BANK CUSTOMER CHURN PREDICTION MODEL: THE CASE OF COMMERCIAL BANK OF ETHIOPIA

A Thesis Presented by

Berhane Gebreegziabher Seyoum

To

The Faculty of Informatics of

St. Mary's University

In Partial Fulfillment of the Requirements for the Degree of Master of Science

In Computer Science

January, 2022

ACCEPTANCE

Bank Customer Churn Prediction Model: The case of Commercial Bank of Ethiopia

By

Berhane Gebreegziabher Seyoum

Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Thesis Examination Committee:

_____

Internal Examiner

Alemebante Mulu (PhD)

_____

External Examiner

Minal Ashagrie (PhD)

Dean, Faculty of Informatics

Alemebante Mulu (PhD)

January, 2022

DECLARATION

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

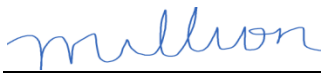
Berhane Gebreegziabher Seyoum
Full Name of Student

Signature

Addis Ababa Ethiopia

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

Million Meshesha (PhD)
Full Name of Advisor



Signature

Addis Ababa Ethiopia

January, 2022

Acknowledgments

First and foremost, I would like to give thanks to God the Almighty who provided me everything to accomplish this thesis.

This thesis would not have been possible without the help, support, and patience of my principal thesis advisor, Dr. Million Meshesha for his guidance and encouragement right from the stage of problem formulation to the completion of the work and I am extremely grateful to him for his unreserved contribution, advice, and encouragement.

I am also grateful to Commercial Bank of Ethiopia especially Ato Girma Lewoyehu Assistant Director of Performance Management System, Career Development and Succession Planning of Commercial Bank of Ethiopia for his unlimited support and understanding.

I am also grateful to all Ethiopian Instructors and Teachers those who have spent their entire lives learning and teaching and to St' Mary University Instructors and employee of the University who has support throughout this Master's program from start to end.

I must heartfelt thanks to my close friends Mesfin Berhane and Gashaw Sisay for their timely review and constrictive feedbacks.

Finally I am deeply grateful for my family, especially my wife Ms. Firehiwot Alemneh who has supported me throughout this Master's program from the start to the end and my children Yemariam Berhane and new born baby Michaela Berhane for encouragement. I would like to dedicate this Master Thesis to them as an indication of their significance in this Thesis.

Table of Contents

List of Abbreviations

CBE	Commercial Bank of Ethiopia
CRISP-DM	Cross-Industry-Standard-Process for Data Mining
CRM	Customer Relationship Management
DT	Decision Tree
K-NN	K-Nearest Neighbor
ML	Machine Learning
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
WEKA	Waikato Environment for Knowledge Analysis

List of Figures

List of Tables

Abstract

In 21st century because of availability and affordability of computer technology, organizations and businesses especially in banking sector are situated in basic requirement to gain a number of key advantages to improve their business using Machine Learning (ML) Algorithm. ML Algorithms is a branch of artificial intelligence based on the idea that systems can learn from data, identify model and make to support decision with minimal human intervention brief about the customer churn. Nowadays industries working with large amounts of data have recognized the value of machine learning in this case Commercial bank of Ethiopia (CBE). CBE is one of such service-giving industries that collects, processes and stores huge amounts of records from time to time and therefore deal with large amount of data. On the other hand, CBE is facing problems in Customer Relationship Management (CRM), specifically it is unable to control the customer churn. Customer Churn is the propensity of a customer to stop doing business with an organization and subsequently moving to some other company. In this study an attempt is made to apply machine learning algorithms for customer churn prediction. After performing business and data understanding the data preparation task is done to clean and make the data ready for experimentation. For the experiment and construct predictive model, machine learning algorithms such as SVM, KNN, Naïve Bayes and Logistic Regression are selected based on their advantages and past performance seen in different literatures, it has been reported that they were widely used classifier algorithms for prediction and classification. The R Studio with R programming was used to simulate all the experiments. Confusion matrix was used to calculate the accuracy, recall and precision and evaluate the performance of the models. The results of the experiment show high accuracy, so that the models can be used to predict customer status accurately. Based on the research findings, the KNN classifier produced an accuracy of 99.91%, the SVM classifier produced an accuracy of 92.4%, Logistic Regression model also produced an accuracy of 93.8%, and Naïve Bayes classifier produced an accuracy of 83.8 %. Therefore, the KNN classifier is proposed for constructing bank customer churn prediction model for Commercial Bank of Ethiopia. Based on the proposed optimal model in this study, we recommend future research to integrate customer churn predictive model with CRM data base management system.

Keywords*: Customer Relationship Management and Customer Churn*

CHAPTER ONE

INTRODUCTION

1.1 Background

Scholars noted that, “Necessity is the mother of invention” [1]. As we all know we are drowning in data, but starving for knowledge, which motivates researchers the need for massive data analysis with the help of Machine Learning (ML) Algorithm Techniques in order to achieve organizational goals.

ML methods have increased since the extent of computing power, and the amount of data gathered has increased tremendously [2].The term machine learning can be defined as “computational methods using the experience to improve performance or to make accurate predictions”.ML methods can be divided into supervised learning, unsupervised learning, where the main difference is that with supervised learning, the data is labeled, and in unsupervised learning unlabeled [2].

Commercial Bank of Ethiopia (CBE) is governmental bank of Ethiopia established with the objective of storing a utility like money, jewelry, and others to the people of Ethiopia. The primary target of the Government is to prevent theft and destruction, and to train the culture of the bank Industries system.

The history of Commercial Bank of Ethiopia (CBE) dates back to the establishment of the State Bank of Ethiopia in 1942 G.C. [3]. It was legally established as a share company in 1963 G.C. In 1974 G.C., CBE merged with the privately owned Addis Ababa Bank. Since then, it has been playing significant roles in the development of the country and pioneer to introduce modern banking to the country. The bank operates under the top level supervision of a board of directors which is composed of a board chairman and nine board members.

A potential objective of every financial organization is to retain existing customers and attain new prospective customers for long term. Nowadays, Banks have realized that customer relationships are a very important factor for their success. Customer relationship management (CRM) is a strategy that can help them to build long-lasting relationships with their customers and increase their revenues and profits.

Customer churn prediction is a field that uses machine learning to predict whether a

customer is going to leave the company or not. Churn prediction models determine customers who stop utilizing a service or product. This is a significant interest for product providers because a large number of churning customers not simply drive to decrease of revenue but can also harm the reputation of a company [4]

In today‘s competitive banking industry, customers can make a choice among various service providers by making a trade-off between relationships and economies, trust and products, or service and efficiency [5].When the customers get lower quality service from the bank the customers will be churned because of these customer churn has become one of the top issues for most banks. Churn is a term used in bank and many other industries referring to customers’ decision to move their subscription from one service provider to another [6].

Losing the customers can be very expensive as it costs to acquire a new customer. It not only leads to opportunity lost because of reduced sales, but also to an increased need for attracting new customers, which is five to six times more expensive than customer retention [6] also noted that churn of good customers have irrecoverable disadvantages for a company.

Hence, there is a pressing need from the bank to understand their customers’ needs, preferences, sentiments, behavior and propensity to switch becoming paramount for banks.

All bank marketing campaigns are dependent on customers' huge Electronic data. The size of these data sources is impossible for a human analyst to come up with interesting information that will help in the decision-making process so, to reduce this complexity Machine Learning models will be useful for a better customer handling. ML allows constructing a prediction model from the historical data, and thereby predict outcomes of future situations in advance. It helps optimize business decisions, increase the value of each customer communication, and improve customer satisfaction.

Currently CBE provides Deposits, Credits, Trade Service, Internet and Card banking products for their customers [3]. Those customer data are the main source of information that can be processed and interpreted in different ways so that, the commercial bank of Ethiopia can predict or classify and know about their customers based on predicted behavior. Predictive Machine learning techniques are useful to predict which customer is likely to churn out based on the customers’ Demographic

data, Behavioral data and Transactional data [4]. The first research question of the study tried to answer a problem area using Demographic (like: income, age, sex, region, education status, marital status, Job type, nationality, occupation) and Behavioral data (like: product, ownership, industry, target); however Transactional data has been done in previous study.

The intention of this research is therefore to apply different ML algorithms for customer churn problem in commercial bank of Ethiopia in order to convert the huge amount of customer dataset into useful model. This will be helpful in predicting customers 'behavior' and supports the CRM processes of the bank.

1.2 Statement of the Problem

Because of basic problem on customer handling, means the existence of shortages in Customer Handling Proficiency, previous experiences in the banking industry, and attention given to the field; handling the needs of customers especially for banks which have more customers like CBE has never been easy. In recent years it has become even more difficult because of the customer have an opportunities and options to choose and join another competitor.

A big problem that encounters in finical businesses, especially in banking sector is customer churn [8] [9]. This occurs when a customer decides to leave the bank because of unsatisfactory product services of the bank.

When banks provide services which is not based on the exact needs of the customer, the bank faces high iteration of customer turnover which is known as customer churn. Existing customer's churn may result in the loss of businesses and thus decline in profit [7]. Customer churn may be caused by several common reasons such as dissatisfaction with the services and high bills. Loss of customers equals the loss of future bank revenue plus loss of initial investment made to acquire those customers. In addition to this losing customers dose not only lead to opportunity lost because of reduced sales, but also to an increase need for attracting new customers, which is five to six times more expensive than customer retention[6].

Therefor to fulfill the aim of the study we mention two local-work researchers' problem and gaps in relation to customer churn prediction, one in banking and the other in

microfinance industries whose highlight of limitations are discussed here under and detailed in the next Literature Review chapter.

Application of Data mining Techniques to Predict Customers 'churn at Commercial Bank of Ethiopia was conducted by Kassahun G. who used 13,172 customer dataset with 9 attributes [8]; and Application Data Mining Techniques for Customers Segmentation and Prediction: in The Case of Buusaa Gonofa Microfinance Institution was conducted by Belachew who used 13,057 customer dataset with 6 attributes[9]. Both researchers used small dataset and less attributes selected. This shows the need for improvement where this research tried to resolve by adding four times more sample size than all the other researchers mentioned here in the study and 25 more attributes to improve the prediction model, accuracy and precision that is from all 34 attributes selected 15 are used by all the four models.

The aim of this study is therefore to apply machine learning algorithms for constructing Customer churn Prediction model for commercial Bank of Ethiopia. To this end, this study attempts to answer the following research questions:

1. Which attributes are more significant to predict customers churning at CBE?
2. Which classification algorithm is more suitable for constructing CBE customer churn prediction model?
3. To what extent the proposed model performs in customer churn prediction?

1.3. Objectives of the Research

13.1 General Objective

The general objective of this research is to construct a prediction model for determining CBE customer churn using Machine Learning Algorithm.

1.3.2 Specific Objectives

For the realization of the general objective stated above, the following specific objectives are formulated.

- To prepare appropriate datasets with relevant attributes for classification and

prediction model.

- To identify and find the main attributes that can help to predict customer churns.
- To collect and prepare data for training and testing during experimentation.
- To select suitable and most widely used machine learning algorithms for customer relationship management.
- To construct customers churn prediction model using the selected machine learning algorithms.
- To evaluate the proposed customer churn prediction model using performance measures on the bases of its accuracy and precision.

1.4 Scope and Limitation of the Research

The proposed research is aimed to investigate and construct a model for determining the probability of CBE customer's churn through specific products and some customers' behaviors and then predict the existing and new customers churn probabilities based on collected CBE dataset. This work encompasses and passes through steps such as Business Understanding, Data understanding, Data Preparation, modeling using Classification algorithms, and evaluation of the proposed model with test datasets. The proposed work only apply machine algorithms to predict CBE product's customers churn; so it is out of the scope of the current research to predict customer churns for other bank products, such as private Banks, private Insurance, Telecommunications, Microfinances, cooperatives associations, pension and social security institutions etc.

1.5 Significance of the Research

Meanwhile the main advantage of this research is to makes several contributions to both model building and practice improvement in customer churns prediction. From this research CBE is the main beneficiary. For the research identified and detected the model of customer status churns or Not Churns, and minimized and limited the CBE customer churns.

The bank can be safe from the loss of its businesses and decline in profit against

customer churns. In addition to this, the research contributes to knowledge building including for other researchers interested in similar area that includes.

The research shows which ML algorithms are more appropriates for predictions of customer churns in banking industry domain; plus it gives good understanding in the concepts of ML model building using R Programming Tools.

1.6 Research Methodology

Research methodology is the general principle that guides the research. In order to conduct a good research, a well-defined approach and principle has to be followed.

1.6.1 Research Design

This research follows an experimental type of research. It is a collection of research design which uses manipulation and controlled testing to understand causal (cause/effect) relationships and to research the relationship between one variable and another.

For proper understanding of the problem under investigation and successful completion of this research, relevant literatures such as books, journals, magazines, conference papers, manuals, and resources from internet, particularly CBE manuals are reviewed for achieving the research objective.

1.6.2 Data Preparation

This phase is also known as data filtering and involves the process of organizing the data for ML. The following steps are completed in order to filter the data to be used in the ML process. These are Business Understanding (Customer Definition, Classification, Account Handling, and Attribute Selection), Data Understanding (Data Representation), Data Collection (Data Quality Assurance), and Data Preprocessing (Data Cleaning and Data Set Splitting).

1.6.3 Implementation tools

This research used experimentation of proposed method and application development environments that are required to produce the output of the dataset using R studio with R programing Tools. R programming tools is the language of Data science that integrated suite of software facilities for data manipulation, calculation and graphical facilities useful for experiment and analysis dataset. It is preferred because of the following reasons: R-Studio has easy and friendly use; Interface-Studio has ability to present datasets in the form of figures with variety of presentations. In R-Studio it is possible to link datasets in common simple format such as .CSV; it is simple and suitable for technical computing; it's also Effective data handling and storage; Suite of operators for calculations on arrays and large, coherent, integrated collection of intermediate tools for data analysis [10] [11]. In general R is selected for rules mining and MS Excel is employed for preprocessing the dataset.

1.6.4 Evaluation

The proposed work was evaluated by comparing the output against the manually observed phenomena using a confusion matrix and also by conducting a comparison with other algorithms that are frequently used by previous works. The confusion matrix is a 2 by 2 table with 4 outcomes, which is true positive (TP), true negative (TN), false positive (FP), and false-negative (FN). These outcome measures are used to describe the performance of a classification model on a set of data.

1.7 Organization of this Thesis

This research thesis will be organized in to five chapters.

The First Chapter presents brief Introduction about the research, Statement of the Problem, General and Specific Objectives, the Scope and Limitations of the Research, Significance, and Research Methodology of the Research.

The Second Chapter starts with an overview of Literature Review, Overview of CBE, and Customer Churn prediction in banking industry, and following by brief overview of Machine Learning Algorithms Techniques how those techniques work, Model

Evaluation, Review of Related works and Gap Analysis.

Chapter Three will present overview of Data Preparation followed by Business Understanding, Data Understanding, and Data Preprocessing.

Chapter Four will present the Overview of Experiment and Discussion of the Results followed by, Proposed Architecture, and Dataset for Experiment, Modeling using KNN, Modeling using Logistic Regression, Modeling using Naïve Bayes, Modeling using SVM, Comparison of Machine Learning Models, and finally Discussion of Result detail.

The Last Chapter will present the Conclusion and Recommendation of the research including future suggestion.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview

Bank is a type of organization dedicated to receive deposits and provide loans to its customers. Banks can also engage in service provision relating to finance like management of wealth, exchange of currencies, and safe deposit boxes. To date, Bank types are classified into two groups namely Commercial and Investment. Usually banks are governed by regulations and commands emanating from governments or Central Banks. Major responsibilities of banks include stabilizing currency, inflation and monetary policy control, and supervision of the overall supply of money [3].

A Commercial Bank is type of bank that provides services such as accepting deposits, making business loans, and offering basic investment products that is operated as a business for profit.

Commercial banks are typically concerned with managing withdrawals and receiving deposits as well as supplying short-term loans to individuals and small businesses. These banks are primarily used by customer for account checking, saving, deposing, and securing loans using home collateral.

Overview of Banks in Ethiopia

One year after it was officially established in 1963 G.C, the National Bank of Ethiopia (NBE) went operational by 1964. Known to have the role of a central bank in Ethiopia, NBE- issues Operation Licenses to newbie banks and other financial institutions; supervise them; regulates money and credit availability, supply, as well as their cost; manage its own foreign reserves and exchange rates.[3]

Commercial Bank of Ethiopia (CBE):

As we introduce the history of CBE in chapter one, CBE provide Deposit Service (Saving Account, Current - Checking Account and Diaspora Account), Credit Service, International Service (Trade Service, Forex Service, Money Transfer, and Correspondent Banking) and Payment Service (Internet Banking, Card Banking,

Mobile Banking and Card Local Transfer). CBE is the owner of 1700 and more branches throughout the nation. According to the home page which is surfed Dec, 2021, CBE is the leading African bank with assets of more than 911.96 billion Birr as on Dec, 2021. This has extreme importance in economic progress & development of the country. As of December 2021 number of Account Holders in CBE reached more than 33.3 million while Mobile and Internet Banking subscribers exceeded 3 million [3].

Even if the bank achieved to this level there are different group of customers which are not satisfied as such, which leads them to leave the bank. Customer Churn can be defined as the loss of clients or customers of a given institution. It can also be termed as- Customer Attrition, Customer Turnover, or Customer Defection, It is the loss of clients or switch iteration of customers called churns. Customer Churn, therefore, is percentage of a company's customers known to have banned to use the service that has been provided to them at a given period of time.

2.2 Background of Customer Churn

According to Bhambri, Churn is defined as propensity of a customer to stop doing business with an organization and subsequently moving to some other company [12]. In the banking world Customer Churn is defined as a customer that closed all his/her bank account and halted further business transaction with that specific bank. Customer churn is not only limited to inducing income loss, however could potentially impose adversity upon current and future engagements of the companies [13].

Today's market competition forced many companies to sell similar products using the same service and quality of product. On other hand with the rising of competition there is an increase in customer expectations and diminishing customer loyalty. Customers now prefer fast feedback in a short time possible; efficient and effective ways of service provision; value added products encompassing their specific demands; comparably low transaction cost; simple or easy procedures; and attractive as well as personalized services.

So nowadays Customer churn is a top-priority case for banks giving due focus in dealing with CRM.

In addition to the fierce competition, customer acquisition cost is greater than retaining the

existing ones. It is revealed in various studies that the cost of acquiring new customers is 5 to 10 times higher than keeping the existing ones satisfied and loyal. In another perspective 10 to 30 percent of customers churn out annually due to poor customer care [14]. This led companies to put a lot of effort in understanding and analyzing their customers' behavior, in order to identify with adequate advance which clients will leave. In particular, these customers are subjected to several rectifying actions (e.g. promotions or gifts) for their retention. This makes customer retention an interesting topic for all businesses. Nowadays, for this same reason companies involved themselves in satisfying and thereby retaining their customers. Especially in industries such as Banks, Insurances, Telecommunications, and etc. which depend on subscribing customer's revenues are results of fees made their customers at times of services. Because these Companies engaged in providing services to huge number of customers such are provided by the payments made by these customers periodically [15].

So to make revenues more sustainable at their lowest corresponding cost, the main issue has to be keeping customers satisfied regularly This is a kind of business strategy known to the business and economics field as "Customer Relationship Management" (CRM) which focuses at ensuring customer's satisfaction. The companies which successfully apply CRM to their business nearly always improve their retention power, that is, the probability that a customer will not leave. Because the customer churn rate has a significant effect on the financial market value of the company. So, most companies keep an eye on the value of the customer at monthly or quarterly periods [16].

2.3 Overview of Machine Learning

Nowadays because of the availability and affordability of computer technology organizations and businesses became able to record their day to day activities. With the advancement of this technology and wide application of database, corporations and organizations gather a huge amount of data record which is stored in different form but identifying valuable information or knowledge for taking a concrete decision becomes very difficult. According to Kumar D. & Bhardwaj D. [17], the fact shows that data is growing at a very rapid rate, but most of data is simply being accumulated in storage device for no further use.

If the collected data from different sources processed properly, can provide hidden knowledge, which can be used for further development and that help to business decisions which bring business to the next level. But as Zen Tut [18] mentioned, the un necessary accumulated data itself is critical to a company‘s growth.

Because of such reasons a new concept of Machine Learning need is important for business organizations for making better decision. Machine Learning refers to extracting knowledge from large set of data whatever may be the nature of data [18]. The data may be web data, multimedia, text data etc. It is also the set of pattern used to find new or unexpected pattern in data using information contained in data warehouse.

The discovered knowledge can be used by the bank managers to acquiring new customers, increasing revenue from existing customers, and retain good customers which is one of the important applications of Machine Learning techniques that can be used in banking organization.

Machine learning is an interdisciplinary field that requires knowledge from artificial intelligence and mathematical statistics to find and extract model from datasets that is beyond the capabilities of the SQL language (Structured Query Language) [20]. Moreover, Machine Learning requires following these steps:

Preprocessing->model->validation

- **Pre-processing**: Preprocessing is one of the important stages in Machine learning because real datasets are noisy, dirty, incomplete, and in different formats.
- **The model:** The model means that the techniques and the algorithms that Machine Learning uses are applied to the data to get results. There are wide range of Algorithms used in Machine Learning described in detail by Chen K [13]; some of the more common ones being K-means clustering and A-priori algorithm.
- **Validation:** This is the final stage of Machine Learnings, which verifies the output Model from the Machine Learning algorithms. All the output model that are found by Machine Learning algorithms are not necessarily valid. To overcome this challenge, the Machine Learning algorithms need to be tested on a test set of data. If the output Model meet the desired output of the Machine Learning algorithm, it will be applied to a larger dataset to discover knowledge. However, if the output model do not fit the desired results, the pre-processing and the Machine Learning algorithm steps need to be re-evaluated [20].

2.4. Machine Learning techniques

There are many Machine Learning algorithms techniques that can be used to classify a problem given a set of features in both supervised and unsupervised learning. These work looks to supervised algorithms to see if any are particularly useful in analyzing the behaviors of CBE customer and give prediction of customer churn. Machine Learning Algorithms investigated in this research are: K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), Logistic regression and Naive Bayes.

2.4.1. Support Vector Machine (SVM)

Support Vector Machines are supervised learning models that can be used for classification, prediction and clustering problems. An SVM takes a set of input observations and associated binary outputs and constructs a model that can classify new observations into one class or the other. Support vector machine is highly preferred by many as it produces significant accuracy with less computation power [21].

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new examples. In two-dimensional space this hyper plane is a line dividing a plane in two parts where in each class lay in either side [22].

The key element is that support vector machine algorithm is to find a hyper plane in an N-dimensional space (N—the number of features) that distinctly classifies the data points.

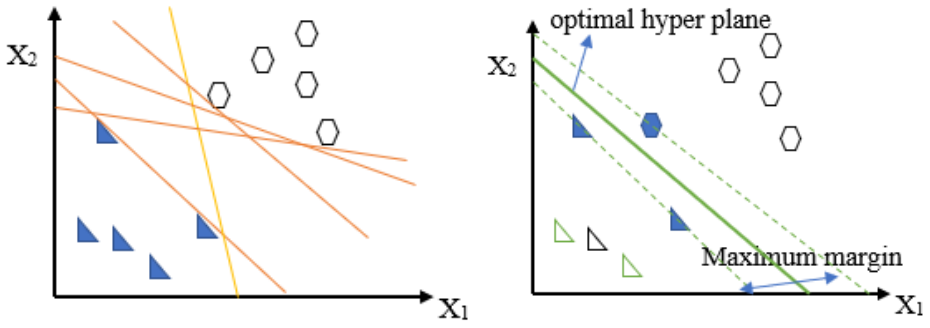


Figure2.1. A linear line separating the data types [22].

To separate the two classes of data points, there are many possible hyper planes that could be chosen. The main thing in SVM is to find a plane that has the maximum margin, which means the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

In real world application, there is trade off on finding perfect class when the two classes are not linearly separable like due to noise, the condition for the optimal hyper-plane can be relaxed by including an extra term: millions of training dataset.

In the case of real-world application, it is not usually possible to get a line that perfectly separates the data within the space. Hence, we might have to use a curved decision boundary. It is possible to get a hyper-plane which could separate the data but this may not be desirable if the data has noise in it. In such cases we need to use the soft margin method [22]. The soft margin method allows for points to appear on the incorrect side of the margin. These points have a penalty associated with them. The penalty increases as the points are farther from the margin. The hyper plane separation looks to minimize the penalty of incorrectly labeled points, while maximizing the distance between the remaining examples and the margin.

The other approach is SVM kernel it employed to separate data that isn't linearly separable, is to map the data into a higher dimensional space. By mapping x= (x, x2) the data will be mapped into two-dimensional space. When this two-dimensional mapping is graphed, an obvious linearly separable line appears [22]. The mapping used to increase the dimensionality of the problem is dependent on the data space being investigated. The above computations, which are used to find the maximum-margin separator, can be expressed in terms of scalar products between pairs of data points in the high-dimensional feature space. These scalar products are the only part of the computation that depends on the dimensionality of the high-dimensional space.

2.4.2. Naive Bayes

Naive Bayes is a widely used classification method based on Bayes theory. Based on class conditional density estimation and class prior probability, the posterior class probability of a test data point can be derived and the test data will be assigned to the class with the maximum posterior class probability [23]. Naive Bayes algorithm is one

of the most effective methods and is a simple but surprisingly powerful algorithm for predictive modeling [24].

The main reason behind its popularity is that it can be written into the code very easily delivering predictions model in very less time. Thus, it can be used in the real-time model predictions. In Naive Bayes probability theory, Bayes theorem relates the conditional and marginal probabilities of two random events [23]. It is often used to compute posterior probabilities given observations.

Let x = (x 1, x2. . . xd) be a d-dimensional instance which has no class label, and our goal could be to build a classifier to predict its unknown class label based on Bayes theorem. Let C = {C1, C2… CK} be the set of the class labels. P (Ck) is the prior probability of Ck (k = 1, 2... K) That are inferred before new evidence, P (x|Ck) be the conditional probability of seeing the evidence x if the hypothesis Ck is true. A technique for constructing such classifiers to employ Bayes' theorem to obtain is given by the following formula: -

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{\sum_{k'} P(x|C_{k'})P(C_{k'})} \qquad [2.1]$$

A naive Bayes classifier assumes that the value of a particular feature of a class is unrelated to the value of any other feature, so that [23]

$$P(x|C_k) = \prod_{j=1}^{d} P(x^j|C_k) \qquad [2.2]$$

2.4.3. K-Nearest Neighbor (KNN)

The nearest neighbor (NN) classifiers, especially the k-NN algorithm, are among the simplest and yet most efficient classification rules and are widely used in practice [22].

The purpose of KNN algorithm is to use a database in which the data points are separated into several separable classes to predict the classification of a new sample point. An object is classified by looking to its nearest examples [25]. In KNN algorithms, K states how many neighbors will be used in voting, it is important to decide the value of k because the accuracy of the classification is dependent on it. K=1 simply

states that an object will be assigned the same class as its nearest example. As the number of K increases then we need to classify a given instant based on the resemblance of all the stated K instances.

The measurement can be performed using any distance metric or similarity function, such as the Euclidean, Cosine or Jaccard [25]. The classifier is developed from the training data documents in this case and to check for the efficiency and accuracy of the classifier the holdout method has been used. In the holdout method the original training dataset is partitioned into two, a major portion to generate the classifier and the minor portion to check the precision of the classifier. It is an assumption that documents of the same class will always be the nearest neighbors of each other i.e. the distance between them will be less as they are in some way related to each other [25].

The major problem of using the k-NN decision rule is the computational complexity caused by the large number of distance computations and the other most critical problem is selecting the value of K in K-nearest neighbor. A small value of K means that noise will have a higher influence on the result i.e., the probability of over fitting is very high. A large value of K makes it computationally expensive and defeats the basic idea behind KNN (that points that are near might have similar classes). A simple approach to select k is k = n^ (1/2) [25].

In the general process of KNN algorithms the first step is preprocessing the data in the database in such a way as to ensure that we can compare observations. Then, our observations become points in space and we can interpret the distance between them as their similarity (using some appropriate metric) One of the most widely used metrics is the Euclidean distance. The Euclidian distance between two instances (X1, X2, X3 … Xn) and (U1, U2, U3 … Un) is given by the following formula: -

$$\sqrt{(X_1 - U_1) + (X_2 - U_2) + \cdots + (X_n - U_n)} \qquad [2.3]$$

2.4.4. Logistic regression

Logistic Regressionis also called logistic model or logistic regression. It is a predictive analysis. It takes independent features and returns output as categorical output. The

probability of occurrence of a categorical output can also be found by Logistic Regression model by fitting the features in the logistic curve [26].

Logistic Regression, falls under Supervised Machine Learning. It solves mainly the problems of Classification to make predictions or take decisions based on past data [26]. It is used to predict binary outcomes for a given set of independent variables. The dependent variable's outcome is discrete.

The output of Logistic Regressionis a sigmoid curve or more popularly known as S-curve. Where the value on the x-axis, independent variable would determine the dependent variable on the y-axis. In Logistic Regression there are only two possible outcomes. 0 and 1. That something occurs, or it doesn't. We use a threshold value to make our prediction easier. If the x-axis' corresponding y-value probability is lesser than the threshold value, the outcome is taken as 0. If it is greater than the value, the outcome is taken as 1.

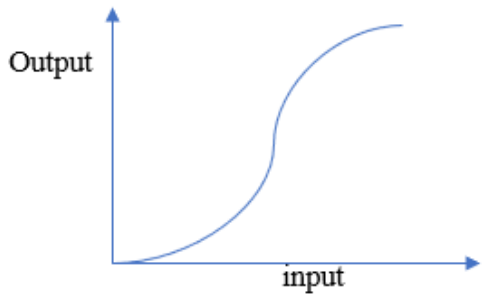


Figure 2.2 General Logistic Curve [26].

The Logistic Regression model can be replaced by the simpler Linear Regression model when the output variable is taken to be continuous. When the output variable is not continuous or is dichotomous another model has to be applied in order to take this difference into consideration.

Logistic Regression model was chosen over the other models because of its mathematical clarity and flexibility. This model can have single or multiple predictors [26].

2.5 Model Evaluation

This section presents the evaluation method that was employed to evaluate the performance of the customer churn prediction models. One of the popular performance

evaluation methods in Machine Learning, Data mining, artificial intelligence and statistics is confusion matrix. In order to quantify the performance of a problem that contains two classes, confusion matrix is usually used [32]. The research used Confusion matrix as evaluation methods because it comprises of evidence about actual and predicted classification performed by the proposed prediction and classification model.

2.5.1 Confusion Matrix

The confusion matrix is used to measure the performance of two class problem for the given dataset. The right diagonal elements TP (true positive) and TN (true negative) correctly classify Instances as well as FP (false positive) and FN (false negative) incorrectly classify Instances.

TN = the number of incorrect classifications that an instance is Negative.

FP = the number of incorrect classifications that an instance is positive.

FN = the number of correct classifications that an instance is Negative.

TP = the number of correct classifications that an instance is Positive.

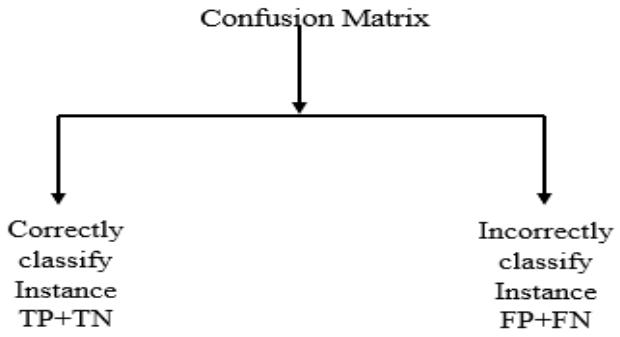


Figure 2.3 Example of Confusion Matrix [32].

Total number of instances = correctly classified instance + incorrectly classified instance

Correctly classified instance = TP + TN.

Incorrectly classified instance = FP + FN

Calculate Value TPR, TNR, FPR, and FNR One can calculate the value of true positive

rate, true negative rate, false positive rate and false negative rate by methods shown below [32].

Table 2.2 Example of Confusion Matrix.

	Predicted	
Actual	Yes	No
Yes	TP	FN
No	FP	TN

Accuracy (AC)

The accuracy is the total number of all correct predictions, TP and TN divided by the total number of the dataset. The best accuracy rate is 1.0, and 0.0 is the worst rate [31].

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad \text{.....(2.1)}$$

Precision

Precision positive prediction is the number of correct positive predictions divided by the total number of positive predictions, TP and FP. The best precision is 1.0, and the worst is 0 [31].

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{.....(2.2)}$$

False Positive Rate (FPR)

The false-positive rate is the number of incorrect negative predictions divided by the total number of negatives, TN and FP. The ideal false positive rate is 0.0, and the worst is 1.0.

$$FPR = \frac{FP}{TN+FP} \quad \text{.....(2.3)}$$

False Negative Rate (FNR)

The false-negative rate is the number of incorrect positive predictions divided by the total number of negatives, FN and TP. The best false negative rate is 1.0, and the worst

is 0.0.

$$FNR = \frac{FN}{FN + TP} \quad \text{.....(2.4)}$$

Error Rate (ERR)

The error rate is the number of all incorrect predictions, FP and FN divided by the total number of the dataset. The best error rate is 0.0, and 1.0 is the worst rate.

$$ERR = \frac{FP + FN}{TN + FP + FN + TP} \quad \text{..... (2.5)}$$

Sensitivity

Sensitivity also called True Positive Rate (TPR) is the number of TP divided by the total number of positives in the dataset, which is a TP and FN.

$$sensitivity = \frac{TP}{TP + FN} \quad \text{..... (2.6)}$$

Specificity

Specificity also called True Negative Rate (TNR) is the number of correct negative predictions divided by the total number of negatives, TN and FP.

$$specificity = \frac{TN}{TN + FP} \quad \text{.....(2.7)}$$

2.6 Review of Related works

As we all know in the global context various works have been done in the area of customer churn prediction for various industries without including the Demographic factor mean profile data.so the study divided the review in to two local work researcher and foreign work researcher and it lists the technique, model, tools, and analyzed problem and gap are summarized in table 2.1.

2.6.1 Foreign Related Works

Customer Churn Analysis in Banking Sector Using Data Mining Techniques [27]: They used Data Mining techniques to track or detect warning signs in customer's behavior such as reduced transactions, account status dormancy and take steps to prevent churn. The paper used a Data Mining model that can be used to predict which customers are most likely to churn (or switch banks). The paper used WEKA, Data Mining software tools for knowledge analysis at the real-life customer records in Nigerian bank to cleaned, pre-processed and then analyzed. The Article used 4958 dataset sample, 8 Attribute, and the performance not mentioned; and the results showed the method determine model in customer behaviors and help banks to identify likely churners and hence develop customer retention modalities.

Data Mining used to predict credit card customer churn [28]: They used 14814 data sample, 22 Attribute and accuracy achieved Multilayer Perception (MLP) 93.25%, Logistic Regression93.25%, Decision Tree (J48 95.97%), Random Forest (RF) 95.59%, Radial Basis Function (RBF) 9.11%, and SVM 93.11% techniques. The dataset was taken from the Business Intelligence Cup organized by the University of Chile in 2004. Highly unbalanced dataset with 93% loyal and 7% churned customers because of this the researchers apply Synthetic Minority Oversampling Technique (SMOTE) for balancing it, Furthermore, tenfold cross-validation was employed. The results indicated that SMOTE achieved good overall accuracy. Classification and Regression Tree (CART) was used for the purpose of feature selection. So the researcher conclude that it is the most important predictor variables in solving the credit card churn prediction problem.

Developing a prediction model for customer churn from electronic banking services using Data mining [29]: Collect data from organizations' Database. In this research, the decision tree technique was applied to build a model incorporating this knowledge. The results represent the characteristics of churned customers.

Customers churn prediction for an insurance company [30]: To identify important churning variables and characteristics, experts within the company were interviewed, while the literature was screened and analyzed. In addition, four promising Data Mining techniques for prediction modeling were identified, i.e. logistic regression, decision tree, neural networks and support vector machine. After performance evaluation, Logistic Regression with a 50:50 (non-churn: churn) training set and neural networks with a

70:30 (non-churn: churn) distribution performed best. In the ideal case, 50% of the churners can be reached when only 20% of the population is contacted.

Predictive model Insurance Industry using R tool [31]: They are tried to find the causes of losing customers by measuring customer loyalty to regain the lost customers. The research paper is using application of Machine Learning technology and R package to predict the results of churn customers on the insurance transaction customer dataset from Insurance Company. The R tool has represented the large dataset churn in form of graphs which depicts the outcomes in various unique pattern visualizations. The paper is logistic regression, decision tree, Random forest, Support vector machine model building to predict churn customer in insurance sector considering churn factor in account to depict various model for churners.

2.6.2 Local Related works

Although much effort was exerted to look into customer churn prediction for Local Research Works, only two researches were discovered one in Banking and the other in Microfinance Industries as mentioned under here.

Application of Machine Learning Techniques to Predict Customers 'churn at Commercial Bank of Ethiopia [8]: He has made a solution for the churn problem in banking sector using DM technique. The researcher use data of 13172 customers with 9 attributes (AccNo, AccType, Comp/Ind (Customer and Individual Customer), CurrCode, AcctStatus, CurrentBalance, DateAcctOpened, DateAcctClosed , DateOfStatus) and their corresponding 628,634 transactions with 10 attributes (Acc no, Acc type, Comp/Ind (Company Customer and Individual Customer), CurrCode, DateOfTxn (date of transaction), Dr/CrFlag (Debit and Credit transaction), OperCode, DrForeignAmt (Debited amount in foreign currency), DrLocalAmt (Debited amount in local currency), CrForeignAmt (Credited amount in foreign currency), CrLocalAmt (Credited amount in local currency)) from the bank. He followed CRISP-DM methodology to conduct the Data Mining process the tool were WEKA for modelling. He used SMOTE (Synthetic Minority Oversampling Technique) has been applied to minimize the class imbalance problem. He used three modelling techniques these are J48, Logistic Regression, and Bagging four. The researcher evaluated a model by F-Measure values. The test result

shows that J48 modelling technique is the best model with a performance of 94.8% followed by bagging (93.9%) and Logistic Regression (76.6%).

Application of Data Mining Techniques for Customers Segmentation and Prediction: in The Case of Buusaa Gonofa Microfinance Institution [9]: He used clustering techniques that resulted in the appropriate number of clusters using data of 13057 customers with 6 attributes. He followed CRISP-DM methodology to conduct Data Mining process by employing the tool WEKA for modelling. Then, a predictive model was developed to predict potential customers where the researcher used K-means, J48. Result indicated that this predictive model achieved an accuracy of 99.95%.

Table 2.1: - Summary of literatures reviewed on customer churn prediction

No	Researcher/ Author	Objective	Classification Technique & Model used	Sample size, Attributes, & Accuracy	Dataset	Gap
1	Kassahun G/meskel	To obtain the best model which can predict the behavior of churning customers	J48, Logistic Regression, and Bagging four, tools used WEKA	13172, 9 Attribute, & 94.8%	CBE	Sample size & Attributes are very small. Accuracy need improvement
2	Belachew Regane	To segment and predict profitable customers of BG MFI for better customer relationship management	K-means, & J48 tools used WEKA	13057, 6 Attribute, & 99.95%	BG MFI	Sample size & Attribute are very small need improvement But accuracy is better than others
3	A. O. Oyeniyi & A.B. Adeyemo	Customer Churn Analysis In Banking Sector	K-means, JRip algorithm WEKA	4958, 8 Attribute,& not mentioned	In Banking Sector	Sample size & Attributes are very small. need improvement
4	Keramati, Abbas	Model for customer churn from electronic banking Services	Decision Tree model	4383, Attribute & Accuracy not mentioned	In Banking Sector	Sample size very small need improvement
5	Chantine Huigevoort	Customer churn prediction for an insurance company	DT, SVM, Logistic Regression	10 000, Attribute not mentioned, & 90%	In Banking Sector	Sample size very small & Accuracy need improvement

2.6.3 Gap Analysis

It is known that a number of commercial banks are in their business action today throughout the world. Banks have their own product design and service delivery strategies [3]. Banks, especially commercial banks in Ethiopia face different challenges and different customer characteristics related to identifying and eliminating customer churns in early stage. As stated in CBE website, one can say that the problem of analysis of CBE customer's characteristics and identifying reasons of customer churns from one banks over another is because of their different types of product design and strategies of customer handling.

Although only one study in Ethiopia has explored in the area of customer churn prediction, most of the other are base in different foreign commercial banks and other industries. A direct implementation of most of those studies output is not practical in specific to CBE. Therefore, this research is conducted based on CBE.

The research Attempt to review different literatures related to customer churn prediction for different commercial banks. In addition, existing studies used different approach such as Different attributes in terms of products and customers characteristics, different datasets and preprocessing activities, and different method for their studies, because of most commercial banks follow different way of handling customers and services provided, strategies and banks structure's and strategies throughout the world.

In Related Works above, suggest that the results will improve and more powerful model could be developed by including different attributes related to customer characteristics and different datasets. In addition to researcher Chantine H [30]., indicated that further improvement can be done and tested through clustering the dataset according to additional attributes like Demographic data of customers, and building models that are cluster specific by categorizing premium and ordinary customers since chances of churn will certainly differ for both groups and by relatively broader datasets.

In this research the problem of customer churn prediction was addressed using specific CBE customer's characteristics which are significant in real world analysis of customer churn practice but not included in previous studies like those of Kassahun G. [8]., Oyeniyi & Adeyemo [27]., Keramati A [29]., & Chantine H [30] to make better prediction and more relevant and practical solutions specific to CBE. The above studies

were conducted on banks customer churn analysis; identified customer characteristics; products characteristics and attributes for customer churn analysis assessment; and identification of churn and Not Churn customer status. Such attributes include Demographic data of customer, products of CBE, customer class, location of customers are included in this research. In addition to that, the research used different datasets (not used before for related works) with balanced and relatively large sample datasets with adequate representation of all geographical locations in Ethiopia by taking the data from CBE branches in different cities unlike to the prior studies. In general, as stated above, this research uses four times more sample size (54,623) datasets than all the other researchers mentioned here in the study and 25 more attributes on Demographical data of customers and customer class and also the research used a new datasets from CBE that is not used before for related research for Customer churn prediction to improve the prediction model, accuracy and precision. The model to be developed will be relevant and practical for Commercial bank of Ethiopia directly.

CHAPTER THREE

DATA PREPARATION

3.1Overview

Data preparation is some of the primary tasks that highly determine the Machine Learning results. The model built mainly depends on how thoroughly and carefully the necessary data is obtained, analyzed and preprocessed. After description of the original data the next subsequent sections present the Business understanding has selected relevant attribute, Data understanding construct the task for relevant attribute used in customer prediction, and preprocessing tasks done to clean attribute in the dataset were performed.

3.1.1 Description of the original data

To conduct the ML project successfully, the researcher try to include all available information about customer related data which are recommended by the Bank managers, Expertise, and literatures. Such as:

- Customer Demographics data.
- Customer Behavioral data.
- Customer Transactional data.
- Customer Account data.

Customer Demographics: is the geographic and population data of a given customer or, information about a group living in particular area. As listed customer-related Demographic predictors (e.g., income, age, sex, region, education status, marital status, Job type, nationality, occupation [33].

Customer behavior: is any behavior related to a customer's bank account. As different studies indicated Behavioral traits in our everyday activities may better explain the phenomenon under investigation. It used to help make key business decisions via market segmentation and predictive analytics. This information is used by businesses for direct marketing, site selection, and customer relationship management. Such predictors are as product, ownership, industry, target, and etc.

Customer Transactional Data: Transactional data is broadly defined as information that records exchanges of number of debits and credits. The transaction details such as date, location, and amount spent, how much amount is debited or credited information and even the category the transaction were left in the research because it has been done by the researcher Kassahun G [8].

3.2 Business Understanding

This phase is also known as Problem understanding and entails the processes used to comprehend the opportunities and business purposes of the company. Business understanding is the second phase of a data preparation, before approaching on data and tools. Understanding of the business domain under investigation is a very important step for clearly stating the business objectives, the ML goal, and for assessing the current situations of the business.

To understand the business domain of Commercial bank of Ethiopia and coin Machine Learning problems- the researcher acquired knowledge through different techniques such as

- Observe CBE's contact center, branches.
- Interviews and discussion also made with CBE senior managers, Domain expert consultation had been made to have brief understanding on the problem area.
- Collected dataset and information from CBE Database and online sources of CBE public website.
- Written documentation used such as fly paper, procedure, Report, and Magazine.

3.2.1 Definition of Customer in CBE

The customer is the person or organization who purchases and uses products and services; customer is a person or legal entity that has a business relationship with the bank; and customer is a person or legal entity who turns to one of the units of a bank to ask for operations and use one, some, or all services offered by banks.

Types of Customers in CBE

Customer: Refer to legal person or natural person with whom the bank agrees to conduct business.

Prospect customer

- Refer to a potential customer in which the bank assumes to deal with business.
- May be recruited during cross selling or any other marketing activities until the customer fulfill the eligibility criteria.

The Prospect and/ or Customer Can be either individual or corporate.

- **Individuals** are a natural person who are illegible for doing business with the bank
- **Corporate** is a legal person or an entity that can do business with the bank.

3.2.2 CBE's customer classification

For the purpose of close follow-ups, proper documentation, and customer due diligence CBE classifies its Customers in to two dimensions:-

- Economic Ownership Dimension
- Volume and value of transaction

3.2.2.1 Economic Ownership Classification

- **Private: -** Individuals, Private Businesses and Other Private entities such as UN, AU, EU, IGAO, NEPAO, Embassies and Funds collected from the community, donors, and other institution which are not legally formed but facilitated by Government entities.
- **Government: -** Federal Government Entities and Regional Government Entities.
- **Public: -** Public enterprises established by proclamation and share companies established by Government.
- **Associations: -** Cooperative societies and Charities, societies and Associations.

3.2.2.2 Volume and Value of Transaction

This subsection volume and value of transaction is classified in to three parts.

I. Retail /consumer/customers

II. Business customer

III. Premium customers

Table 3.1 Customer Classification

Customer Type	Classification Requirement
Retail Customers	Yearly Average Transaction < Birr 100,000.00
Business Customers	Yearly Average Transactions Between Birr 100,000.00 and Birr 1,000,000.00
Premium	Yearly Average Transactions Birr 1,000,000.00 and above

Figure 3.1 below shows the customer type explained in the above section in details of Deposit products and Service in CBE that is Local currency deposit product and foreign currency deposit product.

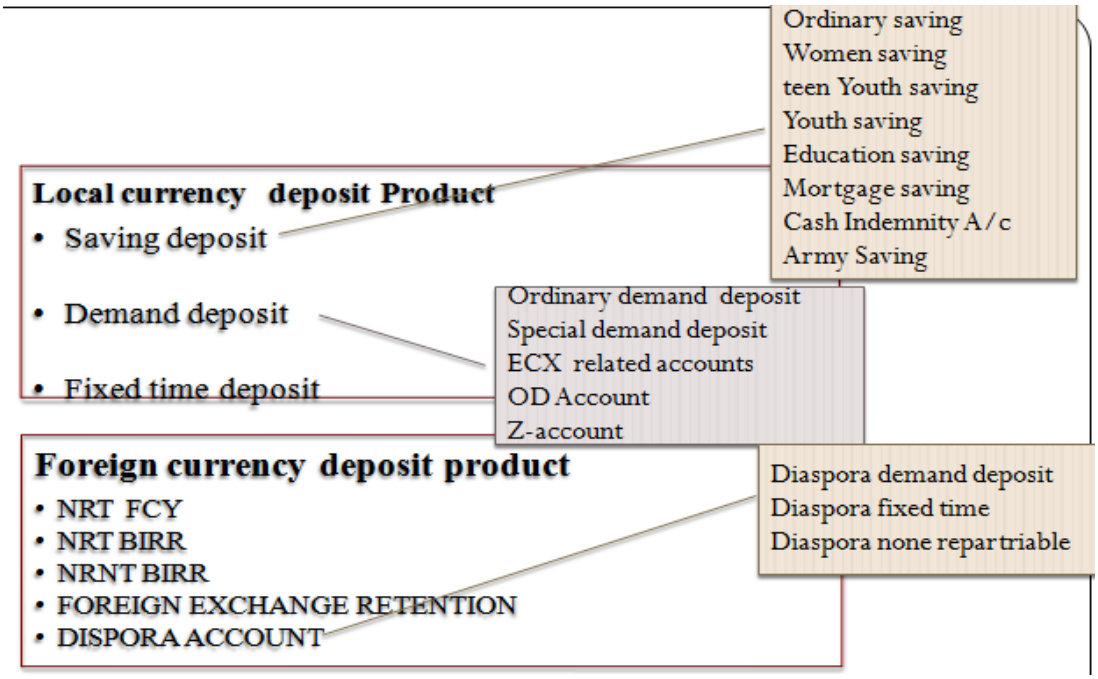


Figure 3.1Local Currency and Foreign Currency Deposit Product services in CBE [3].

3.2.2 Nature of Customer Acquisition by Commercial Bank of Ethiopia

Promotion is a marketing effort whose function is to inform or persuade existing or potential customers about merits of a given product or service and inducing them either to start or continue purchasing firm's products or services. In line with this fact, CBE has been aggressively promoting its products and services to expand its customer base and enhance sales revenue with high cost.

CBE employed different promotional channels to ensure product/services and brand awareness to new, existing and potential customers. Common Promotional mixes includes Public Relation, Advertising, Personal selling, and Sales promotion.

- **Public relation** is one of the promotion tools in CBE. It is mainly used for increasing the banks' image and helps minimize the effect of any information that erodes the bank's image. In this respect, donation and membership are among the activities CBE has been taking part. In CBE's context, the purposes of sponsorship among others are to support business objectives and strategies of the bank, promote product/services of the bank, and reinforce CBE's brand and corporate image.
- **Advertising:** In this regard, conventional media like television and radio are familiar. On the other hand, print ads (magazines, brochures and pamphlets), display ads (signage and screens) and billboard advertisings are among the widely used advertising elements in the bank. However, direct marketing has been rarely used in CBE. There are also Sales Promotion Campaign such as Prize Linked Saving, Fiscal Year (FY) Linked Incentives, Advertising, and publicity expenses.

CBE (1993-2015) ('in million birr')

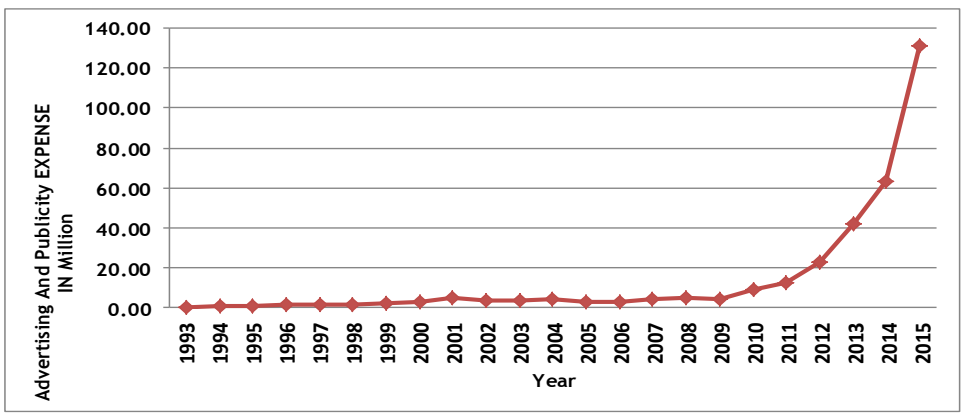


Figure3.2: Annual report of the CBE [3].

Majority of the promotion budget is allotted at the head office level to the Communication, and Promotion and Brand management sub- processes. In addition, districts have their own budget to accomplish their promotional works. Similarly, the E-payment and the TS have their own budget. Figure below shows that promotional expenses incurred by districts; TS, E-payment and Business Development have increased over the past five years. Advertising and publicity expenses of HO, Districts, IBD and E-payment (‘in million birr’).

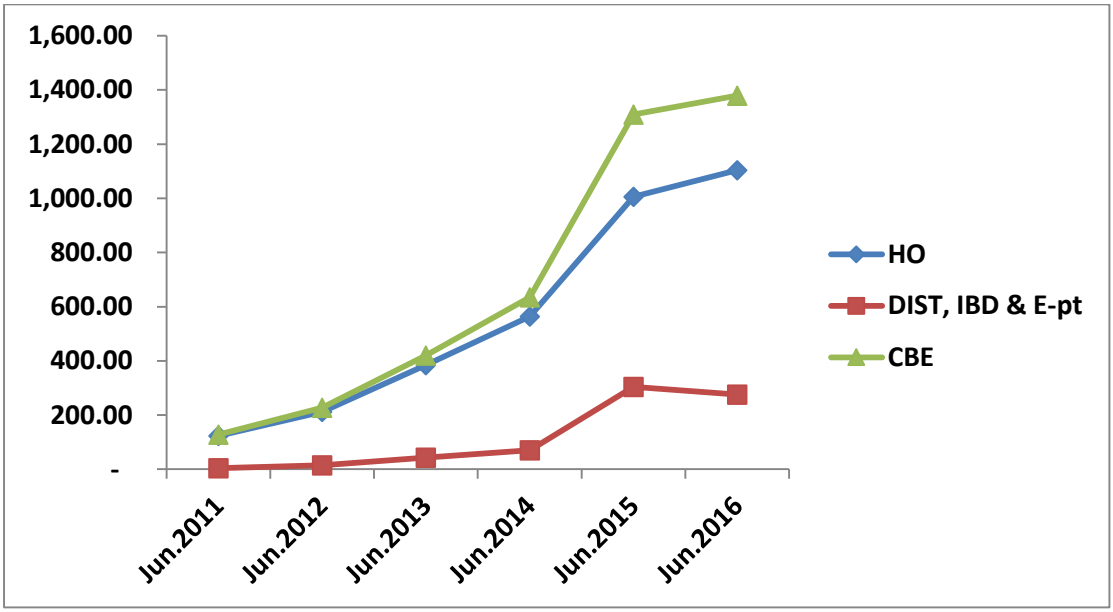


Figure 3.3: Advertising and publicity expenses of CBE [3].

3.2.3 Handling Customer Accounts

3.2.3.1 Handling of Inactive Accounts

- Current and savings accounts, which remain idle or dormant with no movement for a period of six and twelve consecutive months respectively, shall be changed by the system to “Inactive accounts,” regardless of their balances;
- However, these inactive accounts can be reactivated if they show a minimum of three debit transactions in three consecutive months. But, the Customer Service Manager can change the status from inactive to active less than three debits when he/she feels the customer is properly identified and operates his/her account regularly.
    - The customer adequately explains about the previous transactions ;and

- Minimum of four transactions in the last half six months are conducted, excluding interest transactions.

- Inactive current accounts with a balance of less than Birr 500.00 are subject to service charges as per the terms and tariff semi-annually;
- Accounts that carry balances of below the amount of the charge will be closed as part of the collection of the service charges thereon;

3.2.3.2 Handling of Closed Account

- The minimum balance to open a savings account should be Birr 25.00. However, Savings account may be opened and maintained with zero balance for one month.
- Within one month, the account holder should credit at least a minimum of Birr 25.00 in his/her account. However, if the account is not credited in the specified period, it should be closed following the normal account closing procedure.
- Newly opened accounts should be separately kept for three months under the account status assigned for. However, the Customer Service Manager can change the status before the stated period if he/she believes that the customer is properly identified.

As CBE is serving large number of customer they needs regular follow up and addressing their requests on a timely basis is obligatory for business continuity and effective strategy execution. To do this, the Bank has established customer contact center which helps to deliver prompt answer for customers' inquiries. The Contact center has been providing information to external customers through short codes 951. It is the primary customer facing channel to resolve customer complaints and promote the bank new services through. Its key tasks among others are to manage complaints from customers and other stakeholders; receive and respond online or on the spot and receive and refer to the concerned Process or Sub Process.

On the other hand to gain insight into the behavior of customers and the value of those customers CBE is implementing oracle's sizeable CRM system to automate its core support processes. But yet the system is not well organized and did not to start extract knowledge because there is no link between IPCC, customer data base and oracle sizeable.

As the researcher observed, interviews and discussion made with senior managers of CBE, even if CBE is attempting to attract and retain more customers while expanding the market to outreach potential customers and try to make business more successful with latest tips and updates of different technologies, Promote new products and services, until now, the number of closed account has been increasing day by day.

As the researcher understood from the discussion made with managers and operators even if the bank is implementing different technology the bank doesn‘t have an instrument for collecting feedback from churners as to why they are closing their account and customers did not communicated earlier before he/she made a decision. In general as they said there is a very poor traditional means of knowing what is unique features and need for their customers. In CBE branches currently there are a Marketing team for each and all branch do their internal and external marketing business by their own while marketing team do research they do not use appropriate tools and techniques which help them to identify their customers in the market quickly and the root causes of historic churner. When they analyze data they only use excel and the research will took two up to three month.

So, the bank could not pass proper decisions to retain customers, satisfy their customers in providing different services, adding new product and services, based on customer's behavior to stay customers loyal for long period of time.

If the Bank is small the operators can call on customers to understand why their customers are leaving and not made transaction. But such approaches become unmanageable for the bank. CBE which have more than 33.3 million account holders. Additionally, customers may not always disclose the real reasons they are leaving.

Based on the above fact and problems the existing system requires significant improvement and therefore, the researcher initiated to propose a Machine Learning technique which would help to classify and predict profitable customers, so that the bank can pass for proper decisions. Without data, we can do very little to understand customer turnover. So the miner have to capture not just the customer Demographics, the plan type, transaction type and the price charged but also what marketing channel the customer came from, his/her usage metrics and support data. And, most importantly, it has to bring it all together into one consistent view. Because, the study is applying ML to build a model using available customers ‘data in order to predict

those customers who are going to close their accounts. This in turn has a significant impact in improving customer relationship management of CBE next to this we select relevant available attributes.

3.2.4 Attribute Selection

Features Selection is one of the core concepts in data mining which hugely impacts the performance of your model. The data variables that you use to train your models have a huge influence on the performance you can achieve. Variables selection and Data cleaning should be the first and most important step of your model designing. Variables Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in [34].Feature selection is a technique that removes noisy and redundant features to improve the accuracy and generalizability of a prediction model. Although feature selection is important, it adds yet another step to the process of building a bug prediction model and increases its complexity [35].

As stated and suggested in [36] and through made discussion with senior domain experts in CBE we identify, there are a few customers characteristic specific to Ethiopian commercial bank customers and characteristics factor that may be considered as the main factors for customer churn prediction.

This section provides the description of the main factors used to predict the customer churn in Ethiopia commercial bank. The customer's characteristics like Age, marital status, Gender and having other sources of customer class, are the variables that can influence the customer churn's The financial related characteristics include the opening balance, years of customer for CBE, product type, factor also significant in influencing customer churns [35].

Attribute selection from the dataset was done based on the objective of the study at hand. Hence the account number, customer's names and branch code attributes are removed in order to reduce the data to only most important ones; this would minimize the effort required for further processing.

Table 3.2: Attributes of dataset used for CBE customer churn perdition.

No.	Attributes	Description
1	Region Status	Region of customer
2	Current Balance on the Account	Current balance in birr on the account
3	Sex	Gender of the customers
4	Customer minimum opening balance	Customer first product opening balance
5	Marital status	Married or single
6	Categories	Product type (Local Currency and Foreign Currency Deposit Product)
7	Age	Age of the customers
8	Industries	Either individual customer or Agent
9	Target	Retail ,business ,and other customers categories
10	Customer status	Either churn or Not Churn
11	Mobile Banking User	Customers uses mobile banking
12	Internet Banking User	Customer uses internet banking
13	ATM user	Customers uses ATM money withdrawal
14	POS user	Customers uses POS money withdrawal and applied for cashless payment.
15	Grand total number of transaction	Total number of transaction including teller, ATM, POS and internet banking mediums.

3.4 Data Understanding

After understanding the problem to be addressed, the next step was understanding and analyzing the available data. The outcome of Machine Learning and knowledge discovery heavily depends on the quantity of the available data. To do this the researcher included domain expert, data analysts, and Database Administrator for understanding and preparing data for mining. Domain experts and data analysts specify the data required for solving the problem. A Database Administrator for collects the required data and provides it in the format the analyst asked for.

The dataset used in this research was customers' related real data; the data is collected from commercial bank of Ethiopia MIS data warehouse department. The data is

extracted from CORE-Banking system data base in to Excel format by Database Administrator. The dataset consists randomly sampled 54,623 customer's information from June 2017 up to Dec 2021 with 34 attributed (see the attached attribute in Annex part) and two class labels, namely churn and non-churn. The bank as a policy does not expose the privacy of its customers in any way. So, fields such as: Account Number, customer code, Name, Address, Telephone Number and other sensitive information are filtered by the bank's Data Base Administrator themselves. Such data also do not have contribution for the outcomes of the research.

Several fundamental issues related to the data have a significant impact on the quality of the outcome of a knowledge discovery process. In this regard, the initial dataset has been statistically described and visualized using Microsoft Excel to examine the properties of the whole dataset records and to obtain high level information regarding the Machine Learning questions. Simple statistical analysis has been performed to verify the quality of the dataset, addressing questions such as: Does the data cover all cases required? Is the data correct or does it contains errors? Are there missing values in the data? See more about the statistical descriptions of all attributes in the initial dataset.

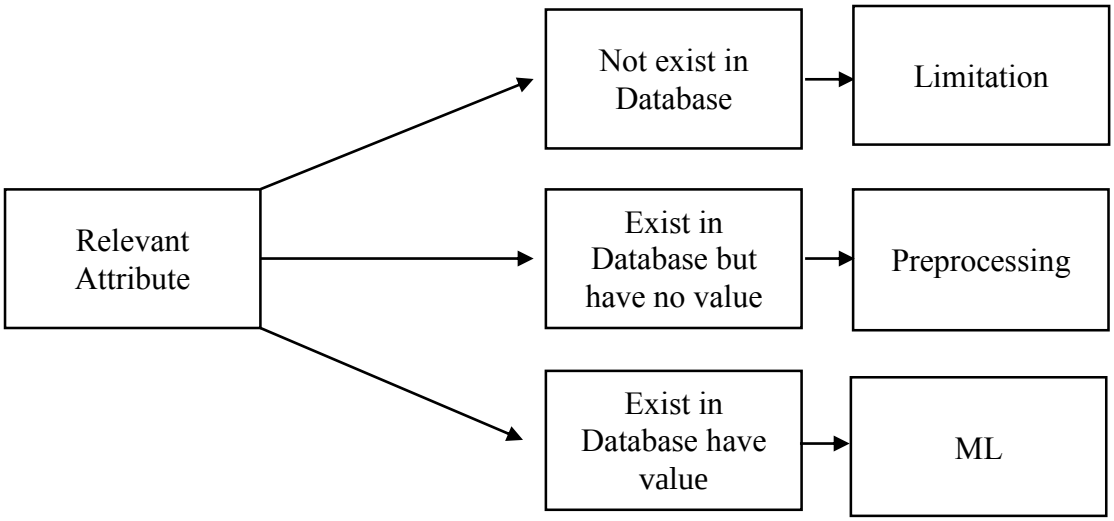


Figure 3.4 Preferred steps of selected Attribute passing through relevant stages.

3.4.1 Data Collection

For this research, all the data was collected from commercial bank of Ethiopia. The data included new attributes which are not considered in prior studies that include

Demographic data of customers, categories of customer's class and total years of customer - ship with CBE. The dataset also included additional attributes such marital status, ownership, industry, product type, and sector. Records contain a dependent field representing either a "churn customer" or "Not Churn customer" field.

The datasets were checked for completeness and correctness of the required attributes and integrity before analysis and prediction. The research explored different non Machine Learning studies related to customer churn analysis and prediction using financial data and customer churn assessment strategies in commercial banks. This helped us to ensure whether attributes (column names) are complete and adequate for prediction of customer churn. For this research the data from CBE contained required customer details for analysis and prediction of customer churn.

3.4.2 Data Representation

Data representation and aggregations of the data is necessary because it minimize the variations of the attribute values in some of the fields and also to make results more expressive and simply interpretable. In the dataset, age attribute is varying in values and it is transformed in to more aggregated values "young, "adult" and old" for the different age groups determined based on experiences and consultation with the domain experts. Hence young is considered to be between 18-25, while adult is considered to be between 26-55 and old age was considered to be those with ages 55 and above.

Similarly, to make it easier for discussions and interpretations of the result, other attributes like opening balance, region and customer class were generalized in to categorical values. The minimum opening balance in CBE is 25 birr so Opening balance below 100 are categorized as Low, those which are between 100 and 500 were categorized as medium and those opening balance above 500 were categorized as high.

From the source of the dataset, the names and values of the attributes have been changed to some generic symbols for the sake of simplicity for experiment and to have a more accurate representation of the variables. The values in the attributes like region status attributes changed to AA, Current balance on his account AB and Sex attributes changed to AC, Customer minimum opening balance changed to AD, Marital status attributes changed to AE, Categories attributes changed to AF, Age attributes changed to AG**,** Industries attributes changes to AH, Target attributes changes to AI and also

customer status changed to AJ. Mobile banking user, internet banking, ATM, POS and grand total number of transaction attributes changed to AK to AO respectively.

The dataset in this experiment and analysis contains categorical values that are transformed to binary values or factors 0s and 1s. Sex variable having values 'M' changed to ‘1’ that representing male and 'F' changed to value ‘2’.

Similarly, target attributes changed to ‘1’ for retail , ‘2’ for bussiness,’3’ for others, and also for industries attributes value changed to ‘1’ for individual,’2’ for agents,

Categories (Product)type attribute changed the variables value to ‘1’ for Women Saving Account, ‘2’ for Special Savings Account, ‘3’ for Youth Saving Account,’4’ for Education Saving Account,’5’ for Amana -Safe keeping saving Depositand,’6’ for others saving products. And region status attributes variables changed to ‘1’ for city administration ,and ’2’ for regions Also, age g attributes value changed to ‘1‘for 18-25,’ and 2’ for greater than 25.

Industries attribute has two values agent and individual, the value of individual variable changed to ‘1’ and for agent changed to ‘2‘’. also Customer minimum opening balance attributes below 100,100 to 500 and above 500 categories changed to ‘1’,’2’, ’3’ values respectively. Current balance attributes also below 5000,5000 to 100000 and above 100000 categories changed to ‘1’,’2’, ’3’ values respectively. and to all Mobile banking user, internet banking, ATM, POS and grand total number of transaction attributes values yes and no changed to ‘1’ for yes and ’2’ for no.

Below it shows the short summary of the changed variable names and values which are used throughout the experiment and analysis.

Table 3.3: Variables and Representation of data types

Attributes	Representation attributes values	Original data type	Transformed data type
region status	AA	Character	Binary
Current balance on his account	AB	Number	Factor
Sex	AC	Character	Factor
Customer minimum opening balance	AD	character	Factor
Marital status	AE	Character	Not changed
Categories	AF	Character	Factor

Age	AG	Number	Factor
Industries	AH	Character	Factor
Target	AI	Character	Factor
Customer status	AJ	Character	Not changed
MOBILE_BANKING_USER	AK	Characters	Factors
Internet banking user	AL	Characters	Factors
ATM user	AM	Characters	Factors
POS user	AN	Characters	Factors
Grand total number of transaction	AO	Number	Not changed

3.5 Data Preprocessing

The data preprocessing refers preparing the dataset in the form that it is ready to Machine Learning task. In this research the processes applied include data cleaning, parsing, data selection and aggregating on the extracted data in order to make the data more suitable for the experiment to improve the overall Machine Learning task.

Data preprocessing which includes Data Preprocessing (Data Cleaning and Data Set Splitting). The Data Preprocessing deployed combining datasets, choosing part of the data as subgroup, combining rows, developing new columns, arranging the data to be used in the modeling, taking care of problematic figures (blank or missing) and dividing into training and test datasets.

3.5.1 Data Cleaning

Here tasks are done to clean attribute that exist in Database but, requires cleaning need to fill the missing value therefore among the total dataset extracted, 420 of them have missed values for attributes such as marital status and age. Instead of filling values on these attributes, it was found easier and more logical to remove the records that make up 0.94% of the dataset. As a result, the remaining 99.06% of the original dataset which amounted to 48,051 records were kept for further processing.

3.5.2 Imbalance Data and Splitting the Dataset

An imbalance dataset is such a case where there is major difference in the number of classification categories [36]. In our dataset domain, the classification categories

consist of “churn customer” and “Not Churn customers”, where the number of “Not Churn customer” cases outnumber the “churn customer” cases. Almost 57 percent of the datasets are non-churn customers (Active). In such a situation a model becomes more inclined to the majority class and cannot properly identify the minority class. To solve this issue, the possibilities are either we can over sample the minority class or under sample the majority class. But under sampling the majority class will act as a difficulty in properly understanding the trends in our independent attributes. Furthermore, only over sampling the minority class will not solve this as the techniques lying behind the over sampling which also matter greatly. Thus, in such a scenario the research used Synthetic Minority Oversampling Technique (SMOTE) and up sampling. This technique is used for both oversampling and under sampling [37]. Synthetic instances of the minority class are created to reduce the margin between the majority and minority class [25]. For the model the research used up sampling to increase the minority class and keep equal number of churned and non-churned. But we have to mention that up sampling was only applied on the training set keeping the test set pure and untouched. And therefore, this helped us to properly classify the borrowers keeping the model aware of both the output classes.

For the purpose of checking the performance of any Machine Learning model in an effective manner, splitting the dataset is a fundamental task. It helps to prevent over fitting by evaluating the performance of the model on a portion of the dataset upon which the model has not been trained. In most empirical studies like [37].it is shown that the dataset is split in to 80:20 ratios which has become most common in many other studies. Therefore, this research used 80:20 train-test split ratio for the supervised model. This has been done using R studio “split_test_train” function from ISRL (Data for an Introduction to Statistical Learning with Applications in R) library. This means that 80 percent of the entire dataset was used to train the model and the remaining 20 percent was used to evaluate the performance of the model.

CHAPTER FOUR

EXPERIMENT AND DISCUSSION OF RESULTS

4.1 Overview

This chapter discusses how the experiment was carried out to construct a model with their analysis based on the steps mentioned in proposed framework architecture. This experimental evaluation ensures the realization of the proposed framework architecture. It presents how the machine learning models were created using SVM, Logistic regression, Navies Bayes and KNN, the major experiments run, interpretations and their performance evaluations of the prediction model. Several subsequent tasks in the experiment are done using R studio tools. The experiments were carried out using the trained and tested with R programming a on a Toshiba Satellite Processor Intel(R) Core(TM)i3-2310M CPU with 2.10GHz, 2.10GHz speed, 3 GB of RAM, 500GB of Hard Disk capacity, with 64-bit Windows 7 operating system.

In chapter Three, all the preprocessing activities performed on the dataset and some of the major tasks performed were presented. This section present proposed framework architecture and focuses on presenting summary of the major experiments made in the process of arriving at the optimal model to achieve the objective set in chapter One.

Once the necessary data is passed through the preprocessing activities as described in the earlier sections, it is then loaded to the R studio for the model building required. The preprocessed data is converted to *csv* format that is suitable for R studio.

Series of Experiments are conducted based on which algorithms prediction models with varying accuracies, sizes and precisions are obtained. This section also presents several activities done related to running and evaluating model building experiments, selecting the best and appropriate model, and providing explanations on the selected model.

4.2 The Proposed Architecture

The proposed architecture in this research illustrated in Figure 4.1.The proposed architecture shows the steps followed in the Prediction and Analysis of Customer status from CBE Customers dataset. In the proposed architecture, the Customer data were taken as input and stored in the dataset and this dataset contains the CBE Customers

data list who are currently active and not active status. To validate the proposed model, we use the CBE Customers dataset. The proposed architecture passes through six main stage that come after Business Understanding meaning understanding the problem to select the relevant attribute and Data understanding it construct tasks for relevant attribute used customer prediction, then Data Preprocessing tasks to clean attribute that exist in datasets, Data Splitting: 80 percent for Training of the entire dataset used to train the model and 20 percent for Test dataset used to evaluate the performance of the model, Classification that used four supervised ML such as SVM, Naïve Bayes, KNN, and Logistic Regression to obtain the Train model, then Train model and the Test dataset are the same pass to the customer prediction output, finally customer prediction model predict customer churn or Not Churn.

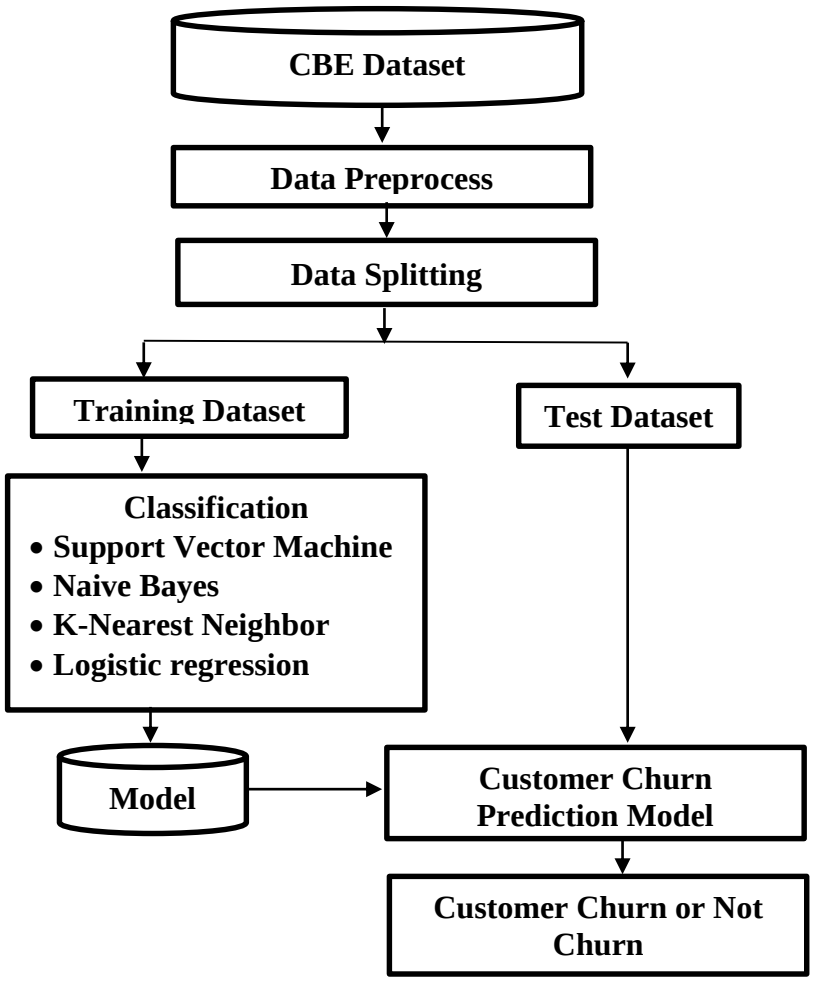


Figure 4.1 proposed architecture for bank Customer Churn Prediction Model

4.2.1 Predictive Modelling

Predictive modeling is the general concept of building a model that is capable of making predictions. Typically, such a model includes a Machine Learning algorithm that learns certain properties from a training dataset in order to make those predictions [38].

Predictive modeling is a name given to a collection of mathematical techniques or models that helps in finding a mathematical relationship between a target or dependent variables and the predictor or independent variables [39]. It helps in predicting the probability of an outcome when a set of independent variables passes through the model. KNN, SVM, Logistic Regression and Naïve Bayes Models can be used for prediction purpose.

The principal goal of this research is to analyze the existing CBE customer behavior and predict customer churns using computational algorithms. With this in mind, the research aimed at identifying Machine Learning technique which is better in predicting customer churn. Throughout this work different Machine Learning algorithms were explored and effective ones were used to find model in the data. The performance of each Machine Learning models was explored and analyzed in previous related works. Then based on successfulness in making predictions and stability in their best performing algorithms were selected. The Machine Learning techniques selected for this thesis were KNN, SVM, Logistic Regression and Naïve Bayes. Each one is selected based on their advantages and past performance seen in other research. In different literatures, it has been reported that they were widely used classifier algorithms for prediction and classification are KNN like in [40], [41], [42], SVM like in [43], [44], [42], [45], and Logistic Regression like in [46] and Naive Bayes like in [47], [48], and [45]. The research was use the above four different algorithms to build four different models for this customer churn prediction and classification model.

Naïve Bayes is selected due to the following reasons; it is easy to implement, Naïve Bayes classifiers can be trained quickly [47], classification process is quick compared to other models [48] [45], it can handle a large and discrete amount of data, it is not sensitive to irrelevant features [42].

Logistic Regression which is also called logistic model or Logistic Regression is a predictive analysis. It takes independent features and returns output as categorical

output. The probability of occurrence of a categorical output can also be found by Logistic Regression model by fitting the features in the logistic curve [38]. Logistic Regression is included in these work because; it is easy to implement and no linear relationship between independent and dependent variable [45], multiple explanatory variables can be used, no confounding effects because Logistic Regression allows quantified values for strength of association between explanatory variables and less prone to over-fitting due to simplicity and low variance [46].

K-nearest Neighbor is selected because; it is difficult to imagine a simpler technique than KNN, where data is classified simply based on its nearest neighbor (or neighbors) in a given training set [42]. Learning does not require making any assumption about the characteristics of the concepts [40] and Complex concepts can be learned by local approximation using simple procedures [42].

SVM is selected to this research because; it is established on the structural risk minimization principle, which seeks to minimize an upper bound of generalization error, and is shown to be very resistant to the over-fitting problem [43] [45]; it uses the kernel trick, so can build in expert knowledge about the problem via engineering the kernel [45]. SVM model is a linearly constrained quadratic program so that the solution of SVM is always globally optimal, while other models may tend to fall into a local optimal solution [44].

4.3 Dataset for Experiment

For this study, the data is collected commercial bank of Ethiopia. As stated in [49]. And also the study specifies the main reasons (attributes) of customer churn), the data from CBE contains customers Demographic characteristics and financial characteristics attributes. The data contained 48,051CBE customer records with 15 attributes one of the attributes is dependent field representing either a churn or not a churn field; the data collected from CBE as of December, 2020 G.C. No single prospective customer is contacted more than once. The attributes, description and categorical values explored for the study are described.

4.4 Modeling using KNN

This subsection deals with how the KNN prediction model is developed and how the information gained was calculated. In this experiment a Prediction model building is done using KNN algorithm.

4.4.1 Experiment One

The dataset consists of 48,051 observations and 15 attributes with one dependent attribute/class (which is customer status) and the remaining predictor (independent) attributes.

4.4.1.1 Normalizing numeric data

Normalization is a technique often applied as part of data preparation for machine learning. The use of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information [50]. This feature is important since the scale used for the values for each variable might be different. The best practice is to normalize the data and transform all the values to a common scale.

> "normalize <- funciotn(x) {return ((x –min(x)) / (max(x) –min(x)))}"

>"CBED_n <- as.data.frame (lapply (CBED [2:15], normalize))"

After run the above code in R studio, it normalized all numeric features in the dataset, instead of normalizing each of the 14 individual attributes.

The final result is stored to CBED_n data frame using as.data.frame () function.

>normalize <- function(x) {return ((x –min(x)) / (max(x) –min(x)))}

>"CBED_n <- as.data.frame (lapply (CBED [2:15], normalize))"

> "Summary (CBED_n)"

4.4.1.2 Creating Training and Test Dataset

The KNN algorithm is applied to the training dataset and the results are verified on the

test dataset. In this study, we divide the dataset into two in the ratio of 80:20 for the training and test dataset respectively. Then CBED_n data frame divided into CBED_train and CBED_test data frames. Our target variable is "AJ" (Customer status) which is not included in the training and but not in test datasets.

> CBED_train <- CBED_n [1:29400,]

> CBED_test <- CBED_n [29400:48,051]

> CBED_train_labels <- CBED_n [1:29400, 1]

> CBED_test <- CBED_n [29400:48051, 1]

4.4.1.3 Training a Model on Dataset

To use knn() function for training a model we need to install a package 'class'. The knn() function identifies the k-nearest neighbors using Euclidean distance where k is a user-specified number. The following command type uses knn().

```
> install.packages("class")
Installing package into 'C:/Users/hp/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/class_7.3-15.zip'
Content type 'application/zip' length 106670 bytes (104 KB)
downloaded 104 KB

package 'class' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\hp\AppData\Local\Temp\RtmpOYelqt\downloaded_packages
> library(class)
```

After library (class) is download and install in R, it is ready to use the knn() function to classify test data. KNN Algorithm is based on feature similarity. Choosing the right value of k is important for better accuracy. As suggested in [40], [41], [42] , the best way to choose the value of K are the square root of the average number of complete cases (the number of observations) (k=sqrt(n)). The reason they suggest this model is that k should be large enough to give a reliable result. The Experiment One conducted by using 200 as a value of K is around the square root of the observations 48,051.

knn() returns a factor value of predicted labels for each of the examples in the test dataset which is then assigned to the data frame CBED_test_pred.

> CBED_train <- CBED_n[1:29400,]

```
> CBED_test <- CBED_n[29400:48,051]

> CBED_train_labels<- CBED_n[1:29400,1]

> CBED_test  <- CBED_n[29400:48,051,1]

>CBED_test_predE1<-knn(train = CBED_train ,test = CBED_test,c1 =
CBED_train_labels, k= 200 )
```

4.4.1.4 Evaluating Model Performance

The model performance was evaluated using Accuracy and Precision. The model was built through KNN algorithm but we also need to check the accuracy and precision of the predicted values in CBED_test_pred as to whether they match up with the known values in CBED_test_labels. To ensure this, we use the CrossTable() function available in the package 'gmodels'.

The test data consisted of 7,980 observations. Out of 7,980 cases, have been accurately predicted (TP->True Positive) as Not Churn in nature which constitutes 7418. Also, 7 out of 48,051 observations were predicted (FN-> False Negative) as Not Churn in nature but got predicted as Churn which constitutes 0.01%.

There is no false positive prediction; that is, churn is not predicted as Not Churn. Also, 555 out of 48,051 observations were predicted as not active (churn) correctly identified not active (churn) in nature is (TN -> True Negative). Accuracy= ((7418 + 555)/7980) as a result KNN with k= 200 achieves 99.91 % and precision of 100%.

4.4.1.5 Improve the Performance of the Model

This can be taken into account by repeating the steps, training a model and Evaluation of model performance by changing the k-value. Generally, the K value is the square root of the observations and in this case, we took k=199 and 201 which is around a square root of 48,051(200). The k-value may be fluctuated in and around the value of 200 to check the increased accuracy of the model and to keep the value of False Negative as low as possible.

4.4.2 Experiment Two

4.4.2.1 Training a Model on Dataset

The experiment two conducted only by changing the value of K to 199 that is around the square root of the observations 48,051 (200).

```
> CBED_train <- CBED_n[1:29400,]

> CBED_test  <- CBED_n[29400:48,051]

> CBED_train_labels   <- CBED_n[1:29400,1]

> CBED_test  <- CBED_n[29400:48,051,1]

> CBED_test _predE1 <- knn(train=CBED_train, test=CBED_test, c1=
CBED_train_labels, k=199)
```

4.4.2.2 Evaluating Model Performance

The test data consisted of 7,980 observations. Out of which 7,980 cases have been accurately predicted (TP->True Positive) as Not Churn in nature which constitutes 7,418. Also, 10 out of 7,980 observations were predicted (FN-> False Negative) as Not Churn in nature but got predicted as Churn which constitutes 0.01%.

There is no false positive prediction again that is churn in nature and predicted as Not Churn. Also, 552 out of 7,980 observations were predicted as not active (churn) correctly identified not active (churn) in nature is (TN -> True Negative). As a result, KNN with K=199 achieves 99.87% accuracy and 100% precision.

4.4.3 Experiment Three

In experiment one and tow, the accuracy and precision of the model is almost close to equal, in this experiment three perform by changing the value of K to 201 that is around the square root of the observations 48,051 (200).

4.4.3.1 Training a Model

> CBED_train <- CBED_n[1:29400,]

> CBED_test <- CBED_n[29400:48,051]

> CBED_train_labels <- CBED_n[1:29400,1]

> CBED_test <- CBED_n[29400:48,051,1]

>CBED_test _predE1<-knn(train=CBED_train, test = CBED_test, c1= CBED_train_labels, k=201)

4.4.3.2 Evaluating Model Performance

The test data consisted of 7,980 observations. Out of which 7,980 cases have been accurately predicted (TP->True Positive) as Active in nature which constitutes 7,418. Also, 10 out of 7,980 observations were predicted (FN-> False Negative) as Not Churn in nature but got predicted as Churn which constitutes 0.01%.

There is no false positive prediction again in experiment three, that is churn in nature and predicted as Not Churn. Also, 552 out of 7,980 observations were predicted as not active (churn) correctly identified not active (churn) in nature is (TN -> True Negative). As a result, KNN with K=201 achieves 99.87% accuracy and 100% precision.

4.4.4 Experiment Four

In this experiment four we test different k values to cheek improvement performance of the model. The following R code shows accuracy level for different K values (185-205).

```
> plot(k.optm, type="b", xlab="K- Value",ylab="Accuracy level")
```

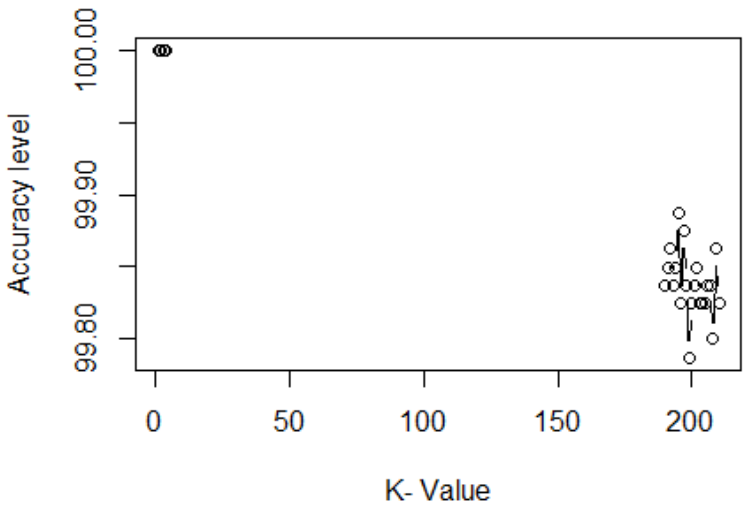


Figure. 4.2: The training model of different K values and result plot

4.4.5 Attribute (variable) Importance

In this section, to identify attribute importance we use the earth package in R to test variable importance based on Generalized cross validation (GCV), number of subset models the variable occurs (nsubsets) and residual sum of squares (RSS).

As observed, eleven attributes are listed as importance variables out of 14 independent attributes, but the two attributes that contributed massively in the prediction are Grand total number of transaction and industries.

4.5 Modeling using Logistic Regression

In this experiment a Prediction model building is done using Logistic Regression algorithm. Logistic Regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name Logistic Regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No [25]. The main thing here is to determine a mathematical equation that can be used to predict the probability of event 1(Not Churn). After the equation is established, it can be used to predict the Y (customer status) attributes when only the X's (other 15 attributes) are known using selected CBE dataset through "mlbench" package.

4.5.1 Experiment one

4.5.1.1 Install additional needed Package in to R studio

Install “mlbenh” package and load the data into R studio and keep only the complete cases.

```
> install.packages("mlbench")
Installing package into 'C:/Users/hp/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/mlbench_2.1-1.zip'
Content type 'application/zip' length 1059296 bytes (1.0 MB)
downloaded 1.0 MB

package 'mlbench' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\hp\AppData\Local\Temp\RtmpCK6Bu8\downloaded_packages
> LG = read.csv(file.choose() , sep = ',')
> View(LG)
```

The dataset has 48,051 observations and 15 attributes. The customer status column is the response (dependent) attributes and telling is it churn or Not Churn.

The following code in R enable to convert the response variable customer status to a factor variable and all other columns to numeric.

```
> LG$AA <- ifelse(LG$AA == "active", 1, 0)
> LG$AA <- factor(LG$AA, levels = c(0, 1))
```

4.5.1.2 Balance the Imbalance Dataset

It is known that the data was split into 80:20 training and test samples ratio. As the response attributes (customer status) is a binary categorical variable, there is a need to balance the training data into equal proportion of classes.

The customer status churn and Not Churn are split approximately in 1:2 ratios. As shown in above code result, clearly there is a few customer status class imbalances. So, before building the Logistic Regression model, it is needed to build the samples such that both the 1's and 0's (churn and Not Churn) that are in approximately equal proportions. This concern is normally handled with a couple of techniques called: Down sampling, Up sampling and Hybrid Sampling using SMOTE and ROSE [51]. In Down sampling, the majority class is randomly down sampled to be of the same size as the smaller class. That means, when creating the training dataset, the rows with the churn Class will be picked fewer times during the random sampling.

Similarly, in Up Sampling, rows from the minority class, that is, churn is repeatedly sampled over and over till it reaches the same size as the majority class (Not Churn). But in case of Hybrid sampling (SMOTE and ROSE packages), artificial data points are generated and are systematically added around the minority class [51]. In this study we use up sampling techniques to balance the imbalance data. So first we create the Training and Test data using "caret" Package in R Studio.

```
> library(caret)
Loading required package: lattice
Loading required package: ggplot2
Warning messages:
1: package 'caret' was built under R version 3.5.3
2: package 'ggplot2' was built under R version 3.5.3
```

```
> '%ni%' <- Negate('%in%')
> options(scipen=999)
> set.seed(100)
```

```
> trainDataIndex <- createDataPartition(LG2M$AA, p=0.7, list = F)
> trainData <- LG2M[trainDataIndex, ]
> trainDataIndex <- createDataPartition(LG2M$AA, p=0.8, list = F)
> trainData <- LG2M[trainDataIndex, ]
```

```
> testData <- LG2M[-trainDataIndex, ]
```

The above R studio code snapshot shows installed caret package and used the "createDataPartition" function to generate the row numbers for the training dataset. As stated in the above code p=0.8, that is 80% of the dataset rows to go inside "trainData" for training the model and the remaining 20% to go to "testData" for test. There is around 2,000 rows more Not Churn samples than churn sample. So, it needs up sampling to balance the dataset using the "upSample" function in R. As a result of the above up sampling code result, churn and Not Churn status are now in the same ratio.

4.5.1.3 Interpretation of the Logistic Model Result

After building the Logistic Regression model, then we analyze the fitting and interpret what the model is telling us. In above R code Logistic Regression model result, the "glm" function internally encodes categorical variables into n – 1 distinct level. The column "Estimate" represents the regression coefficients value. Here, the regression coefficients explain the change in log (odds) (The odds are itself the ratio of two probabilities, p and 1-p) of the response variable for one-unit change in the predictor variable.

The column "Std. Error" in above logistic model result represents the standard error associated with the regression coefficients. And z value is analogous to t-statistics in multiple regression output. Z value > 2 implies the corresponding attributes is significant.

P value determines the probability of significance of predictor variables. A variable having $p < 0.5$ is considered an important predictor [25]. A predictor that has a low p-value is likely to be a meaningful addition to the model because changes in the predictor's value are related to changes in the response variable [25].

In the existence of other attributes, attributes like sex, Marital status, Age, are statistically not significant ($p>0.5$). As for the statistically significant variables, Grand total number of transaction, POS user, ATM user, Internet banking user, Mobile Banking User has the lowest p-value that indicate a strong association of those attributes to the customer with the probability of having Not Churn customer. AIC (Akaike Information Criterion) value of this model is 13,271, then we create another model to achieve a lower AIC value without including such not-significant attributes.

As a result of the above model (with excluding not significant attribute) we achieved with a lower AIC value (12,612) and a better model.

4.5.1.4 Predict on Test Dataset

To predict the observation using Logistic Regression we need to set type=" response" in order to compute the prediction probabilities.

```
> pred <- predict(logitmod, newdata = testData, type = "response")
```

The prediction “pred” contains the probability that the observation is Not Churn or churn for each observation. The common practice is to take the probability cutoff as 0.5. If the probability of Y is > 0.5, then it can be classified an event (Not Churn). So, if “pred” is greater than 0.5, it is Not Churn else it is churn. The following R code shows the prediction of the observation.

```
> pred <- predict(logitmod, newdata = testData, type = "response")
```

```
> y_pred_num <- ifelse(pred > 0.5, 1, 0)
> y_pred <- factor(y_pred_num, levels=c(0, 1))
> y_act <- testData$AA
> mean(y_pred == y_act)
```

4.5.1.5 Evaluating Model Performance

Then after prediction we compute the accuracy using proportion of y_pred that matches with y_act.

```
> mean(y_pred == y_act)
[1] 0.9242911
```

So, the total accuracy rate of the model using proportion of y_pred with y_act is 92.4% as shows in the above R code result of the accuracy.

```
> CrossTable(x = y_pred,y= y_act,prob.chisq=FALSE)

   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  7476

             | y_act
      y_pred |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |      3327 |       393 |      3720 |
             |  1443.281 |  1270.494 |           |
             |     0.894 |     0.106 |     0.498 |
             |     0.951 |     0.099 |           |
             |     0.445 |     0.053 |           |
-------------|-----------|-----------|-----------|
           1 |       173 |      3583 |      3756 |
             |  1429.447 |  1258.316 |           |
             |     0.046 |     0.954 |     0.502 |
             |     0.049 |     0.901 |           |
             |     0.023 |     0.479 |           |
-------------|-----------|-----------|-----------|
Column Total |      3500 |      3976 |      7476 |
             |     0.468 |     0.532 |           |
-------------|-----------|-----------|-----------|
```

Figure 4.3: the cross table result first snap shot

To create the confusion matrix table and calculate the accuracy of the model, the package “e1071” is installed and run the confusion matrix function like below code in R Studio.

install.packages('e1071', dependencies=TRUE)

```
> confusionMatrix(y_pred, y_act)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3327  393
         1  173 3583

               Accuracy : 0.9243
                 95% CI : (0.9181, 0.9302)
    No Information Rate : 0.5318
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.8485

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.9506
            Specificity : 0.9012
         Pos Pred Value : 0.8944
         Neg Pred Value : 0.9539
             Prevalence : 0.4682
         Detection Rate : 0.4450
   Detection Prevalence : 0.4976
      Balanced Accuracy : 0.9259

       'Positive' Class : 0
```

Figure. 4.4: confusion matrix results first snap shot

Precision of the model evaluates through the following equation. Precision =TP/TP+FP

Precision = 3583/3583+393= 0.901

To plot the results of sensitivity and specificity of the model and to check what threshold achieves (calculate the AUC) we used rock curve as a tool using “ROCR” function. The ROC is a curve generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings while the AUC (area under the curve) is the area under the ROC (Receiver Operator Characteristic) curve. A model with good predictive ability should have an AUC closer to 1 (1 is ideal) than to 0.5[52].

```
> plot(y_act,y_pred)
> library(ROCR)
```

```
> ROCRpred = prediction(pred, y_act)
> ROCRperf = performance(ROCRpred, "tpr", "fpr")
> plot(ROCRperf, colorize = TRUE)
```

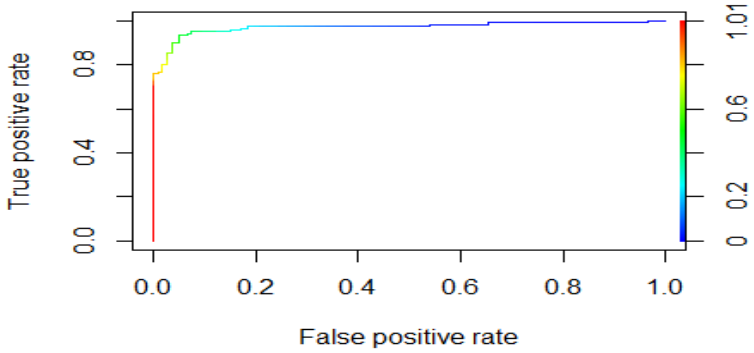


Figure. 4.5: confusion matrix results snap shot plot

```
> library(InformationValue)
> plotROC(y_act, pred)
```

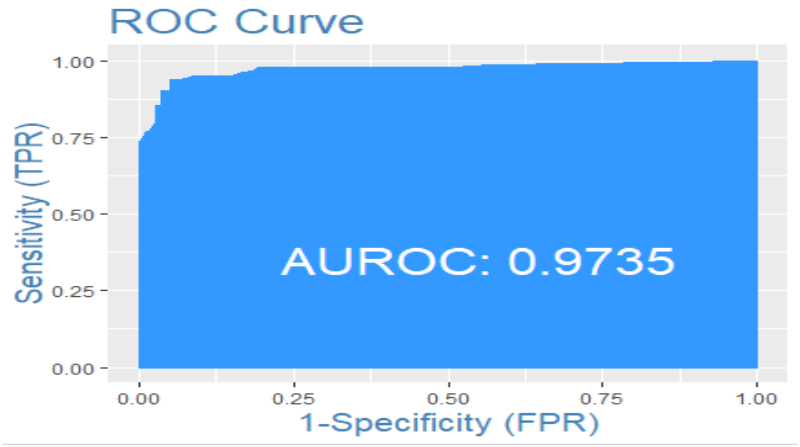


Figure. 4.6: ROC Curve result snap shot

ROC determines the accuracy of a classification model at a user defined threshold value. It determines the model's accuracy using Area under Curve (AUC). The area under the curve (AUC), also referred to as index of accuracy (A) or concordant index, represents the performance of the ROC curve. In a good model, the curve should rise steeply, indicating that the TPR (Y-Axis) increases faster than the FPR (X-Axis) as the cutoff score decreases. Greater the area under the ROC curve, better the predictive ability of the model. (Higher the area, better the model) [52]. as showed in the above diagram ROC is plotted between True Positive Rate (Y axis) and False Positive Rate (X Axis). In the above plot, the area under curve cover the maximum area (ROC curve 97.35%) and the curve is higher and close to the left corner (true positive) that indicate the model is pretty good.

4.5.1.6 Experiment two

The experiment two is conducted by using unbalanced dataset unlike to the experiment one.

4.5.1.6.1 Building the Logistic Regression Model with Imbalance Dataset

```
> pred <- predict(logitmod, newdata = testData, type = "response")

> y_pred_num <- ifelse(pred > 0.5, 1, 0)
> y_pred <- factor(y_pred_num, levels=c(0, 1))
> y_act <- testData$AA
> mean(y_pred == y_act)
[1] 0.9388711
```

Precision of the model with imbalanced dataset evaluates through the following equation. Precision =TP/TP+FP

Precision = 3692/3692+284= 0.928

As shown in above prediction model result, the accuracy of the model is 93%. It shows that the unbalanced dataset in this Logistic Regression model predict more accurately by 1% than the balanced (up balance) dataset in experiment one.

4.6 Modeling using Naive Bayes

In this experiment a Prediction model building is done using Naive Bayes Classification algorithm and RStudio used to predict the customers with R programing.

4.6.1 Data description

For this Navies Bayes model also, we use samples of about 48,051 CBE customers' details with 14 independent and one dependent attributes. The dataset and its structure are as follows:

Then we do Convert the customer status field values churn and Not Churn to '1'and '0' respectively and also Convert the customer status field as factor as shown below: -

```
> NBMFIL$AA <- ifelse(NBMFIL$AA == 'Active',1,0)
> NBMFIL$AA <- factor(NBMFIL$AA,levels = c(0,1))
```

4.6.2 Partitioning the Dataset

In predictive modeling, the data needs to be partitioned into train and test sets. 80% of the data is partitioned for training purpose and 20% of the data for testing purpose as we do in above other models. In this section after data splitting, we apply Feature scaling to standardize the range of independent variables.

4.6.3 Classification Using Naive Bayes

In this section Naïve Bayes classification model is executed in R Studio on top of the CBEs dataset to classify Not Churn and churn borrowers. To do Naive Bayes classification model, we perform the following: - first we Install and load "e1071" package before running Naive Bayes; Test the models built using train datasets through the test dataset and then Using accuracy, precision and error rate, we analyses how these models are behaving for the test dataset.

```
> classifier = naiveBayes(x= traning_set,y = traning_set$AA)
> y_pred = predict(classifier,newdata = test_set)
```

The above code used to classify the dataset using Naïve Bayes and predict it using test dataset. The result of the model also evaluates through accuracy and precision as we do in previous two models.

4.6.4 Visualize Test Set Result

We visualize the test set result using "crossTable" function in RStudio and summary of it using table.

```
> CrossTable(x= y_pred,y=y_act)

   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  7476

             | y_act
      y_pred |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |      3500 |      1204 |      4704 |
             |   764.747 |   673.193 |           |
             |     0.744 |     0.256 |     0.629 |
             |     1.000 |     0.303 |           |
             |     0.468 |     0.161 |           |
-------------|-----------|-----------|-----------|
           1 |         0 |      2772 |      2772 |
             |  1297.753 |  1142.388 |           |
             |     0.000 |     1.000 |     0.371 |
             |     0.000 |     0.697 |           |
             |     0.000 |     0.371 |           |
-------------|-----------|-----------|-----------|
Column Total |      3500 |      3976 |      7476 |
             |     0.468 |     0.532 |           |
-------------|-----------|-----------|-----------|
```

Figure. 4.7: The cross table result second snap shot

```
> tb <- table(y_act,y_pred)
> View(tb)
```

	y_act	y_pred	Freq
1	0	0	3500
2	1	0	1204
3	0	1	0
4	1	1	2772

Figure. 4.8: The cross table result third snap shot

4.6.5 Evaluating Model Performance

The above Naïve Bayes model result evaluated using accuracy and precision techniques as we do in previous two models before. The test data consisted of 7476 observations. Out of which 3500 cases have been accurately predicted (TN->True Negative) as (Not Churn) churn in nature which constitutes 46.8%. Also, 1204 out of 7476 observations were predicted (FP-> False Positive) as churn in nature but got predicted as Not Churn which constitutes 16.1%.

Also,2772 out of 7476 observations were correctly predicted (TP -> True Positive) as Not Churn in nature which constitutes 37.1% and there is no (False Positive ->FP) prediction that mean churn in nature but got predicted as Not Churn. The total Accuracy of the model is 83.9%.

As a result Accuracy = 6272 /7476 =83.89% and Precision = 2772/2772+0= 1 = 100%.

Accuracy of the model using confusion Matrix table function in RStudio.

```
> confusionMatrix(table(y_act,y_pred))
Confusion Matrix and Statistics

     y_pred
y_act    0    1
    0 3500    0
    1 1204 2772

               Accuracy : 0.839
                 95% CI : (0.8304, 0.8472)
    No Information Rate : 0.6292
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6831

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7440
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.6972
             Prevalence : 0.6292
         Detection Rate : 0.4682
   Detection Prevalence : 0.4682
      Balanced Accuracy : 0.8720

       'Positive' Class : 0
```

Figure .4.9: confusion matrix results second snap shot

4.7. Modeling using SVM

SVM (Support Vector Machine) is a supervised machine learning algorithm which is mainly used to classify data into different classes. Unlike most algorithms, SVM makes use of a hyper plane which acts like a decision boundary between the various classes. These closest data points to the hyper plane are known as support vectors [52]. As we do in other previous three models, we use R programming language to build an SVM classifier model.

4.7.1 Install Caret Packages in to RStudio

The caret package is also known as the Classification and Regression Training, has tons of functions that helps to build predictive models. It contains tools for data splitting, pre-processing, feature selection, tuning, and unsupervised learning algorithms [51].

> install.packages("caret")

4.7.2 Data Description and Load the Dataset

The dataset we use in this model have about 48,051 borrowers record details with 14 independent and one dependent attributes, it stored in a CSV or Comma Separated Version.

The output shows that the dataset consists of 48,051 observations each with 15 attributes.

The below code is used to convert the data frame's "AJ" column to a factor variable.

4.7.3 Split the Data in to Training and Test Set

In this section the dataset split into training set and testing set with 80:20 ratio, this is also called data splitting. The training set specifically used for the model building and the testing set for evaluating the model.

The caret package in R provides a method createDataPartition() which is basically for partitioning our data into train and test set.

```
> intrain <- createDataPartition(y = SVMMFIL$AA, p= 0.8, list = FALSE)
> training <- SVMMFIL[intrain,]
> testing <- SVMMFIL[-intrain,]
```

In the above R code, the "y" parameter takes the value of variable according to which data needs to be partitioned. In our case, target variable is at AA, so we are passing SVMCBEL$AA. The "p" parameter holds a decimal value in the range of 0-1. It's to show the percentage of the split. We are using p=0.8. It means that data split should be done in 80:20 ratios. So, 80% of the data is used for training and the remaining 20% is for testing the model. The createDataPartition() method is returning a matrix "intrain". This "intrain" matrix has training dataset and we're storing this in the 'training' variable and the rest of the data (20%) of it stored in the testing variable. Then the dimensions of training data frame and testing data frame is looks like the following.

4.7.4 Training Model

To train the model, first we need to implement the trainControl() method provided by the Caret package. The training method is used to train the data on specific algorithms.

```
> trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
```

In the above code we used the “number” parameter to holds the number of resampling iterations. And the “repeats” parameter contains the sets to compute for the repeated cross-validation. We are using setting number =10 and repeats =3.

```
> svm_Linear <- train(AA ~., data = training, method = "svmLinear",trControl=trctrl,preProcess = c("center",
"scale"),tuneLength = 10)
```

In the above train model code, the “AA~.” denotes a formula for using all attributes in the classifier and AA as the target variable. And the “preprocess” parameter is set for preprocessing the training data with passing two values parameter “center” and “scale.

We are passing 2 values in our “pre-process” parameter “center” & “scale”. These two parameters help for centering and scaling the data. After pre-processing, these convert the training data with mean value as approximately “0” and standard deviation as “1”. The “tuneLength” parameter holds an integer value. This is for tuning the algorithm. The result of the train () method is save in the svm_Linear variable.

The model tested at value “C” =1. The model is trained with C value as 1. Then we predict classes for the test set using predict () method. The caret package provides predict () method for predicting results. We are passing 2 arguments. Its first parameter is the trained model and second parameter “newdata” holds the testing data frame. The predict() method returns a list, then we save it in a test_pred variable.

```
> test_pred <- predict(svm_Linear, newdata = testing)
```

4.7.5 Model Performance Evaluation

To check the performance of the model through accuracy and precision we use the confusion matrix. The result of confusion matrix is shown in below code.

```
> confusionMatrix(table(test_pred, testing$AA))
Confusion Matrix and Statistics


test_pred    0    1
         0 3246  314
         1  254 3662

               Accuracy : 0.924
                 95% CI : (0.9178, 0.9299)
    No Information Rate : 0.5318
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8476

 Mcnemar's Test P-Value : 0.0133

            Sensitivity : 0.9274
            Specificity : 0.9210
         Pos Pred Value : 0.9118
         Neg Pred Value : 0.9351
             Prevalence : 0.4682
         Detection Rate : 0.4342
   Detection Prevalence : 0.4762
      Balanced Accuracy : 0.9242

       'Positive' Class : 0
```

Figure. 4.10: Confusion matrix snap shot

The output of confusion matrix shows that the model accuracy for test set is 92.4%.

Precision of the model also evaluates through the following equation. Precision =TP/TP+FP

Precision = 3662/3662+314= 0.921

The SVM model variable importance in percentages, according to this, the top four massively importance attributes are AO (Grand total number of transaction), AH (Industries), AH (and AM (ATM user).

4.8. Comparison of Machine Learning Models

The below (table 4.1) shows a summary of the result obtained from all the four models when all models are trained on around 29400 instances. It has been found that KNN resulted into highest accuracy. But at the prediction level Logistic regression, SVM and Naive Bayes model performed well in terms of accuracy and precision. These machine learning models evaluations (comparison) are done through the model accuracy and precision. The four Machine learning models (SVM, logistic regression, Naive Bayes, and KNN) were used to predict the Ethiopian CBE data on the 18 selected attributes.

Table 4: 1 model comparison

	KNN		SVM	Logistic regression		Naive Bayes
	KNN (Experiment one)	KNN (Experiment two)		with balanced data	With imbalance dataset	
Total number of instances (test dataset)	7980	7980	7980	7476	7476	7476
Correctly classified instance	7973	7970	7973	6910	566	6272
Incorrectly classified instance	7	10	7	7019	557	173
Accuracy	99.91	99.87	92.4	92.89	93.8	83.89
Precision	1	1	0.921	0.9	0.928	1

As stated in above table machine learning algorithm K-NN recorded an extremely higher accuracy (99.9%) than the rest three models but Logistic Regression(93.8%) and SVM (92.4%) show a comparable performance. The result of the study is presents to the domain experts those are senior CBE managers and experts and they proved technical evaluation to assurance of the results of the study and quality of the data.

4.9 Discussion of result

The following section presents the summary of the main findings of this work. For easy readability it is presented in table form. As stated earlier (in chapter One) the main objective of building the prediction model is to develop and construct for each customer churn status (churn, Not Churn) that would help in predicting the likely status of a new customer in terms of these attributes. The experiment for the study is extensively tested on the dataset. The dataset in which testing is performed are those which are described in the data preprocessing section. As it is described above, the CBE dataset contain 48,051 out of it 20% of the dataset is taken as test sample by keeping the remaining 80% of it as training.

The prediction is checked on KNN, SVM, Naïve Bayes and Logistic Regression algorithms separately, and the result is presented & discussed as follows. The summary of the result also includes confusion matrix, which is a table layout to visualize the performance of a model. A typical confusion matrix consists of rows and columns where each column represents the number of instances in the predicted class and each row represents the number of instances in an actual class. In predictive analytics, a confusion matrix represents the total number of true positives, false positives, false negatives and true negatives.

4.9.1 Results on KNN (Nearest neighbor algorithm)

A separate check conducted in KNN model by changing the value of k to 199,200,201 and 185-205 is done during experimentation, and the percentage of correctly identified customer status among the total is taken as result of the study. Performance of the study is checked through accuracy and precision using confusion matrix. The result obtained from the experiment is stated in the following manner. The following table gives the summary of the four-experiment conducted in KNN model.

Table 4.2 Summary of KNN experiments result

no	Experiment label	Attribute used	K value	Accuracy level	Precision
1	Experiment one	15	199	99.91	1
2	Experiment two	15	200	99.87	1
3	Experiment three	15	201	99.87	1
4	Experiment four	15	185-205	Average of (99.89)	1

These accuracy values are very good as stated in [30] [31] estimates high accuracy to be in the range of 0.75 and 1 representing 75% and 100% respectively.

4.9.1.1 Confusion Matrix of KNN Classifier for customer status Prediction

Table 4.3 Confusion matrix result of KNN experiment one

KNN (Experiment one with value of K 193)		Predicted class	
		Not Churn	churn
Actual class	Not Churn	7418	0
	Churn	7	555

Table 4.4 Confusion matrix of KNN experiment two

KNN (Experiment two with value of K 192)		Predicted class	
		Not Churn	churn
Actual class	Not Churn	7418	0
	Not Churn	10	552

4.9.1.2 Attribute Contribution for KNN Prediction Model

According to the KNN prediction model, only eleven attributes are listed as important attributes out of 14 independent attributes, out of these eleven important attributes, two attributes, A15(Grand total number of transaction) and A8(industries), have contributed massively in the prediction with the result showing below in descending order. As a result of this, the results analysis and discussion on this section would be made around the two most contributing attributes.

4.9.2 Results on SVM (Support Vector Machine)

After the development of the SVM classification model, a number of evaluation techniques were implemented. Confusion matrix is used for evaluate model accuracy and crossable used to visualize the result, and also precision of the model calculated using equation. The SVM classification model was developed to identify prospective Not Churn and Churn borrowers in Ethiopia CBEs, it yielded with accuracy values of 92.4%. These accuracy values are very good as [49] [50] estimates high accuracy to be in the range of 0.75 and 1 representing 75% and 100% respectively.

4.9.2.1 Confusion Matrix of SVM Classifier for customer status Prediction

Table 4.5 Confusion matrix of SVM experiment

SVM		Predicted class	
		Not Churn	churn
Actual class	Not Churn	3662	254
	Churn	314	3246

4.9.3 Results on Logistic Regression Model

Like the experiment held on using nearest neighbor algorithm and SVM, conducting the experiment through CBEs dataset with 48,051 observations and 15 attributes. The model performance evaluates through accuracy and precision, sensitivity and specificity. Confusion matrix is used for evaluate model accuracy and crossable used to visualize the result. And also, precision of the model calculated using equation. The Logistic Regression model was developed to identify prospective Not Churn and churn borrowers in CBE like in other models. In this study also the imbalance data and balanced data is tested. It yielded with accuracy values of 92.4% and 93.8%. These accuracy values are very good as [49] [50] estimates high accuracy to be in the range of 0.75 and 1 representing 75% and 100% respectively.

4.9.3.1 Confusion Matrix of Logistic Regression Classifier for customer Status Prediction

Table 4.6 Confusion matrix of Logistic Regression experiment one

Logistic Regression(with balanced data)		Predicted class	
		Not Churn	churn
Actual class	Not Churn	3583	173
	Churn	393	3327

Table 4.7 Confusion matrix of Logistic Regression experiment two

Logistic Regression(with imbalance data)		Predicted class	
		Not Churn	churn
Actual class	Not Churn	3692	173
	Churn	284	3327

4.9.3.2 Attribute contribution for Logistic Regression prediction model

AO(Grand total number of transaction), AN(POS user), AM(ATM user), AL(Internet banking user),AK(MOBILE_BANKING_USER) These five attributes have the lowest p-value that indicate a strong association of those attributes to the customer with the probability of having Not Churn customer.

4.9.4 Results on Naive Bayes Model

For the Naive Bayes prediction model, we use samples of about 40,051 customer borrowers' details with 14 independent and one dependent attributes. After the development of the model, evaluation techniques were implemented. Confusion matrix is used for evaluate model accuracy and crossable used to visualize the result, and also precision of the model calculated using equation. The model accuracy is 83.9%, these accuracy values are very good as estimates high accuracy to be in the range of 0.75 and 1 representing 75% and 100% respectively.

4.9.4.1 Confusion Matrix of Naive Bayes Classifier for customer status Prediction

Table 4.8 confusion matrix of Naive Bayes experiment

Naive Bayes		Predicted class	
		Not Churn	churn
Actual class	Not Churn	2772	0
	Churn	1204	3500

To summarize this chapter, the processes involved in building the four prediction model SVM, KNN, Naïve Bayes and Logistic Regression Machine Learning algorithms Techniques; as well as the performance evaluation procedures were discussed meanwhile the cross table was presented to visualize the model result and the confusion matrix was presented for the accuracy, also precision was calculated.

In relation to research question one "Which attributes are more significant to predict customers churning at CBE?" AO (Grand total number of transaction), AN (POS user), AM (ATM user), AL (Internet banking user), AK (MOBILE_BANKING_USER) and AH (industries) are found to be effective for predicting customers churn in CBE.

In addition, in relation to research question two, "Which classification algorithm is more suitable for constructing CBE customer churn prediction model?" KNN was emerged the most suitable for customer churn prediction model Accuracy 99.91% and precision 100%.

In relation to research question three "To what extent the proposed model performs in customer churn prediction?" KNN by 99.91%, Logistic Regression by 93.8%, SVM by 92.4 %, and Naïve Bayes by 83.89% were modeled to find out how machine learning can be utilized in making decision whether to extend customer or not.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDETIONS

5.1 Overview

This chapter summarizes the entire findings of the study. It is divided into two main sections conclusion, recommendation it provides the conclusion and recommendations the future works.

5.2 Conclusions

The necessary data for the experiment was obtained from the commercial bank of Ethiopia on a scattered excel sheets which totally amounted to a dataset of 48,051. The necessary preprocessing activities were applied on the dataset after which 48,051 data was prepared for the experimentation. Classification and prediction model building was experimented with KNN, SVM, and Naïve Bayes and Logistic Regression algorithm. And the tools were used to simulate all the experiment is R Studio using R programing. The confusion matrix was used and then accuracy, and sensitivity were calculated. The three models (KNN, SVM and Logistic Regression) yielded remarkable results when it comes to correctly classifying instances, KNN with accuracy of 99.91%, 99.87%, 99.874% in the first, second and third experiments respectively, SVM with accuracy of 92.4%, and Logistic Regression with accuracy of 92.8 % and 93.8% on balanced and imbalanced dataset respectively.

Based on the above conclusion all the researcher questions are addressed as follows.

- Question one “Which attributes are more significant to predict customers churning at CBE?” AO (Grand total number of transaction), AN (POS user), AM (ATM user), AL (Internet banking user), AK (MOBILE_BANKING_USER) and AH (industries) are found to be effective and relevant (important) predictors for the target class customer status (Not Churn and churn) for predicting customers churn in CBE.
- In addition, in relation to research question two, “Which classification algorithm is more suitable for constructing CBE customer churn prediction model?” KNN was emerged the most suitable for customer churn prediction model Accuracy 99.91%

and precision 100%.

- In relation to research question three “To what extent the proposed model performs in customer churn prediction?” KNN by 99.91%, Logistic Regression by 93.8%, SVM by 92.4 %, and Naïve Bayes by 83.89% were modeled to find out how machine learning can be utilized in making decision whether to extend customer or not.

From the results of the experiments it can be concluded that the Machine Learning algorithms Techniques can be effectively applied on the banks in order to generate customer churn predictive models with an acceptable level of accuracy.

Therefore, banks in Ethiopia can apply these Machine Learning models to detect possible customer churn probabilities in advance. By adopting this way, it is expected that CBE would operate very well to identify and predict their customer churn and will be able to make the expected investment returns.

However, due to the quality and size of dataset used, in addition to the Machine Learning tools and techniques are essential factors for the modeling performance, an increased size in the dataset with increase in amount and diversity of attributes involving even all banks in Ethiopia, could have resulted in an improved modeling. It could have enabled the research to make use of larger data with more attributes than those used in this study to help address other areas of problem in the banking industries of the country. The model building experimentation conducted is based on collected CBE datasets with just 48,051 dataset but it is the researcher’s belief that it would have resulted in improved model if it includes other banks dataset in Ethiopia with big data size and all type of products customers, and if other techniques were also utilized. Hence following section presents some recommendations drawn based on the result of the research.

5.3 Recommendations

Even though the investigation undertaken is mainly for academic purpose, it will have important contribution for commercial banks and for other researchers interested in similar area. Although the results of this study are inspiring, there are problem areas that need further investigation for future work to attain better

inclusive model and also bring it to an operational level. Therefore, the researcher forwards the following issues as a future research direction based on this study: -

- Based on the proposed optimal model in this study, we recommend future research to integrate customer churn predictive model with CRM data base management system.
- We, future researches, can incorporate even all the 33.3 million dataset owned by CBE to build more useful customer churn predictive model and elevate the overall Costumer Churn rate.
- Machine Learning algorithms can be used to limit or possibly completely extinguish to the prediction of customer churns in CBE customers. But there are generally other problems that exist in every commercial banks. Therefore, interested researchers can look into different natured problems in commercial banks including CBE from Machine Learning perspective.
- Different Classification Algorithms such as Neural Network, Decision Tree and Ensemble Learning can be employed for banks in Ethiopia with increased dataset Sample size and different attributes that include all type of products and customers by different researchers to see if it could result in different findings.
- All Banks applied the same rule and regulation of NBE (National Bank of Ethiopia) but internally they implement different Marketing strategies and customer handling policy because of this the model need minor adjustment; therefor this customer churn predictive model is specific to CBE, we recommend future researches to Private Banks.

References

[1] Gudela Grotea and José M. Cortinab, Necessity (not just novelty) is the mother of invention: using creativity research to improve research in work and organizational psychology, 2018.

[2] Oskar. Stucki, predicting the customer churn with machine learning methods-case: Private insurance customer data, 2019.

[3] Commercial Bank of Ethiopia (CBE) web site" https://combanketh.et/" December 2021.

[4] Arwa A. Jamjoom The use of knowledge extraction in predicting customer churn in B2B, 2021.

[5] Kazi Imran Moin ," Use of Data Mining in Banking" International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 2,Mar-Apr 2012.

[6] Hajar Ghaneei, Seyed Mohammad" Developing a prediction model for customer churn from electronic banking services using data mining" 22 August 2016.

[7] Anjali Ganesh Jivani," The Novel k nearest neighbor algorithm" 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Jan. 04 – 06, 2013, 2013 IEEE.

[8] Kassahun Gebremeskel, Application of Data mining Techniques to Predict Customers 'churn at Commercial Bank of Ethiopia,2013.

[9] Belachew Reganie, Application of Data Mining Techniques for Customers Segmentation and Prediction: The Case of Buusaa Gonofa Microfinance Institution, 2013.

[10] Sylvia Tippmann, Programming tools: Adventures with R, December 2014.

[11] Ken Kelley, Keke Lai, and Po-Ju Wu, Using R for Data Analysis a Best Practice for Research, 2008.

[12] Bhambri, V. (2012). Data Mining as a Tool to Predict Churn Behavior of Customers. International Journal of Computer & Organization Trends, 2(3), 85–89.

[13] Chen K, Hu Y-H, Hsieh Y-C Predicting customer churn from valuable B2B

customers in the logistics industry: a case study. IseB 13:475–494. Doi: 10.1007/s10257-014-0264-1, (2014).

[14] Kotler, P., Keller, K. L. Marketing Management. Pearson Prentice Hall, 2009.

[15] Karaelmas Fenve, Review of Customer Churn Analysis Studies in Telecommunications Industry Karaelmas, 2017.

[16] Seker, S.E, Müşteri Kayıp Analizi (Customer Churn Analysis). YBS Ansiklopedi, 2016.

[17] Kumar, D., & Bhardwaj, D. (2011). Rise of Data Mining: Current and Future Application Areas. International Journal of Computer Science Issues, 8(5), 256–260.

[18] ZenTut. (2013). Data Mining Techniques. Data mining. Retrieved April 09, 2013, from http://www.zentut.com/data-mining-

[19] Iqbal, N. & Islam, M. (2016).From Big Data to Big Hope: An outlook on recent trends and challenges. Journal of Applied Computing, 1(1):1-12.

[20] Mark Stamp," A Survey of Machine Learning Algorithms and Their Application in Information Security: An Artificial Intelligence Approach," San Jose State University, San Jose, California, September 2018.

[21] M.A.R. Khalid, M.A.H. Farquad," Comparative Analysis of Support Vector Machine: Employing Various Optimization Algorithms" 14th International Conference on Information Technology, 2015 IEEE

[22] Jorma Laaksonen, Erkki Oja," Classification with Learning k-Nearest Neighbors", Helsinki University of Technology Laboratory of Computer and Information Science, ©1996 IEEE.

[23] Jiangtao Ren, Sau Dan Lee, Xianlu Chen," Naive Bayes Classification of Uncertain Data" 2009 Ninth IEEE International Conference on Data Mining.

[24] Yuguang Huang, Lei Li," Naïve Bayes Classification Algorithm Based on Small Sample Set" Beijing University of Posts and Telecommunications, Proceedings of IEEE CCIS2011.

[25] Rahul Saxena," Introduction to k Nearest Neighbor Classification and Condensed Nearest Neighbour Data Reduction" Data Science, Machine

Learning journal monthly blog, December 23, 2016.

[26] Ashlesha Vaidya," Predictive and Probabilistic Approach Using Logistic Regression: Application To Prediction of Loan Approval" 8th ICCCNT 2017 July 3-5, 2017, IIT Delhi. IEEE – 40222.

[27] A. O. Oyeniyi & A.B. Adeyemo (2015): Customer Churn Analysis in Banking Sector Using Data Mining Techniques. Afr J. of Comp & ICTs. Vol 8, No. 3. Pp 165-174.

[28] Anil Kumar D, Ravi V (2008) Predicting credit card customer churn in banks using data mining. Int J Data Anal Tech Strateg 1(1):4–28

[29] Keramati, Abbas; Ghaneei, Hajar; Mirmohammadi, Seyed Mohammad, Developing a prediction model for customer churn from electronic banking services using data mining, 2016.

[30] Chantine Huigevoort, Customer churns prediction for an insurance company using Data mining technique: 2015.

[31] Sunita ,Yashaswi , Dr. Smita Chavan2; National Monthly Refereed Journal of Research in Commerce & Management predictive model Insurance Industry using R tool:2013.

[32] Shiv. Maharaj: Machine Learning Model Evaluation available on www.analyticsvidhya.com/blog/2021/05/machine-learning-model-evaluation/2021

[33] Michael D. Clemes, Christopher Gan and Min Ren Synthesizing the Effects of Service Quality, Value, and Customer Satisfaction on Behavioral Intentions in the Motel Industry: An Empirical Analysis, [2010]

[34] Maritza Mera-Gaona, Framework for the Ensemble of Feature Selection Methods, 2021.

[35] Tianpei Xu, Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping, 2021.

[36] S. B. Kotsiantis, Feature selection for machine learning classification problems: a recent overview 2011

[37] Available on https://medium.com/analytics-vidhya/three-steps-in-case-of-

imbalanced-data-and-close-look-at-the-splitter-classes-8b73628a25e6

[38] Raschka, Sebastian," Predictive modeling, supervised machine learning, and pattern classification" 2014 IEEE.

[39] R. David A. Dickey," Introduction to Predictive Modeling with Examples," SAS Global Forum, 2012.S

[40] Ram Babu, Rama Satish," Improved of K-Nearest Neighbor Techniques in Credit Scoring," International Journal for Development of Computer Science & Technology, Volumn-1, Issue-2, Feb-March-2013.

[41] Abbas Keramati, Amin Omidvar," Default Probability Prediction of Credit Applicants Using a New Fuzzy KNN Method with Optimal Weights," IGI Global, ch024, 2015.

[42] Mark Stamp," A Survey of Machine Learning Algorithms and Their Application in Information Security: An Artificial Intelligence Approach," San Jose State University, San Jose, California, September 2018.

[43] Jesper De Groot, "Credit risk modeling using a weighted support vector machine" Utrecht University, Master Thesis, September 23, 2016.

[44] Tony Van Gestel, Bart Baesens, Dr. Ir. Joao Garcia, Peter Van Dijcke," A Support Vector Machine Approach to Credit Scoring, "impact research, Credit Methodology Global Market Risk, Dexia Group,2015.

[45] Bolarinwa Akindaini, Martti Juhola," Machine Learning Applications in Mortgage Defaults Predictions" University of Tampere, November 2017.

[46] Ansen Mathew," Credit Scoring Using Logistic Regression" Master's Theses, San Jose State University Spring May, 2017.

[47] Olatunji J. Okesola, Kennedy O. Okokpujie, Adeyinka A. Adewale, Samuel N. John, Osemwegie Omoruyi," An improved Bank Credit Scoring Model a Naïve Bayesian Approach," International Conference on Computational Science and Computational Intelligence,2017.

[48] AidaKrichene," Using a naive Bayesian classifier methodology for loan risk assessment," Journal of Economics,Finance and Administrative Science Vol.22No.42,2017 pp. 3-24.

[49] Nicholas Gray and Scott Ferson, Logistic Regression through the Veil of Imprecise Data, 2021.

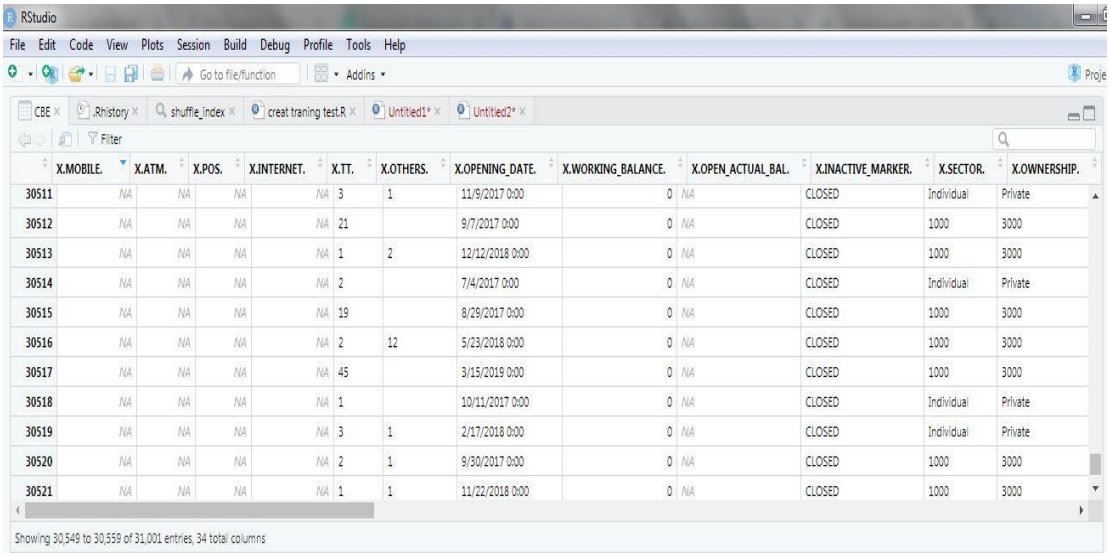
[50] Saradnici “data normalization” Microsoft document, Microsoft azure, docs.micrsoft.com 1st Jan, 2019. Available: https://docs.microsoft.com. [Accessed: Nov, 2021].

[51] Eréndira Rendón, Data Sampling Methods to Deal with the Big Data Multi-Class Imbalance Problem, 2020.

[52] Mohammad Hassan Almaspoor, Support Vector Machines in Big Data Classification: A Systematic Literature Review, 2021.

Annexes

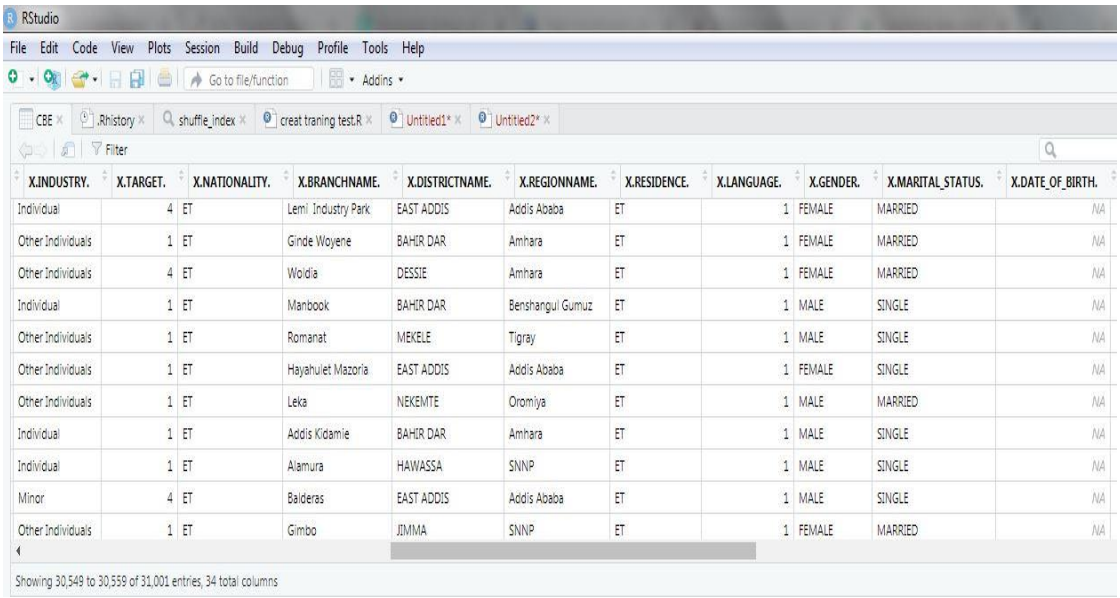
A-1: CBE Dataset Prepared 34 attribute for the Research



	X.MOBILE.	X.ATM.	X.POS.	X.INTERNET.	X.TT.	X.OTHERS.	X.OPENING_DATE.	X.WORKING_BALANCE.	X.OPEN_ACTUAL_BAL.	X.INACTIVE_MARKER.	X.SECTOR.	X.OWNERSHIP.
30511	NA	NA	NA	NA	3	1	11/9/2017 0:00	0	NA	CLOSED	Individual	Private
30512	NA	NA	NA	NA	21		9/7/2017 0:00	0	NA	CLOSED	1000	3000
30513	NA	NA	NA	NA	1	2	12/12/2018 0:00	0	NA	CLOSED	1000	3000
30514	NA	NA	NA	NA	2		7/4/2017 0:00	0	NA	CLOSED	Individual	Private
30515	NA	NA	NA	NA	19		8/29/2017 0:00	0	NA	CLOSED	1000	3000
30516	NA	NA	NA	NA	2	12	5/23/2018 0:00	0	NA	CLOSED	1000	3000
30517	NA	NA	NA	NA	45		3/15/2019 0:00	0	NA	CLOSED	1000	3000
30518	NA	NA	NA	NA	1		10/11/2017 0:00	0	NA	CLOSED	Individual	Private
30519	NA	NA	NA	NA	3	1	2/17/2018 0:00	0	NA	CLOSED	Individual	Private
30520	NA	NA	NA	NA	2	1	9/30/2017 0:00	0	NA	CLOSED	1000	3000
30521	NA	NA	NA	NA	1	1	11/22/2018 0:00	0	NA	CLOSED	1000	3000

Showing 30,549 to 30,559 of 31,001 entries, 34 total columns

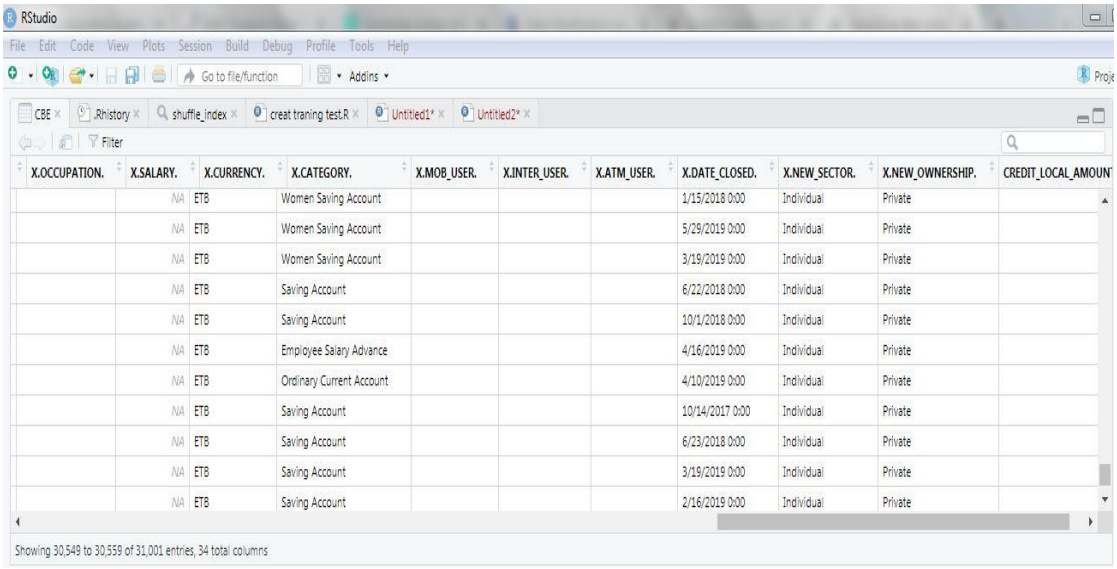
A-2: CBE Dataset Prepared 34 attribute for the Research cont.



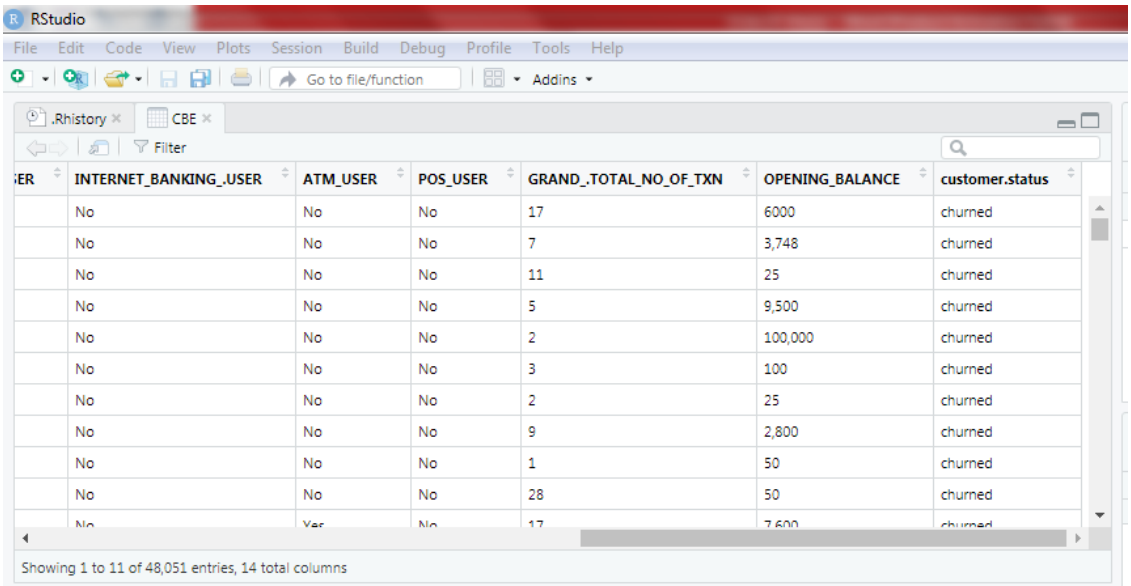
	X.INDUSTRY.	X.TARGET.	X.NATIONALITY.	X.BRANCHNAME.	X.DISTRICTNAME.	X.REGIONNAME.	X.RESIDENCE.	X.LANGUAGE.	X.GENDER.	X.MARITAL_STATUS.	X.DATE_OF_BIRTH.
Individual	4	ET	Lemi Industry Park	EAST ADDIS	Addis Ababa	ET	1	FEMALE	MARRIED	NA
Other Individuals	1	ET	Ginde Woyene	BAHIR DAR	Amhara	ET	1	FEMALE	MARRIED	NA
Other Individuals	4	ET	Woldia	DESSIE	Amhara	ET	1	FEMALE	MARRIED	NA
Individual	1	ET	Manbook	BAHIR DAR	Benshangul Gumuz	ET	1	MALE	SINGLE	NA
Other Individuals	1	ET	Romanat	MEKELE	Tigray	ET	1	MALE	SINGLE	NA
Other Individuals	1	ET	Hayahulet Mazoria	EAST ADDIS	Addis Ababa	ET	1	FEMALE	SINGLE	NA
Other Individuals	1	ET	Leka	NEKEMTE	Oromiya	ET	1	MALE	MARRIED	NA
Individual	1	ET	Addis Kidamie	BAHIR DAR	Amhara	ET	1	MALE	SINGLE	NA
Individual	1	ET	Alamura	HAWASSA	SNNP	ET	1	MALE	SINGLE	NA
Minor	4	ET	Balderas	EAST ADDIS	Addis Ababa	ET	1	MALE	SINGLE	NA
Other Individuals	1	ET	Gimbo	JIMMA	SNNP	ET	1	FEMALE	MARRIED	NA

Showing 30,549 to 30,559 of 31,001 entries, 34 total columns

A-3: CBE Dataset Prepared 34 attribute for the Research cont.



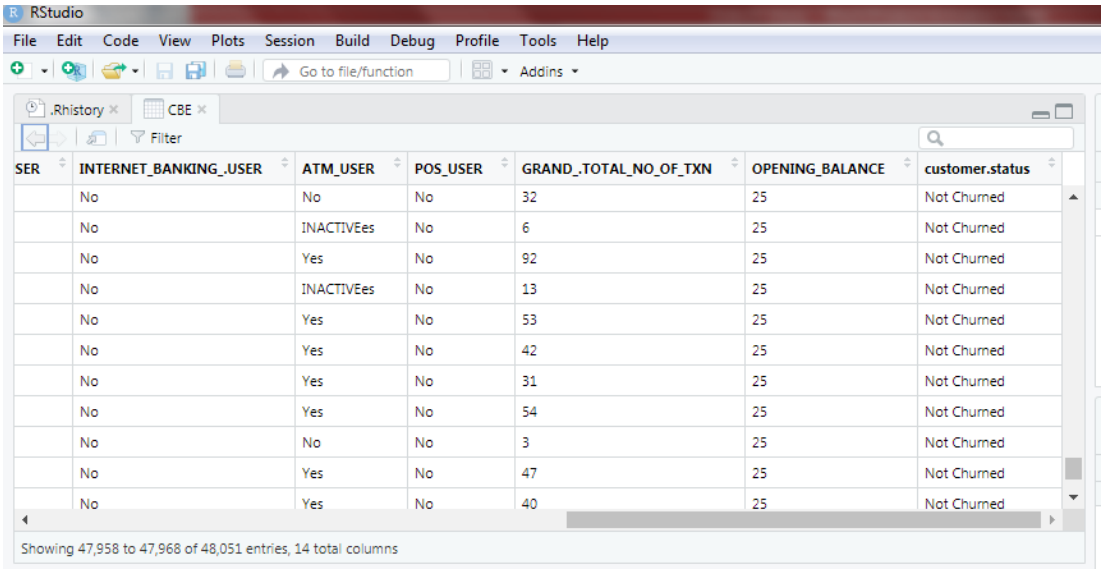
B: Final CBE Dataset Prepared for the Research



C: Structure of CBE Dataset

```
> str(CBE)
'data.frame':   48051 obs. of  14 variables:
 $ SEX                      : chr  "MALE" "MALE" "FEMALE" "MALE" ...
 $ MARITAL_STATUS           : chr  "SINGLE" "OTHER" "SINGLE" "PARTNER" ...
 $ REGION                   : chr  "Benshangul Gumuz" "Somale" "Amhara" "Oromiya" ...
 $ SECTOR                   : chr  "Individual" "Individual" "Individual" "Individual" ...
 $ INDUSTRI                 : chr  "Other Individuals" "Other Individuals" "Other Individuals" "Individual" ...
 $ TARGET                   : chr  "Business Customer" "Business Customer" "Business Customer" "Business Customer" ...
 $ CATAGORIES..PRODUCT_TYPE.: chr  "Youth Saving Account" "Saving Account" "Women Saving Account" "Special Savings Accoun
t" ...
 $ MOBILE_BANKING_USER      : chr  "No" "No" "No" "No" ...
 $ INTERNET_BANKING_.USER   : chr  "No" "No" "No" "No" ...
 $ ATM_USER                 : chr  "No" "No" "No" "No" ...
 $ POS_USER                 : chr  "No" "No" "No" "No" ...
 $ GRAND_.TOTAL_NO_OF_TXN   : chr  "17" "7" "11" "5" ...
 $ OPENING_BALANCE          : chr  "6000" "3,748" "25" "9,500" ...
 $ customer.status          : chr  "churned" "churned" "churned" "churned" ...
```

D: Not Churned Customers in CBE



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

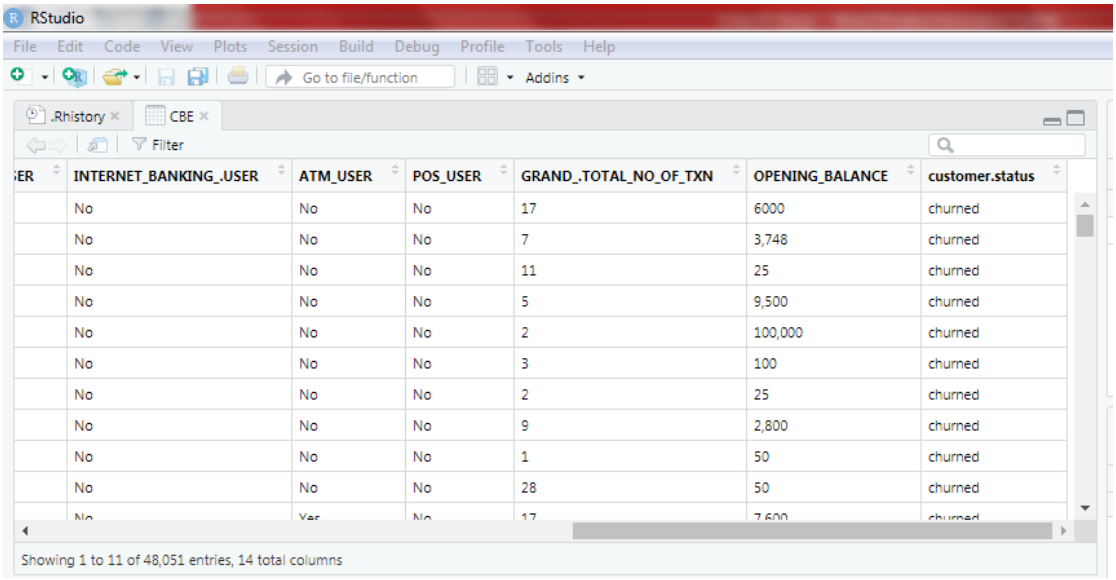
.Rhistory | CBE

Filter

SER	INTERNET_BANKING_.USER	ATM_USER	POS_USER	GRAND_.TOTAL_NO_OF_TXN	OPENING_BALANCE	customer.status
	No	No	No	32	25	Not Churned
	No	INACTIVEes	No	6	25	Not Churned
	No	Yes	No	92	25	Not Churned
	No	INACTIVEes	No	13	25	Not Churned
	No	Yes	No	53	25	Not Churned
	No	Yes	No	42	25	Not Churned
	No	Yes	No	31	25	Not Churned
	No	Yes	No	54	25	Not Churned
	No	No	No	3	25	Not Churned
	No	Yes	No	47	25	Not Churned
	No	Yes	No	40	25	Not Churned

Showing 47,958 to 47,968 of 48,051 entries, 14 total columns

E: Churned Customers in CBE



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

.Rhistory | CBE

Filter

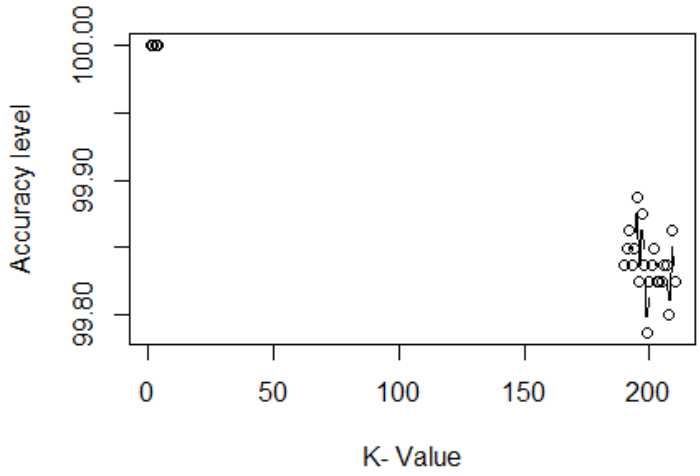
ER	INTERNET_BANKING_.USER	ATM_USER	POS_USER	GRAND_.TOTAL_NO_OF_TXN	OPENING_BALANCE	customer.status
	No	No	No	17	6000	churned
	No	No	No	7	3,748	churned
	No	No	No	11	25	churned
	No	No	No	5	9,500	churned
	No	No	No	2	100,000	churned
	No	No	No	3	100	churned
	No	No	No	2	25	churned
	No	No	No	9	2,800	churned
	No	No	No	1	50	churned
	No	No	No	28	50	churned
	No	Yes	No	17	7,600	churned

Showing 1 to 11 of 48,051 entries, 14 total columns

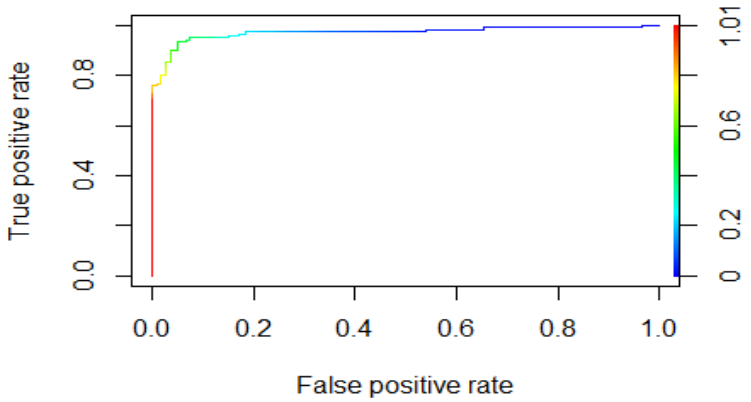
F: Summary of Attributes in CBE

```
> summary(CBE)
     SEX            MARITAL_STATUS       REGION            SECTOR            INDUSTRI
 Length:48051       Length:48051       Length:48051       Length:48051       Length:48051
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
    TARGET          CATAGORIES..PRODUCT_TYPE. MOBILE_BANKING_USER INTERNET_BANKING_.USER
 Length:48051       Length:48051               Length:48051        Length:48051
 Class :character   Class :character           Class :character    Class :character
 Mode  :character   Mode  :character           Mode  :character    Mode  :character
   ATM_USER           POS_USER          GRAND_.TOTAL_NO_OF_TXN OPENING_BALANCE
 Length:48051       Length:48051       Length:48051           Length:48051
 Class :character   Class :character   Class :character       Class :character
 Mode  :character   Mode  :character   Mode  :character       Mode  :character
 customer.status
 Length:48051
 Class :character
 Mode  :character
```

G: Training model of different K values and result plot



H: confusion matrix results snap shot plot



I: Cross table result snap shot

```
> CrossTable(x = y_pred,y= y_act,prob.chisq=FALSE)

   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  7476

             | y_act
      y_pred |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |      3327 |       393 |      3720 |
             |  1443.281 |  1270.494 |           |
             |     0.894 |     0.106 |     0.498 |
             |     0.951 |     0.099 |           |
             |     0.445 |     0.053 |           |
-------------|-----------|-----------|-----------|
           1 |       173 |      3583 |      3756 |
             |  1429.447 |  1258.316 |           |
             |     0.046 |     0.954 |     0.502 |
             |     0.049 |     0.901 |           |
             |     0.023 |     0.479 |           |
-------------|-----------|-----------|-----------|
Column Total |      3500 |      3976 |      7476 |
             |     0.468 |     0.532 |           |
-------------|-----------|-----------|-----------|
```

J: ROC Curve result snap shot