective

buyers in value-added service packages market for the three services. Four classifiers (SVM, NN, KNN, and J48) have good performance classifiers in the classification of prospective buyers of value-added packages. As described earlier, consideration of precision, recall, and area under ROC is important to clearly understand weather the classifier is classifying instances to majority class or not. Using these three measures and number of instances in the confusion matrix, the best performance classifier has been selected. Specifically, Neural network classifier identified the highest number of prospective buyers of voice and SMS service packages and Decision tree (J48) for data service package buyers classification case.

Potential buyer identification and targeted marketing process discussed in this study can be realized by designing an automated system. This automated system takes data from the billing system with the implementation of aggregation strategies used in this research and classifiers identified more accurate by this study. The system's output (potential buyers list) can be communicated to SMS/MMS gateway for contacting the respondents with a high probability of buying.

6.2 Recommendation

Better insight can be obtained if further studies focus on the behavioural differences between users that buy packages and non-buyers or those that buy packages rarely. Such studies indicate market basket analysis opportunities like services that can be provided together with some packages or in the design of recommender systems.

It is also better to investigate package buyers use of applications and content access. This helps ethio telecom to provide special packages based on used applications and accessed contents in order to enhance the provision of specially designed packages based on user's preference.

Appendix A

Explanation of Raw CDR and Aggregated Fields

A.1 Explanation of Raw CDR Fields

CDR Column Name	Description
CDR_ID	CDR service type (Voice, Data, SMS) identifier id
RE_ID	Distinguishes between records of services like voice
BILLING_NBR	Identification of billing
CDR_TYPE	The service type the CDR is generated for
CALLING_NUMBER	Identity of call initiator
CALLED_NUMBER	Identity of call termination
CALLING_IMEI	Equipment identity of calling party
CALLING_IMSI	Subscriber (calling) identity
THIRD_PARTY_NUMBER	Third party equipment (like MSC, SMSC) numbers
CALL_START_TIME	Call start time stamp
CALL_END_TIME	Call end time stamp
CALL_DURATION	Length of duration on call

CDR Column Name	**Description**
CALL_FEE	Fee charged for service usage
CALLED_COUNTRY	Country code where call is terminated
CALLING_CARRIER	Carrier identification where call is started
CALLED_CARRIER	Carrier identification where call is terminated
CALLING_DISTRICT	The visited area of calling party
CALLED_DISTRICT	The visited area of called party
STATUS_DATE	CDR status update date
CALLING_SUB_ID	ID of a sub CDR generated after CDR splitting.
BILLING_CYCLE_ID	The time when the charging event occurs
CHARGE_1	Total call fee
CHARGE_2	Reserved call fee
RATE_ID1	Rate service usage id
ACCOUNT_ITEM_ID1	Default account id of the charged party
UPLOAD_TRAFFIC	Volume of data traffic used during uploading
DOWNLOAD_TRAFFIC	Volume of data traffic used during downloading
BILLING_OFFERING_ID	Primary offering id
ERROR_CDR_TYPE	Error type
CALLFORWARDINDICATOR	Call forwarding flag
HOTLINEINDICATOR	Special number to indicate the number is common, barred, free, special-tariff, or voice mail
CALLING_TRUNK_ID	Network id of calling party
CALLED_TRUNK_ID	Network id of calling party

A.2 Explanation of Aggregated Fields

Data Feature	**Feature name**	**Aggregation strategy**
Total service usage fee	CFEE	Sum of fee for service payment
Outgoing call duration	OGD	Sum of duration of outgoing call in a month
Outgoing call frequency	OGF	Count of outgoing calls made in a month
Incoming call duration	ICD	Sum of duration for incoming calls
Incoming call frequency	ICF	Count of incoming calls made in a month
Off-peak usage frequency	OFPF	Count of outgoing calls made during off-peak hour (10:00 PM - 6:59 AM)
Off-peak usage duration	OFPD	Sum of duration for outgoing calls during off-peak hour (10:00 PM - 6:59 AM)
Off-peak usage fee	OFPFEE	Service charge paid during off-peak time usage
Peak hour usage frequency	PF	Count of outgoing calls made during peak hour i.e. before 10:00 PM and after 6:59AM
Peak hour usage duration	PD	Sum of duration of outgoing calls made during peak hour i.e. before 10:00 PM and after 6:59AM
Peak hour usage fee	PFEE	Service charge paid during peak hour
International call usage frequency	ICF	Count of frequency of international calls made
International call usage duration	ICD	Sum of duration of international calls made
International call usage fee	ICFEE	Service charge paid for international call usage

Data Feature	**Feature name**	**Aggregation strategy**
Upload traffic volume	UTR	Sum of upload traffic volume
Upload traffic fee	UTRFEE	Service charge paid for upload traffic
Download traffic volume	DTR	Sum of download traffic volume
Download traffic fee	DTRFEE	Service charge paid for download traffic
Morning time upload volume	BNU	Sum of upload traffic volume in the time interval 7:00AM - 11:00 AM
Morning time download volume	BND	Sum of download traffic volume in the time interval 7:00AM - 11:00 AM
Afternoon time upload volume	AFNU	Sum of upload traffic volume in the time interval 6:00PM - 18:00 PM
Afternoon time download volume	AFND	Sum of download traffic volume in the time interval 6:00PM - 18:00 PM
Evening time upload volume	BMNU	Sum of upload traffic volume in the time interval 18:00PM - 23:59 PM
Evening time download volume	BMNU	Sum of download traffic volume in the time interval 18:00PM - 23:59 PM
After midnight download volume	AMNU	Sum of download traffic volume in the time interval 00:00AM - 6:59 PM
After midnight upload volume	AMND	Sum of upload traffic volume in the time interval 00:00AM - 6:59 PM
Sent SMS count	SENDFRQ	Count of SMS generated by user
Received SMS count	RECFRQ	Count of SMS received by user
Bulk SMS response count	BSRESP	Count of bulk SMS message responses made by user
Bulk SMS received count	BSRECV	Count of bulk SMS messages received by user

Bibliography

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

[2] S. Jaroszewicz, "Cross-selling models for telecommunication services," *Journal of Telecommunications and Information Technology*, pp. 52–59, 2008.

[3] G. Maji and S. Sen, "Data warehouse based analysis on cdr to retain and acquire customers by targeted marketing," in *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 2016, pp. 221–227.

[4] L. Liu, Z. Yang, and Y. Benslimane, "Conducting efficient and cost-effective targeted marketing using data mining techniques," in *2013 Fourth Global Congress on Intelligent Systems*. IEEE, 2013, pp. 102–106.

[5] Z.-Y. Chen, Z.-P. Fan, and M. Sun, "A multi-kernel support tensor machine for classification with multitype multiway data and an application to cross-selling recommendations," *European Journal of Operational Research*, vol. 255, no. 1, pp. 110–120, 2016.

[6] H. Ahn, J. J. Ahn, K. J. Oh, and D. H. Kim, "Facilitating cross-selling in a mobile telecom market to develop customer classification model based on hybrid data mining techniques," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5005–5012, 2011.

[7] A. Bhadani, R. Shankar, and D. V. Rao, "A computational intelligence based approach to telecom customer classification for value added services," in *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)*. Springer, 2013, pp. 181–192.

[8] I. Purnamasari, R. B. Bahaweres *et al.*, "Naive bayes classifier algorithm and particle swarm optimization for classification of cross selling (case study: Pt

telkom jakarta)," in *2016 4th International Conference on Cyber and IT Service Management*. IEEE, 2016, pp. 1–4.

[9] F. Thuring, J. P. Nielsen, M. Guillén, and C. Bolancé, "Selecting prospects for cross-selling financial products using multivariate credibility," *Expert systems with Applications*, vol. 39, no. 10, pp. 8809–8816, 2012.

[10] A. Rahman and M. N. A. Khan, "A classification based model to assess customer behavior in banking sector," *Engineering, Technology & Applied Science Research*, vol. 8, no. 3, pp. 2949–2953, 2018.

[11] J. D. Kelleher, B. Mac Namee, and A. D'arcy, *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT Press, 2015.

[12] C. Alexandre, P. Bocsanean, J. Mangana, J. Santos, D. Monteiro, and P. Santos, "Marketing behaviors analysis in a mobile wallet solution using data mining," in *2017 9th International Conference on Computational Intelligence and Communication Networks (CICN)*. IEEE, 2017, pp. 88–92.

[13] J. Kelleher, B. Namee, and A. D'Arcy, "Machine learning for predictive data analytics," 2015.

[14] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, 2018.

[15] M. M. Al-Rifaie and H. A. Alhakbani, "Handling class imbalance in direct marketing dataset using a hybrid data and algorithmic level solutions," in *2016 SAI Computing Conference (SAI)*. IEEE, 2016, pp. 446–451.

[16] T. Liang, B. Zeng, J. Liu, L. Ye, and C. Zou, "An unsupervised user behavior prediction algorithm based on machine learning and neural network for smart home," *IEEE Access*, vol. 6, pp. 49 237–49 247, 2018.

[17] J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, and G. Cheng, "Characterizing user behavior in mobile internet," *IEEE transactions on emerging topics in computing*, vol. 3, no. 1, pp. 95–106, 2014.

[18] B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo, "Analysis of gsm calls data for understanding user mobility behavior," in *2013 IEEE International Conference on Big Data*. IEEE, 2013, pp. 550–555.

[19] Z. Guo and F. Wang, "Telecommunications user behaviors analysis based on fuzzy c-means clustering," in *International Conference on Future Generation Information Technology*. Springer, 2010, pp. 585–591.

[20] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

[21] A. Sell and P. Walden, "Segmentation bases in the mobile services market: Attitudes in, demographics out," in *2012 45th Hawaii International Conference on System Sciences*. IEEE, 2012, pp. 1373–1382.

[22] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[23] A. S. Halibas, A. C. Matthew, I. G. Pillai, J. H. Reazol, E. G. Delvo, and L. B. Reazol, "Determining the intervening effects of exploratory data analysis and feature engineering in telecoms customer churn modelling," in *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)*. IEEE, 2019, pp. 1–7.

[24] C. C. Aggarwal, *Data mining: the textbook*. Springer, 2015.

[25] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, pp. 3–24, 2007.

[26] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2017, pp. 79–85.

[27] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in *2008 Fourth international conference on natural computation*, vol. 4. IEEE, 2008, pp. 192–201.

[28] G. Bonaccorso, *Machine learning algorithms*. Packt Publishing Ltd, 2017.

[29] D. F. Nettleton, A. Orriols-Puig, and A. Fornells, "A study of the effect of different types of noise on the precision of supervised learning techniques," *Artificial intelligence review*, vol. 33, no. 4, pp. 275–306, 2010.

[30] Ethiotelecom. (2019) Home. Ethiotelecom. [Online]. Available: https://www.ethiotelecom.et