Bank Customer Churn Prediction Based On Transaction Behavior

Fernando GPR

179460U

Dissertation submitted to the Faculty of Information Technology,University of Moratuwa, Sri Lanka for the partial fulfillment of therequirements of the Honours Degree of Bachelor of Science inInformation Technology.

2020

Declaration

We declare that is our own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has beenacknowledged in the text and a list of references is given.

Name of Student	Signature of Student
G.P.R Fernando	……
	Date: ……..

Supervised by

Name of Supervisor	Signature of Supervisor
Mr. ChamanWijesiriwardana	……
	Date:……...

Acknowledgements

I would first like to thank my supervisor, Mr.ChamanWijesiriwardanaof the Senior Lecturer, Faculty of Information Technology, University of Moratuwa. He consistently allowed this thesis to be my own work, but steered me in the right the direction whenever he thought I needed it.his guidance, supervision, advices and sparing valuable time thorough the research project.

I would also like to thank the experts who were involved in the data extraction for this research project. Without their passionate participation and input, the extraction could not have been successfully conducted.

Finally, I must express my very profound gratitude to my parents and to my spouse for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Abstract

Customer churn has become a huge problem in many banks because it costs a lot to acquire a new customer than retaining an existing one. Possible churners in a bank can be identified with the use of a customer churn prediction model and as a result the bank can take necessary actions to prevent those customers from leaving the bank. In order to set up such a model in a bank, few things have to be considered such as how a churner in a bank is defined and which variables and methods should be used. This proposes that a churner for that bank should be defined as a customer who has not been active for the last three months as per the bank's definition of an active customer. Behavioral and demographic variables should be used as an input for the model and classification should be used as a technique.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Businesses are more caring about customer oriented approaches, to survive their business as the business environment has become extremely competitive. This has been affected by the banking sector as well. The financial industry has become a fast growing technology often shortened to fintech, which aims to deliver the financial needs in a more effective, accurate manner with the involvement of modern technology. Over the last decade, private venture capital skyrocketed and the share of investment dollars going into fintech increased from 5% to nearly 20% [8]. In order to fintech growth, customer financial service engagement has been increased. Recently most banks have taken steps to move on to digital banking platforms [9]. It has been the key fact which delivers products and services to the customer in a competitive manner. With the help of digital marketing, banks could reach whenever easily. Switching to a new bank and trying out new products and services is just a matter of a few clicks nowadays. Rather than being loyal it is noticed most of the customers are churners nowadays.

While focusing on customer acquisition, the key feature of business success is customer retention. According to the stats on customer acquisition and retention,[10] Acquiring a new customer can cost five times more than retaining an existing customer. Increasing customer retention by 5% can increase profits from 25-95%. The success rate of selling to a customer, the bank already has is 60-70%, while the success rate of selling to a new customer is 5-20% [11]. Loyal customers are 5x as likely to repurchase, 5x as likely to forgive, 4x as likely to refer, and 7x as likely to try a new offering.

Whether the bank has massive customer centric information, it still couldn't get the maximum usage of quantitative analysis for customer relationship management. Customer loyalty becomes lower and lower too. The 5% customer loyalty decline will lead to a reduction of 2% of the bank profits [4]. In this study these issues are identified and trying to propose a model that can predict bank churn customers. One of the basic short-term revenue goal is increasing the customer base and faster the way to grow a

business. But it is more of a great deal of profit considering up-selling to existing customers makes it more profitable than new customers. While developing new customer acquisition, retaining customers is the key success of a business. Hence this study is about the usage of transaction data (including demographic customer data) to predict and understand customer churn in the banking domain and proposing measures to retain and strengthen customers for better customer relationship management.

1.1 Background

As the showcasing of the banking industry of Sri Lanka on the central bank report [12], the country's economic growth in the Sri Lankan context [12] has a population of approximately 21 million. The real GDP growth is around 4.5%-5% in the recent past. The banking sector continued to expand during the recent years while exhibiting resilience amidst challenging market conditions both globally and domestically. The financial sector is also recognized as one of the fastest-growing sectors of the economy. The banking sector consists of 33 banks, out of which 26 are LCBs (License Commercial Banks) (2 state banks, 11 domestic private banks, and 13 foreign banks) and 7 are LSBs (Licensed Specialized Banks) (6 state banks and 1 private bank). The number of banking outlets and Automated Teller Machines (ATMs) operated by the banking sector was around 7000 and 4000, respectively, whereas the banking density (branches only) per 100,000 persons was 17 during the recent years.

While going through the information it has been noticed that there is a huge competition among the banks and there is a fair market opportunity. Every bank is providing demanding products and services to more customer acquisition. But simultaneously the number of accounts are being closed, why do customers switch from one bank to another? Number of reasons can be considered on customer leave. That can be categorized uncontrollable and controllable reasons. Less engagement is the number one reason, According to Bank Clarity Report stats 61% of the leaving customers are out of a Bank control [13].

When discussing about bank account lifeline. In the context of Sri Lankan commercial banks there are various products that will be suited to customer needs. Normal Saving account products have been selected on this project. It is a general savings account, not specialized for age, profession or income level. Accounts are being on active status means customer engagement is being happened. When any withdrawal had not been made in the recent consecutive two years the account status is changed to Dormant. There should be a withdrawal is made to get back into active status. If any withdrawal had not been made in the recent consecutive ten years the account status is changed to 'Abandoned'. There is a process of transferring a percentage of funds into the central bank from abandoned accounts. Less customer interaction accounts lead that accounts to get into Dormant or Abandoned status. The worst case is account closed status. One of the reasons is that according to customer requests the account gets closed. Another is the accounts couldn't maintain a minimum balance for consecutive 3 months will be getting closed.

Customer churn is defined as the inclination of customers brings to an end doing business with an organization or turn to the services provided by other banks. Accounts which get closed will directly affect product profitability. In order to provide a better insight of churning customer, this study is going to provide a banking customer churn prediction model.

1.2 Problem

Customer churn has become a common problem that many banks having to face. Most of the banks have been facing a problem of closed accounts because, less customer interaction is a painful problem for every bank. Therefore retaining a customer is more profitable same as the new customer acquisition. Less customer engagement leads more closed accounts. There is no method on early identifying churning customers, if it is, bank can engage more on focused customer retention activities easily. Hence there should be a way to early identify churn customer based on their account behavior.

1.3 Aim and Objectives

The aim is creating a customer churn prediction model using the transaction data of customers who have already left the bank and who are already active.

The major objective is to identify customers who are about to leave their bank accounts by studying their transaction behaviors.

Other objectives can be listed as follows. Critically identify the reasons for customers who have already left the ordinary savings. Identify the most likely customers to be left out of the account and find out what remedies are required to maintain the account.

1.4 The scope of research

The main concern is to determine if the customer is going to leave the bank with the help of his behavior. One of the leading Sri Lankan commercial banks is chosen to collect data for this project, using nearly 1,000 customers who left the bank in 2019 and 1,000 active customers, using transactions that have persisted over the past year. Using those transaction data a model will be provided that will beenablingto customers who have already left the bank and who may be expected to leave the bank in the future.

1.5 Summary

This chapter has discussed the problem domain and how a valuable solution is going to be implemented to banking sector. Andclearly defend the objectives and the scope of the research is considering. The data set has been collectedfrom one of the Sri Lankan commercial banks. Based on this, next chapter is going to discuss about the researches which have conducted to address a similar problem domain.

Chapter 2

Literature Review of Bank Customer Churn Prediction

2.1 Introduction

In this chapter, we will delve deeper into how to use data mining techniques on bank customer churn prediction. Then in that regard, we discuss data mining concepts and their uses for popular banking and financial sector customer churn prediction use-cases. It then identifies how data mining techniques can be applied to issues that are already unsolved and dubious. Eventually, this chapter identifies the conceivable data mining techniques that will be used to solve the problem recognized. Lately the chapter has been discussed summary of the problem and challenges identified.

2.2 Data Mining Techniques

Data mining is the process of discovering information by looking at a huge data set. This is not meant by extraction of new data. Instead, this technique involves extrapolating of patterns and new knowledge of data which was already collected. There are several disciplines and procedures, depending on the variety of data mining task. Data mining can be classified into Association, Classification, Clustering, Predictions, Sequential Patterns, and Similar Time Sequences. Machine learning, statistics, neutral network and database are the generally categorized segments, which are depends on different domains.

This can be further explained in detail.Based on the task it is mainlydivide into supervised and unsupervised learning. Supervised learning is a process of learning algorithm based on a training dataset. Input variables and output variables are provided for mapping the algorithm. The goal is to produce an approximate prediction model to outcome the output variable for a given new input data. Unsupervised learning models underlying or hidden structure or distribution in order to learn more about the data. Only provides input variables no correlate output variables are produced. In statics, it can be divided more detailed into regression analysis, clustering, and discriminant analysis and so on.For the neural network methods, it can be divided into self-organizing neural

networks and feed-forward Neural Networks.The main method in database is Multidimensional data analysis and On Line Analytical Processing. The distribution of data mining techniques is figured as shown below.

Although no one type of data mining method is suited for all purposes, algorithm models with unique characteristics are used for each.

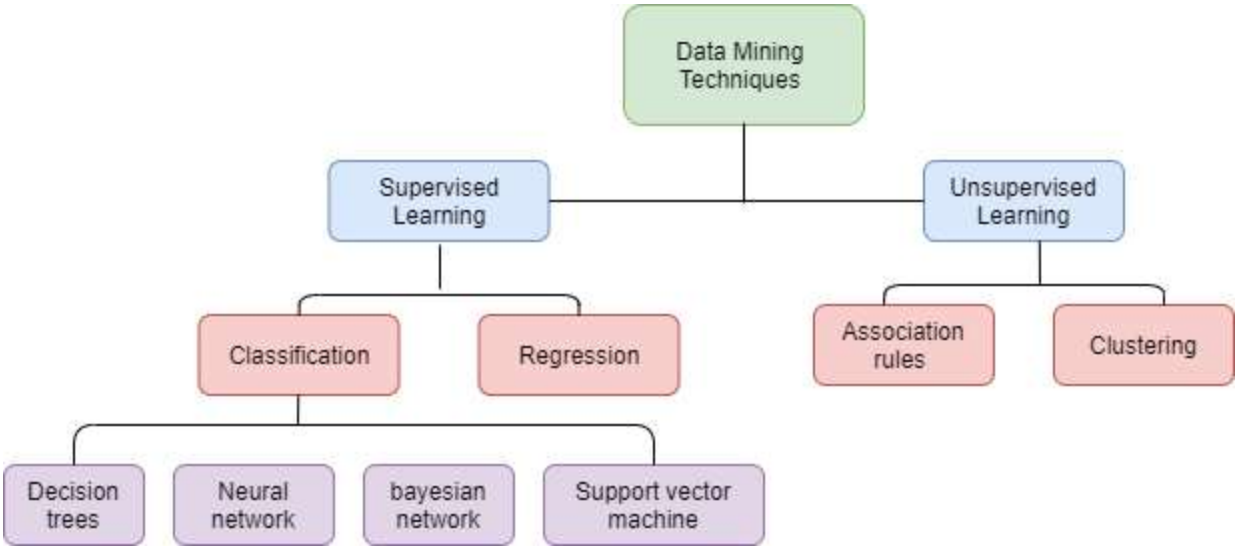


Figure 1: Distribution of data mining techniques

The following section is going to describe in detail of different characteristics and usage of their involvement in the classification technique. In predicting churn customer the class label is selected as churn. Hence the goal is to predict discrete values, e.g. {1,0}, {True, False}, {spam, not spam}.Below is described some of the methods that will be related to our classification problem domain.

2.2.1 Decision Tree

Decision tree algorithm is one of the supervised learning algorithms. Regression and classification problem can be solved by use of this algorithm. The focus of using a decision tree algorithm is to create a model using training data then predict class variable.Based on the type of target variable there are two types of decision trees. Categorical variable decision tree holds target variable as categorical. Decision tree has continues target variable belongs to continuous variable decision tree.Decision tree classify the data set by sorting down from the root node. The very top of the tree is called

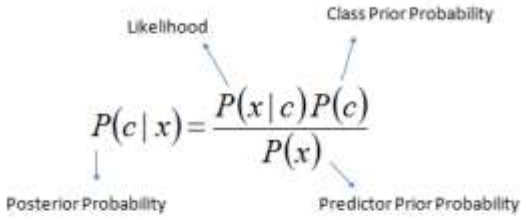
the “Root Node”. Following the nodes from “root node” are called “internal nodes”. Lastly are called “Leaf Node”.Identifying which attributes need to consider as the root node and each level is the main challenge of the decision tree.

2.2.2 Neural Network

Neural network consist of input, output and hidden layers. These layers main task is transformation of input into variables output unit. By considering information flows neural network happens in two ways. Feed forward networks signals only travel in one direction without any loop. This consists of single input layer and a single output layer. This extensively used in pattern recognition. Feedback recurrent or interactive networks can use their internal state (memory) to process sequences of inputs. Signals travel in both directions with loops [16]. Neural networks have a unique ability to extract meaning from imprecise or complex data to find patterns and detect trends that are too convoluted for the human brain for other computer techniques.

2.2.3 Naïve Bayes

Naive Bayes classifiers are a collection of classification algorithms based on Bayes’ Theorem. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c).



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).

P(c) is the prior probability of class.

P(x|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of a predictor.

2.2.4 Support Vector Machine

Support vector machine (SVM) is one of the supervised machine learning methods. That is used for classification problems. Support vector machine is highly preferred by many as it produces significant accuracy with less computation power. Support Vector Machine can be used for both regression and classification tasks. But, it is widely used in classification objectives. By plotting each data item as a point in n-dimensional space with the value if each feature then perform classification by finding the hyper-plane that differentiates the two classes very well. The SVM classifier is a best segregates the two class problems[17].

2.3 Literature Review

Whether the banks have vast dataset of the customers still most of them are in an ideal state. Most of the applications currently being used is not contributing much accepted manner for gaining better customer relationship management. Data mining techniques are the key field to extract valuable information nowadays. Existing research extraction may focus on the domain of customer churn:

ShrishaBharadwaj[5] put forward a Customer Churn Prediction in Mobile Networks, which helps to predict Marketing research is conducted to predict customer needs. They have developed two predictive models using logistic regression and Multilayer Perceptron (MLP) to produce research data, so that organization can study the customers and get what they want. This enables customers to identify their needs and develop their business accordingly before they can request a customer. WojewnikP [6] has been used K-mean and classification algorithms to produce hybrid algorithm for customer churn prediction. Guoxun Wang [1] has shown how to apply different classification algorithms in a credit card holder's dataset. Furthermore a performance evaluation has been made on each classification algorithm.

JussiAhola and EsaRinta-Runsala[2] have stated prediction model problem was solved using two different methods. The main purpose of this case study was to get insight of information and identify patterns. It has been used for clustering and then precedes a prediction analysis using a set of identified variables. Here, logistic regression and classify the customers of the bank with decision trees are used to create a model and determine the profitability of its members.

Peters, Edward M.L. et al.[3] have implemented a model for understanding Service Quality and Customer Churn by Process Discovery for a Multi-National Banking Contact Center. The first step was to identify the quality and speed of the service of front office employee. Many publications have shown that collection of data manually. But where the employee's behavior may change under the observation and resulting data may error prone. Because of that automated behavioral collection method was used with 5000 employees. This paper shows that the behavior of the bank's front office and back office employees affects the equality and speed provided by the service delivery, where a number of important outcomes have been achieved through the identified factors stated there was an intervention on customer churn reductions leads, retains customers.

Zhao Jing and Dang Xing-hua [7] put forward a VIP customer churn model using support vector machine. Churners are defined as average daily balance of deposits is less than 20000 Yuan. Observation period was 3 consecutive months.

Shaoying Cui and Ning Ding [4] have implemented customer churn prediction using improved FCM algorithm, Using Cluster, they have extracted patterns into a real-world data set, adopted a variety of methods, and graded multiple profiles.The algorithm has defined clustercenters, means clustering method has been most commonly used. In this study 3 months' historical data and basic customer information is used. Customer behavior was not critically analyzed; the study was mainly focused on improving the FCM algorithm.

At the point of summary data mining technology can be used to successfully identify important information in large business databases such as banks. It can also be used for customers churn prediction more efficiently by extracting valuable hidden information. In

order to fulfill that most of the researchers contributed on different view point. But the accuracy rate of by analyzing the customer behavioral sense is not high enough.

The research direction goes to predict the bank customer churn according to their behavior. Based on the behavioral data, the most commonly used and effective methods in the customers churn prediction classification [1] are going to be used.

2.4 Summary

In this chapter we have discussed about the machine learning techniques and how they are going to be addressed to solve customer churn prediction. While gone through the series of similar research. It has observed there is a research gap to detect churning customers based on their transaction behaviors. Furthermore the next chapter is discussing about technologies going to be aligned with this research.

Chapter3

Technologies Adapted

3.1 Introduction

In the previous chapter, we have discussed different findings in the area of customer churn predictions, its developments, issues and future challenges and we define our research problem and also identified classification data mining as the technology to address the problem. This chapter is discussing technologies which are going to be applied to the model the bank customer churn predictor.

3.2 Data Mining Tools

Describedbelow are some of the popular data mining tools and WEKA has been selected as the tool for modeling bank churn predictor and the applicability if it is going to provide in details in further sections.

3.2.1 Rapid Miner

Rapid Miner is one of the best predictive analysis system developed in JAVA programming language. It provides an integrated environment for deep learning, text mining, machine learning & predictive analysis. The tool can be used for over a vast range of applications including for business applications, commercial applications, training, education, research, application development, machine learning [18]. Fast visual workflow designer is leading key feature of its user-friendliness.

3.2.2 Orange

The orange program is written in python language. It is best performing at data visualization, the component base software perfect suite for machine learning and data mining. It also provide tools for analytic prospective. It performs simple data analysis with clever data visualization and explore statistical distributions, box plots and scatter plots, or dive deeper with decision trees, hierarchical clustering, heat maps, MDS and

linear projections, even your multidimensional data can become sensible in 2D, especially with clever attribute ranking and selections [19].

3.2.3 WEKA

WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can be applied directly to the processed dataset. WEKA contains tools for areas such as data pre-processing, classification, regression, clustering, association rules, and visualization. Here WEKA is used for generating classification mining results of the banking customer transaction dataset.

Also known as Waikato Environment which is machine learning software developed at the University of Waikato in New Zealand. It is best suited for data analysis and predictive modeling. It contains algorithms and visualization tools that support machine learning. WEKA has a GUI that facilitates easy access to all its features. It is written in JAVA programming language. It works on the assumption that data is available in the form of a flat file. WEKA can provide access to SQL Databases through database connectivity and can further process the data/results returned by the query [18].

3.3 Java

Java is a general-purpose powerful programming language. Java desktop application has been implemented named “churnapp” for data preparation. To transfer raw data in the process of filtering and proposing, resulting data has been taken as the input for WEKA. Here churnapp is used to read data from churndbmysql database (Staging database on the transaction) after going through the preprocessing business logic data is served that can be directly exported into WEKA.

3.4 MySQL

The database server MySQL is used for staging purposes. A managed database named **churndbprodml** is used to upload selected transactions and customer data. Relational tables are prepared to cleanse the data set. Raw data are imported via .csv format and then processed data are exported via .csv in order to do model building with WEKA.

3.4 Summary

This chapter have detail discussed the technology proposed to customer churn prediction model. While going through the compatibility and scale up java and WEKA have been selected the main component for the data modeling. The next chapter will going to show the approach of the banking customer churn prediction of the technologies listed here.

Chapter 4

A Novel Approach to Predict Banking Churn Customer

4.1 Introduction

In my approach, it is going to implement a prediction model that can be used for the early identification of bank customers who are going to churn. In the third chapter, we have considered the technologies that are required to be solved the research problem properly. In this chapter the approach is to find out how the presented technology is going to be applied into the problem domain of building a model for early identification of banking churn customers. Bank churn is the main part of the solution. The hypothesis was under consideration. The bank and prediction model, which builds on a variety of inputs and their processing, also describes its features critically.

4.2 Hypothesis

Through the hypothesis the customer churn prediction in the banking and finance sector does not have a proper mechanism, and the customers in the financial sector are analyzing the transaction that they associate with that entity as their behavior to predict their churn probability is an added value for sustain the business. This hypothesis was impact to use massive amount of customer behavioral data is still belong hidden information of their customer. That will be competitive advantage to extract and use to retain customers early as possible.

The hypothesis of this research is early identification of the customers who are going to be churned recently from the bank can be achieved by using classification analysis using there transaction behavior.To conclude, the classification algorithm is applied in a variety of ways, comparing their accuracy and ultimately using the most accurate technique to build the model.

4.3 Methodology

4.3.1 Data Collection

This is done by predicting banking customers' churn using their transaction behavior. To do this, the bank will obtain information about customers who have already left and are already active, and their transaction behaviors. From the CORE banking system, the data has extracted about the account holders in a selected branch for a period of one year. All the transactions related to the behavior of the customer is recorded in the core banking system and so that the required data files can be downloaded directly.

4.3.2 Data Preprocessing

Generally, this data is preprocessed so as to eliminate the meaninglessness of some data. Since the data is directly extracted from the core banking system for the banking customer churn prediction model, these are most likely to be accurate.

4.3.3 Feature Extraction

While obtaining the data set, customer information and transactions data are downloaded. Not all parameters are necessary to build this prediction model; hence there should be a feature extraction process to identify the most effective variables. The accuracy of the Predictive model is directly affected by the feature list, so selecting the best fit feature list is a challenge.

4.3.4 Classification

The Classification rule is used to build the predictive model. Classification is the best-known and most used method of data mining. The aim of the classification method is to accurately predict the target class of objects of which the class label is unknown [14].

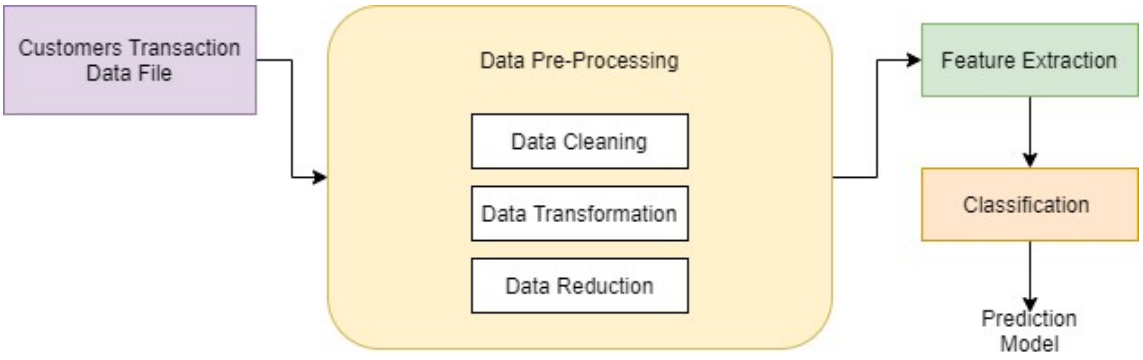


Figure 2 : Model building methodology

4.4 Input

Customer behavior is defined as customer transactions. A branch is selected which indicates heavy transaction volume, within the selected time period closed accounts and active customer lists are extracted. The transactions related to appearing accounts are separately extracted as the customer behavior list.

	D	I	J	K	L	N	O	P	Q	R	S	T	U	V	W	X
1	SNAME	DATE_OPEN	DATE_LAST_ACT	DATE_CLOSED	DATE_OF_BIRTH	ID_TYPE	SEX	CUS	PROFESSION	AVG_BALANCE (6 M)	BALANCE	ACS.	BALANCE	ACS.	BALANCE	ACS.
2	GD AMITH	09/09/1996	22/02/2017	01/06/2019	03/06/1960	NIC	F	7		18.2	0	0	0	0	0	0
6	RN EAPASINGHE	29/03/1990	03/10/2011	01/04/2019	25/12/1954	NIC	M	7	Pensioner	1.2	5236.05	2	0	0	163880.93	1
7	DM SIRISENA	16/07/1996	26/01/2015	01/01/2019	18/04/1943	NIC	M	7		0	172304.15	3	1500000	8	0	0
8	P GNANASIRI	07/10/1996	13/01/2010	01/07/2019	22/01/1954	NIC	M	7		40.25	1109205.61	5	4852865.71	9	0	1
9	DB KANNANGARA	25/08/1998	08/03/2014	01/01/2019	23/10/1952	NIC	M	7	Pensioner	0	44477.71	2	1230000	4	36701.82	3
10	JKKC THILAKARATHNA	22/10/2004	10/05/2019	10/05/2019	14/11/1961	NIC	M	7		297.33	485354.97	8	1121180	2	0	2
11	HLDJDL SENAVIRATNE	24/01/1983	11/07/2017	01/06/2019	01/05/1943	NIC	F	7		13.76	0	0	0	0	0	0
12	MKDS GUNARATNE	01/02/1983	25/10/2016	01/05/2019	07/02/1976	NIC	M	7		9.91	0	0	0	0	0	0
13	WK DISSANAYAKE	01/06/1986	18/05/2016	01/07/2019	11/08/1953	NIC	M	7		21.76	0	0	0	0	0	0
14	C SAMARAWEERA	14/07/1987	16/03/2009	01/09/2019	10/07/1955	NIC	F	7		110.23	4163260.88	3	6468318.02	4	0	0
15	KGD PEREIRA	08/12/1987	12/09/2019	12/09/2019	12/08/1960	NIC	F	7		137602.76	0	0	0	0	0	0
16	GA GURUSINGHE	04/08/1992	17/12/2014	01/04/2019	04/01/1954	NIC	F	7		0.19	0	0	0	0	0	0
17	HAP PERERA	27/11/1992	17/01/2019	17/01/2019	26/06/1953	NIC	M	7		0	0	0	0	0	0	0

Figure 3 : Customer master data

According to my approach, a bank's customers are predicted to churn, using their transaction behaviors, to obtain data from customers who have already closed accounts and already have active accounts. Transaction data is extracted separately for customers who receive their transactions behavior for one year.

	A	B	C	D	E	F	G
1	ACCTNO	TRN_DT	DR/CR	AMT	TRANCD	AUXTRC	TRREMK
2	8200	30/09/2018	C	14.63	160		
3	8200	30/09/2018	D	0.73	198		
4	8200	03/10/2018	C	3100	1	6072	0330CB01
5	8200	04/10/2018	D	30005	2	6054	4693429008188530
6	8200	05/10/2018	C	40000	1	6072	0181CS11
7	8200	05/10/2018	D	10005	2	6059	0002AB01
8	8200	07/10/2018	C	3100	1	6072	0330CB01
9	8200	08/10/2018	D	46005	2	6054	4693429008188530
10	8200	09/10/2018	D	7005	2	6059	0355AB01
11	8200	15/10/2018	C	10000	1	6072	0088CB01
12	8200	17/10/2018	C	10000	1	2111	
13	8200	18/10/2018	C	10000	1	6072	0088CB01
14	8200	18/10/2018	D	505	2	6054	4693429008188530
15	8200	19/10/2018	D	1000	2	6151	4693429008188530
16	8200	22/10/2018	C	15322	1	2111	
17	8200	24/10/2018	D	1000	2	6151	4693429008188530
18	8200	28/10/2018	D	40005	2	6059	0185AB01
19	8200	31/10/2018	C	3.51	160		
20	8200	31/10/2018	D	0.17	198		
21	8200	03/11/2018	D	305	2	6054	4693429008188530
22	8200	07/11/2018	D	7.5	2	6170	4693429008188530
23	8200	07/11/2018	D	7.5	2	6170	4693429008188530
24	8200	07/11/2018	D	2530	2	6174	4693429008188530

Figure 4 : Transaction data

The “master data file” which holds customer demographic variables is extracted from core banking system. Not only that, the“transaction data file” is extracted that consists of selected customer transaction for specific time duration. This means that customers' behavior with a bank product can be displayed accordingly. Then identify most effective transaction variables on churn customers such as number of debits, withdrawals, average balance etc. and customer age, engagement period etc. are used for the conditions for demographic variables for predict churn customer.

Positivist paradigm is used as research paradigm test hypothesis by quantitative data concluded from transaction category count. Churn variable is identified as dependent variable and transaction categories are identified as independent variables. Elaborative logical sense of relationships among extracted variables has been identified.

4.5 Output

In order to predict churning customers, identified closed accounts and active accounts customer transaction behavior is used. By considering above logic and data, the records of bank customer behavior are filtered using WEKA tool, classification rule mining, predict churning customers based on the transaction behavior. Best performing classification algorithm is going to be used to build the prediction model.

4.6 Process

In order to build a churn prediction model by using extracted feature set of selected customer segment WEKA data mining tool is being supported. By directing the data set to the WEKA mining tool different classification techniquesare applied. While going through the result check and compared the accuracy level for best suited technique for predicting customer churn. The process is preparing the dataset that could showcase the customer behavior form their transaction. Outcome prediction model is exported; to predict the customers who will be going to churn recently is done by extracting the identified features transaction records and applying to the churn prediction model.

4.7 Features

The data set is used according to the selected banking segment. Transaction records are generally available information on every financial sector. Hence to identify churners, build a mechanism thatusesthese transaction records is profitable and this will help to gain exact understanding of the customer behaviors. Accuracy of the data set for model building is a highlighted key feature. WEKA has been used as a data mining tool because it is open source, freely available and its expandability.In addition to the transaction features; we can further improve by including more demographic features.

4.9 Summary

This chapter has presented the data mining concept that is going to apply to sole our research problem, while going through the methodology of implementing a churn predictor. Hypothesis, input, output and process have been presented. Java based data mining tool named WEKA has been identified and selected as the model building tool. The analysis and design of proposed bank churn customer predictor will be discussed in the next chapter.

Chapter 5

Analysis and Design

5.1 Introduction

Last chapter we have discussed the data mining concepts and what are selected to solve the bank customer churn prediction, with combination of all extracted knowledge and data on the banking domain. Here it considers how the design and analysis is done.

5.2 Top level architecture

This research is based on quantitative data which is presented in a numerical format, collected in a standardized manner, analyzed using statistical techniques. For archive above objectives following design method was engaged.

Listed herewith are the main components of banking customer churn prediction model.

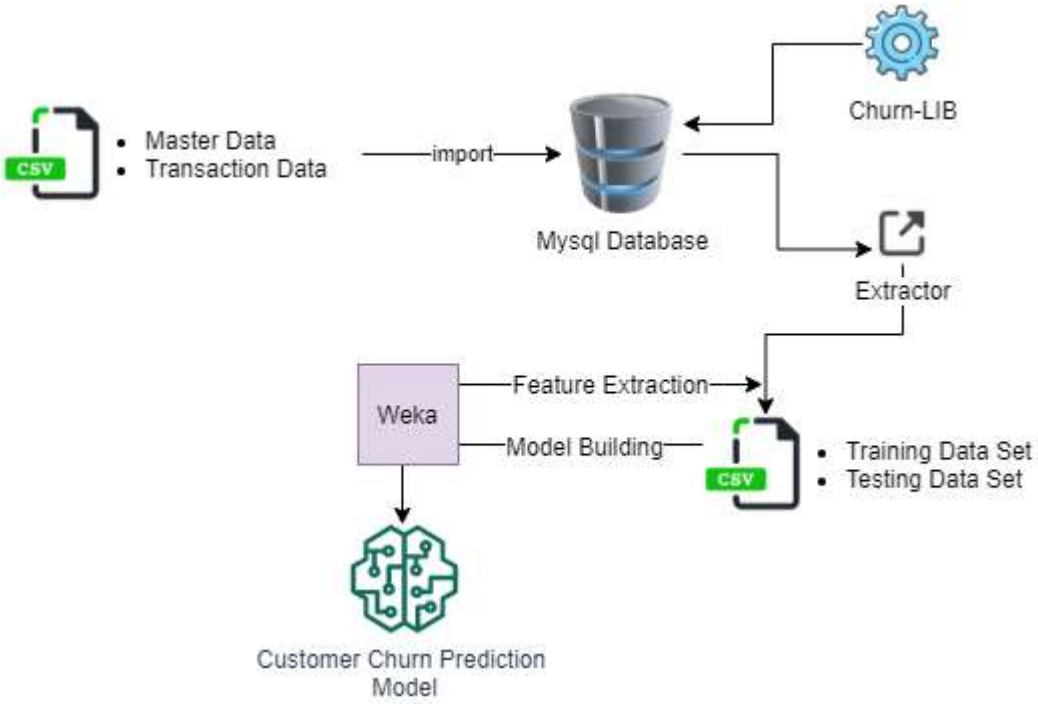


Figure 5 : Churn prediction design

5.3 Data set

Customer demographic and transaction data are extracted from the CORE banking system. Data are separately extracted for already closed accounts and active accounts from a defined time stamp. The fields on master data file are.

ACCTNO,BRANCH,CIFNO,SNAME,SCCODE,ACTYPE,STATUS,DDCTYP,DATE_OPEN,DATE_LAST_ACT,DATE_CLOSED,DATE_OF_BIRTH,ID_NO.,ID_TYPE,SEX,CUST_CLASS,PROFESSION,AVG_BALANCE_6M,BALANCE,ACS.,BALANCE,ACS.,BALANCE,ACS

Transaction details which related to master data account holders are separately extracted as follows:

ACCTNO,TRN_DT,DR_CR,AMT,TRANCD,AUXTRC

Those records are imported into the MySQL database for management in a more convenient way. A single transaction belongs to a particular transaction type. There are two main categories of transaction code: TRANCD and AUXTRC. The transaction which involves the teller are appeared under AUXTRC and system level transactions are categorized under TRANCD. Detailed list of TRANCD and AUXTRC are extracted via AUX_TRAN_CODE_PARAMETERS.csv

5.4 Database

5.4.1 Staging Data

Collected records are stored in the staging database. There should be history transaction records of customers who have already closed accounts and those who have an active account. Before analyzing the transaction behavior data of customers, it is needed to preprocess the data by using preprocessing techniques.

The preparation has to be involved for arranging transaction records for mining format. TRN and AUX transaction codes are separately considered. The research is going as a quantitative analysis, so each transaction related count is taken as a quantitative value. It is noticed that a large number of transaction codes were recorded on customer

transactions. There, a separate desktop application (churn-LIB) is built to arrange the record set horizontally. Data are loaded into listed staging tables on MySQL.

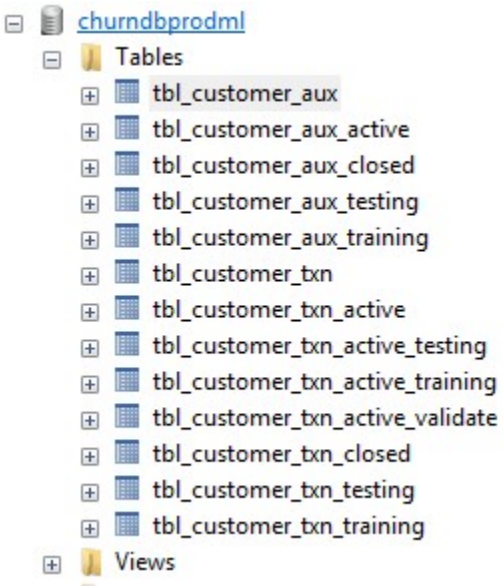


Figure 6 : Records staging tables

5.4.2 Data Preparation

Data preparation is the fundamental stage of data mining. Data are collected from various sources. Before applying those to a data mining processthe data should be arranged in acceptable format. The preparation is a critical process in any data mining process as it leads to reduce complexity of unclean real world data.In predicting churn customer forecasts, data collected from the core banking system is distributed vertically.A standalone java program “churnlib” has been implemented for convert a data segment to horizontal tabular formatted data. Standard SQL is used to read and modify the identified features.

5.5 Churn-LIB

The churn-LIB is a java application which has been written for preprocessing customer transaction data. Some demographic data has been made that will be used for treats as independent variables. Account age (ACC_AGE) is calculated by means of account created date. This variable can give the information how long the customer is being

served with the bank product. Total debit (TOTAL_DR) count is made as a new variable that can showcase how many instances of debit transactions happened on particular customer instances. Total credit count (TOTAL_CR) is a newly introduced variable which represents the number of credit instances on a customer. Maximum debit amount (MAX_DR_AMOUNT) is recorded as the maximum debit amount of a customer on the dataset. Same as Maximum credit amount (MAX_CR_AMOUNT) is recorded as the maximum credit amount of a customer on the dataset. Customer age (CUST_AGE) is calculated for data extracted timestamp.

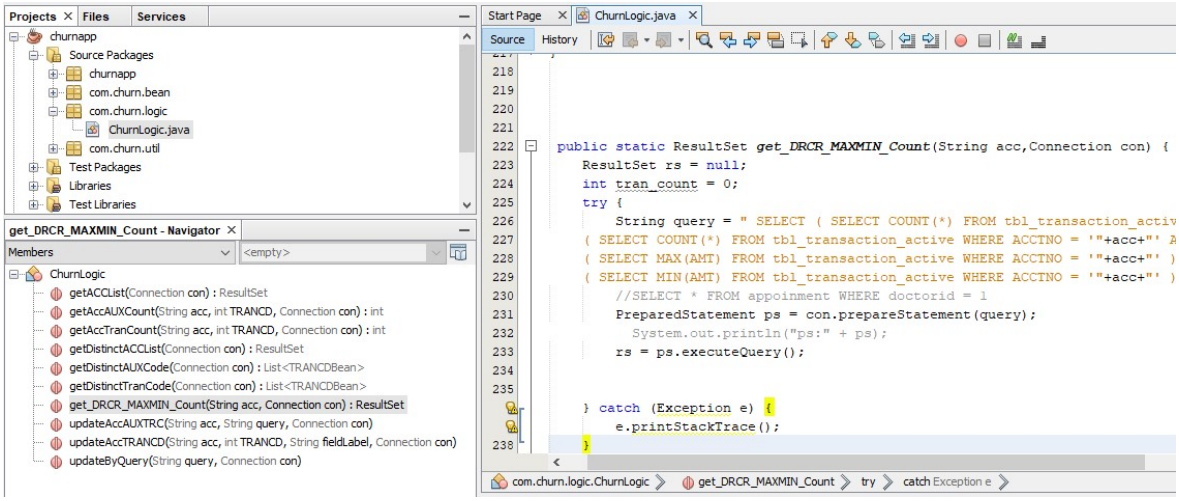


Figure 7 : Churn-LIB application

5.6 Data Preprocessing

Preprocessing the data is an important task. And it is very important for the accuracy, this is because the data is mostly noisy and it sometimes has missing values as well as false values. Data is rarely clean and often can have corrupt or missing values. It is important to identify, mark and handle missing data when developing machine learning models in order to get the very best performance. Before getting the data into the prediction process, the data had to be preprocessed. So feature selection and feature extraction is two major processes of preprocessing that directly impact the accuracy of the model. Feature selection is a method of selecting some data and discarding irrelevant.Transaction data

are presented on relating to selected customers and therefore no irrelevant and redundant information exists.

5.7 Exploring Dataset

The transformed dataset is extracted into .csv file format which is going to directly import into WEKA data mining tool.

5.7.1 About Modeling

Modeling is a statistical technique to predict some kind of given outcome, in this instance it is bank customer churn prediction. In data terminology we divide inputs and outputs; inputs are predictors and these are also called independent variables. From the extracted features,customer churn is used as dependent variable.

5.8Classification Model Building

Preprocessed data are used to rank the attributes. It resulted in a large number of attributes which are representing each TRANCD and AUXTRC within those best effecting attributes should be filtered out. Classification rule has been applied as supervised machine learning algorithm to build the prediction model. After that need to validate the model with the test dataset. It needs to apply and check if the model is working properly with fresh data. Customer churn variable is identified as dependent variable while simulating of TRANCD and AUXTRC data checks how it is changed while applying values from currently being active customers.

5.9 Summary

In this chapter we have discussed how to analyze and plan to solve the problem of building churn predicting model of bank customers.The data set was identified. Data preparation is done using an implemented standalone java program. In the next chapter we are going to discuss in detail on how to implement it.

Chapter 6

Implementation

6.1 Introduction

In the last chapter we have discussed the research design in detail. Data collection, data preparation steps, preprocessing, machine learning algorithms and how those will be applicable into customer churn prediction. In this chapter we are going to discuss the way of these identified design steps can be applied practically, the way of source data are being collected, implementation of churnLib as data preparation library, data preprocessing techniques that were used to customer transaction dataset and the process of model building algorithms and evaluation.Finally best performing algorithm is going to be selected.

6.2 Data Transformation

First, the active customer's demographic data and their transaction data are extracted throughout the year in related to the selected branch. The two datafiles (csv) obtained are saved separately. And the same way two files related to customers who have already left the account will be saved. The files are named as master saved-data.csv , transaction-data,csv. And then imported to a mysql database where all the records in the master data, transaction data are appeared.

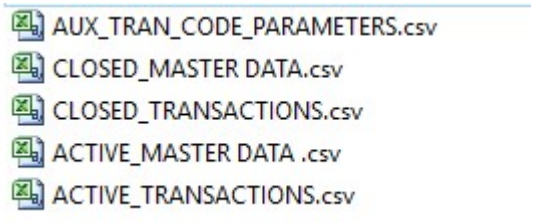


Figure 8 : Customer master – transaction data files

One of the main dataset of the prediction model is the history details of the customer's transactions. While examining the details it was noticed every transaction is under a transaction code. Therefore, all the transaction codes in the core bank are imported to the staging database using a separate file, to identify the transaction codes separately. These transaction codes will be identified in the future as customer behavior. That is, the code that accompanies a relevant transaction can be used to see if the customer has been engaged to a bank product. To convert appeared transaction codes as feature variables it is required to perform a data transform.

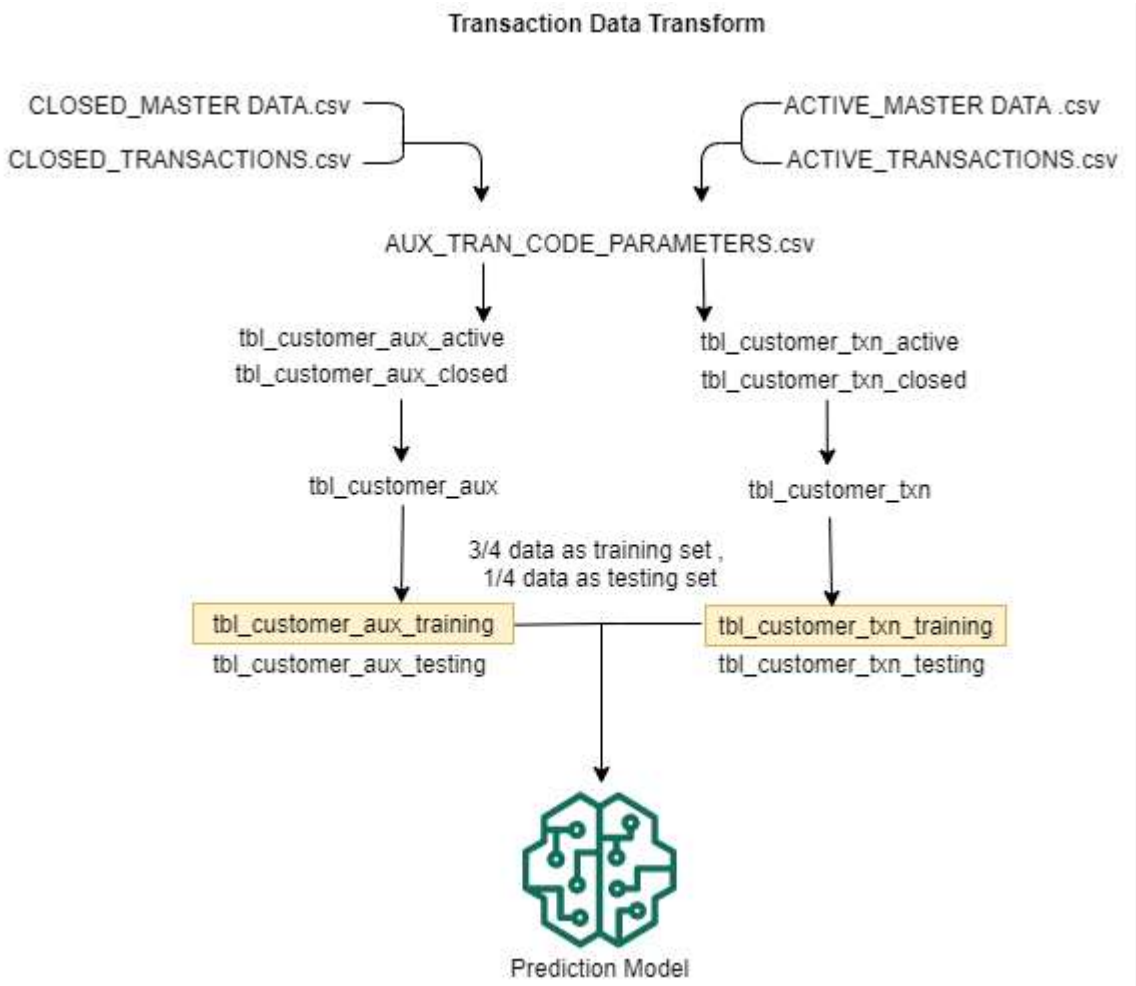


Figure 9 : Transaction Data Transform

The transaction count for each customer is included in the feature variable that identifies the counter. This numerical count is very important in advancing research as quantitative analysis.

This procedure works for both active customers and churn customers separately and store their results in separate tablets.

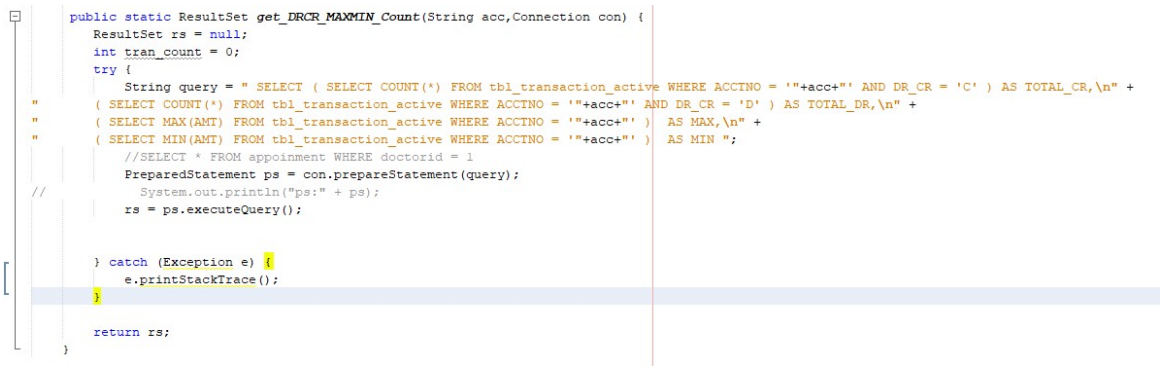


Figure 10 : Statistical logic in churn-LIB

Transaction codes can be identified under two main categories: txn and aux. The aux code associated with a transaction, is the way the customer has dealt with a teller transaction, be it ATM machines, branch banking, counter transactions on the branch network. TXN refers to systematic transactions, such as when a transaction occurs or the banking system automatically, allows further account-to-account transactions. The research is carried out in two main ways: using aux and txn to separately analyze transaction records. It sets the data-file from the appropriate MySQL table to determine the customer behavior. It uses a churn program to obtain transaction counts for its field as quantitative values. Resulting tables are tbl_customer_aux_closed,tbl_customer_aux_active,tbl_customer_txn_active, tbl_customer_txn_close. Transaction history files all the aux, txn codes shown have been attached to these tables. As a result, a significant number of features are contained in the data set. The aux and txn features together form a table of data, instead of the now considered close and active customer records.

6.3 Explorer the Data

Before proceeding with data mining activity it is good to explore the dataset. It is important for getan understandon how the extracted data are distributed. Then we can identify missing value, irrelevant features for our preprocessing task. The selected data mining tool is WEKA.

Tbl_customer_txn.csv has been loaded into WEKA explorer. There are 1541 instances; those correspond to 1541 customer's transaction behaviors that were filtered out under 'txn' transaction code. On each transaction we got 78 attributes.

ACCTNO	CUST_CLASS	7_OVERBOOKING_ACCOUNT_CLOSING	143_AUTOMATIC_TRANSFER	215_NBT_ON_SAFE_DEPOSIT_RENT
BRANCH	PROFESSION	15_TRANSFER_OUT	144_AUTOMATIC_TRANSFER	400_AGF_AUTO_GRAB_FUND
CIFNO	AVG_BALANCE_6M	20_CLEARANCE_PAYMENT	145_AUTOMATIC_TRANSFER	472_MIN_BAL_ADMIN_FEE_PASSIVE_ACC
SNAME	OTHER_CASA_ACC_BALANCE	48_VAT_ON_POSTAGE_FEE	145_BONUS_INTEREST	551_BATCH_SWEEP
SCCODE	OTHER_CASA_ACC_ACS	50_CHEQUE_PAYMENT	155_DEPOSIT_INTEREST_PAYMENT	553_BATCH_SWEEP
ACTYPE	OTHER_FD_ACC_BALANCE	51_MISCELLANEOUS_FEE	158_MAIN_DEPOSIT_WITHDRAWAL	950_DEBIT_CHARGE_STATIONARY
STATUS	OTHER_FD_ACC_ACS	55_OVERBOOKING_ACCOUNT_CLOSING	160_INTEREST_ACCOUNT	960_DEBIT_VAT_STATIONARY
DDCTYP	OTHER_LOAN_ACC_BALANCE	70_BILL_PAYMENT_CREDIT	180_MIN_BAL_ADMIN_FEE_PASSIVE_ACC	970_DEBIT_NBT_STATIONARY
DATE_OPEN	OTHER_LOAN_ACC_ACS	71_BILL_PAYMENT_DEBIT	198_WITHOLDING_TAX	NON_ACTIVE_PERIOD
DATE_LAST_ACT	1_CASH_PAYMENT	73_B_PYMNT_SERVICE_CHARGE_DEBIT	203_NBT_ON_CHARGES	ACC_AGE
DATE_CLOSED	2_CASH_WITHDRAWAL	75_MINI_STATEMENT_CHGS	205_POSTAGE_FEE	TOTAL_DR
DATE_OF_BIRTH	3_CHEQUE_PAYMENT	82_LOCAL_CHEQUE_CREDIT	206_VAT_ON_POSTAGE_FEE	TOTAL_CR
ID_NO	4_TRANSFER	89_DEBIT_CORRECTION_TRANSACTION	207_VAT_ON_AFT_FEE	MAX_DR_AMOUNT
ID_TYPE	5_TRANSFER	133_AUTOMATIC_TRANSFER_FEE	208_VAT_ON_SAFE_DEPOSIT_RENT	MAX_CR_AMOUNT
SEX	6_ACCOUNT_CLOSING	141_AUTOMATIC_TRANSFER	209_NBT_AFT_FEE	CUST_AGE
		142_AUTOMATIC_TRANSFER	214_NBT_ON_POSTAGE_FEE	churn

Figure 11: Feature list

The class attribute is churn, two distinct values YES, NO is distributed with 769 and 772 and instances. This will be the classification problem that comes to predict "churn" variable. This is also called a supervised learning because we get to know the class value from training instances. These instances are independent with the class value is attached. The idea is to produce automatically some kind of model that can classify new customer transaction behavior. While going through the attribute some of the fields have identified those are not given a high contribution for analysis those attributes are removed. ACCTNO,BRANCH,CIFNO,SNAME,SCCODE,ACTYPE,STATUS,DDCTYP,DATE_OPEN,DATE_LAST_ACT,DATE_CLOSED,DATE_OF_BIRTH,ID_NO,ID_TYPE, PROFESSION and CUST_CLASS attributes are removed from the data set. At removing unrelated attributes, 62 attributes we have to have in our dataset. These attributes are continuous ("numeric"). The class is discrete and it consist whether the customer churn status yes or no. The data file is saved as .arff file initial attributes are listed as flowed.

@relation 'tbl_customer_txn-weka.filters.unsupervised.attribute.Remove-R1-14,16'
@attribute SEX {M,F}
@attribute AVG_BALANCE_6M numeric
@attribute OTHER_CASA_ACC_BALANCE numeric
@attribute OTHER_CASA_ACC_ACS numeric
@attribute OTHER_FD_ACC_BALANCE numeric
@attribute OTHER_FD_ACC_ACS numeric
@attribute OTHER_LOAN_ACC_BALANCE numeric
@attribute OTHER_LOAN_ACC_ACS numeric

@attribute 1_CASH_PAYMENT numeric
@attribute 2_CASH_WITHDRAWAL numeric
@attribute 3_CHEQUE_PAYMENT numeric
@attribute 4_TRANSFER numeric
@attribute 5_TRANSFER numeric
@attribute 6_ACCOUNT_CLOSING numeric
@attribute 7_OVERBOOKING_ACCOUNT_CLOSING numeric
@attribute 15_TRANSFER_OUT numeric
@attribute 20_CLEARANCE_PAYMENT numeric
@attribute 48_VAT_ON_POSTAGE_FEE numeric
@attribute 50_CHEQUE_PAYMENT numeric
@attribute 51_MISCELLANEOUS_FEE numeric
@attribute 55_OVERBOOKING_ACCOUNT_CLOSING numeric
@attribute 70_BILL_PAYMENT_CREDIT numeric
@attribute 71_BILL_PAYMENT_DEBIT numeric
@attribute 73_B_PYMNT_SERVICE_CHARGE_DEBIT numeric
@attribute 75_MINI_STATEMENT_CHGS numeric
@attribute 82_LOCAL_CHEQUE_CREDIT numeric
@attribute 89_DEBIT_CORRECTION_TRANSACTION numeric
@attribute 133_AUTOMATIC_TRANSFER_FEE numeric
@attribute 141_AUTOMATIC_TRANSFER numeric
@attribute 142_AUTOMATIC_TRANSFER numeric
@attribute 143_AUTOMATIC_TRANSFER numeric
@attribute 144_AUTOMATIC_TRANSFER numeric
@attribute 145_AUTOMATIC_TRANSFER numeric
@attribute 145_BONUS_INTEREST numeric
@attribute 155_DEPOSIT_INTEREST_PAYMENT numeric
@attribute 158_MAIN_DEPOSIT_WITHDRAWAL numeric
@attribute 160_INTEREST_ACCOUNT numeric
@attribute 180_MIN_BAL_ADMIN_FEE_PASSIVE_ACC numeric
@attribute 198_WITHOLDING_TAX numeric
@attribute 203_NBT_ON_CHARGES numeric
@attribute 205_POSTAGE_FEE numeric
@attribute 206_VAT_ON_POSTAGE_FEE numeric
@attribute 207_VAT_ON_AFT_FEE numeric
@attribute 208_VAT_ON_SAFE_DEPOSIT_RENT numeric
@attribute 209_NBT_AFT_FEE numeric
@attribute 214_NBT_ON_POSTAGE_FEE numeric
@attribute 215_NBT_ON_SAFE_DEPOSIT_RENT numeric
@attribute 400_AGF_AUTO_GRAB_FUND numeric
@attribute 472_MIN_BAL_ADMIN_FEE_PASSIVE_ACC numeric
@attribute 551_BATCH_SWEEP numeric
@attribute 553_BATCH_SWEEP numeric
@attribute 950_DEBIT_CHARGE_STATIONARY numeric
@attribute 960_DEBIT_VAT_STATIONARY numeric
@attribute 970_DEBIT_NBT_STATIONARY numeric

@attribute NON_ACTIVE_PERIOD numeric
@attribute ACC_AGE numeric
@attribute TOTAL_DR numeric
@attribute TOTAL_CR numeric
@attribute MAX_DR_AMOUNT numeric
@attribute MAX_CR_AMOUNT numeric
@attribute CUST_AGE numeric
@attribute churn {no,yes}

6.4 Preprocessing

The real world data is pretty dirty. It has mismatches, noise and lost values. In the absence of attribute values of incomplete data, there is no point of interest, which can lead to poor decision making or inappropriate results when using a machine learning algorithm. Data noise is the value of attributes generated when performing data entry or when a program or hardware is failing. Also may contain Inconsistent containing discrepancies in codes or names. If these problems arise, quality data for quality mining results will not produce results. Quality decision must be based on quality data. Duplicate or missing data may cause incorrect or even misleading statistics.

These observations are made using the preprocessing techniques used to make the data ideal for data mining.Data has been observed with missing values and noisy values, in addition to that redundancy data is the major problem when obtain an accuracy of the prediction result.Therefore, a number of data preprocessing methods have been used before applying a machine learning algorithm. While checking the dataset, some fields had corresponding missing values.ReplaceMissingValues option has a proper option to work with the missing values.This has used to replace all the missing values for nominal and numeric attributes in a dataset with the modes and means from the training data.

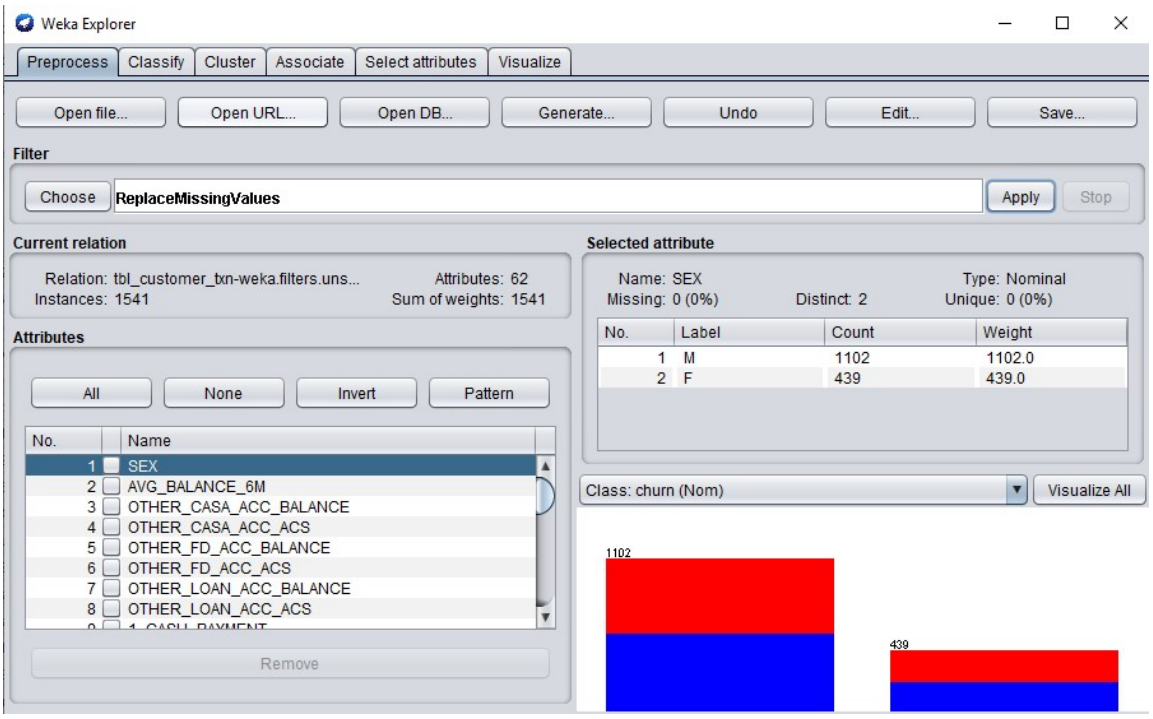


Figure 12 : Replace missing values

6.5 Feature selection

Due to the fact that the best prefabricated technical feature selection is implemented through PCA (Principle Component Analysis) dimensionality reduction, it is observed that a large number of features should have a mechanism for selecting the most efficient features.PCA can take more measurements (more dimensions of data) and make a 2-D PCA plot. This plot will show similar data cluster together.

Have some statistical distribution and what to uncover into low dimensional pattern to build models. It is well established method. It is a data driven hierarchical coordinating system. It captures the maximum amount of variance in the data. The idea here is to find if the data has statistical distribution and try to uncover dominant combination of features that describe much of the data as possible. That's PCA is going to do ellipsoid maximum variance standard deviation and quantify if arise a new point how likely is it given the distribution to old point. Singular value decomposition in the PCA will tell what direction account for the most variance second most variance, third most variance if the data so and so for. This is very useful statistical technique to best suited to attribute selection on customer transaction dataset.

With the large number of features it automatically constructs high dimensionality. What PCA does is reducing high dimensional data and reducing something that can explain in few-dimensional. Calculate the average measurement for feature A and the average measurement for feature B. With the average values calculate the center of the data. Shift the data so that the center on top of the origin in the graph. Shifting the data did not change how the data points are positioned relative to each other. The data centered on the origin that can try to fit a line to it. Draw a random line that goes through the origin. Then rotate the line until it fits the data as well it can, given that it has to go through the origin.To quantify how good this line fits the data, PCA projects the data onto it. And then it can either measure the distances from the data to the line and try to find the line that minimizes those distances or it can try to find the line that maximizes the distances from the projected points to the origin. That's how PCA is done value decomposition.

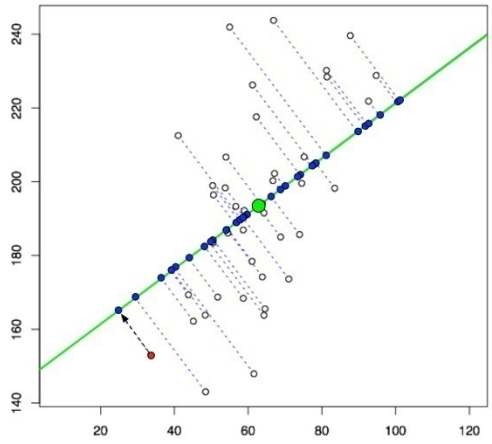


Figure 13 : PCA value decomposition

For analysis of the principle components of the customer transaction dataset, the principle component on the Select Properties panel has been selected. Then the variance coverage parameter is set to 0.95 to determine the accuracy. Leave the class label as churn. The result list as follows. When critically evaluating the correlation matrix in which the process is completed.

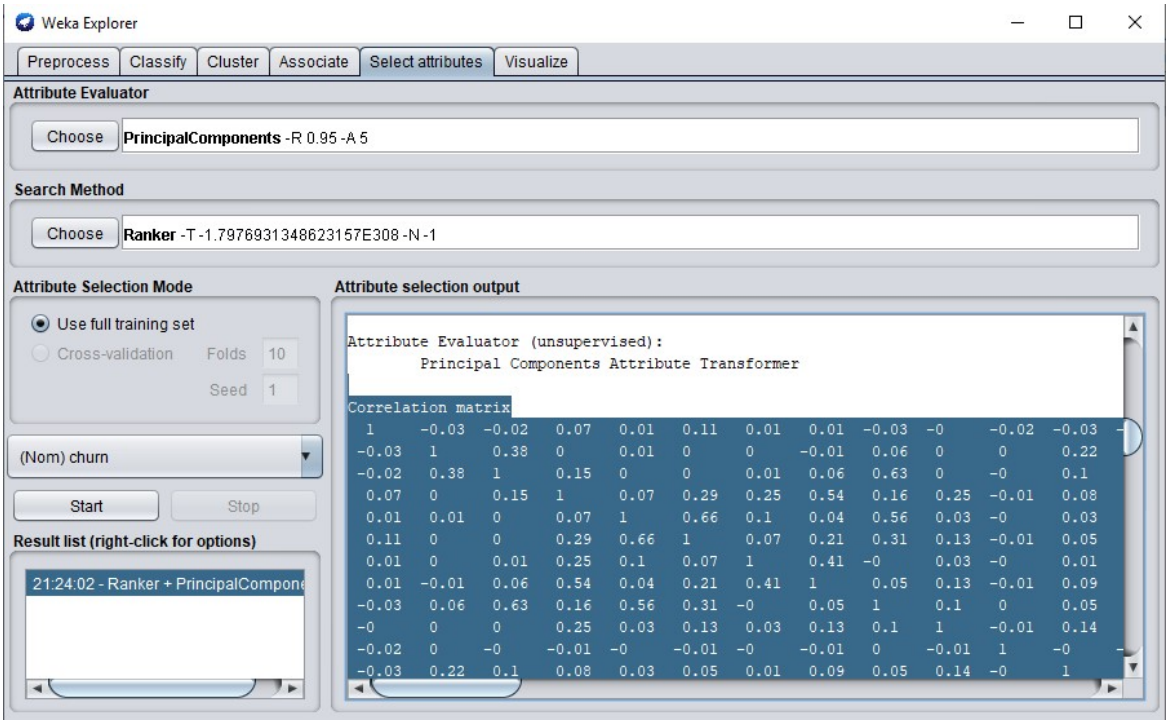


Figure 14 : WEKA Principal components evaluator

This is a scatter plot matrix using the dataset component features. The diagonal shows standard deviation. The bigger numbers more positively correlated smaller the negative numbers more negatively correlated.Measure zero completely unrelated. PCA does is results combination of features.

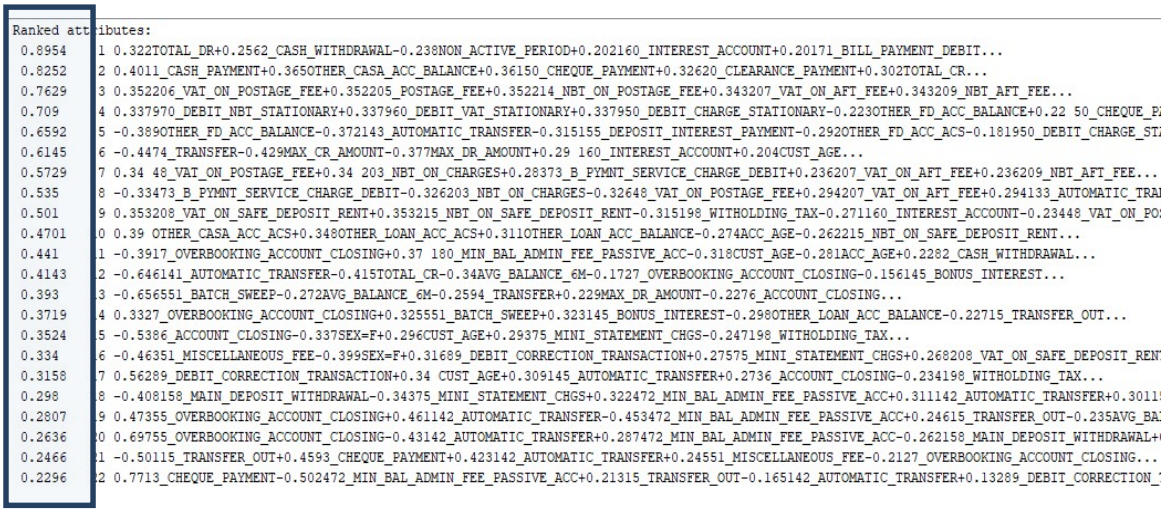


Figure 15: PCA results combination of features

According to the variance the attributes are ranked. This has to do how much of the variance it's covered. The variance in the data set is interesting stuff that has something

to effect on predictive class. What we do this from PCA is get large variances.We can end up with a set of information that has largely contributed to us by selecting the highest-order features from the big data set.Finally we can throw away a lot of things that make slower and complicated.We need to change the search methods ranking parameter to eliminate the less relevant features without limiting the removal of tool properties. We can make place a threshold value and find out what needs to be removed. To do that principle component filter has used in preprocess tab. That has resulted

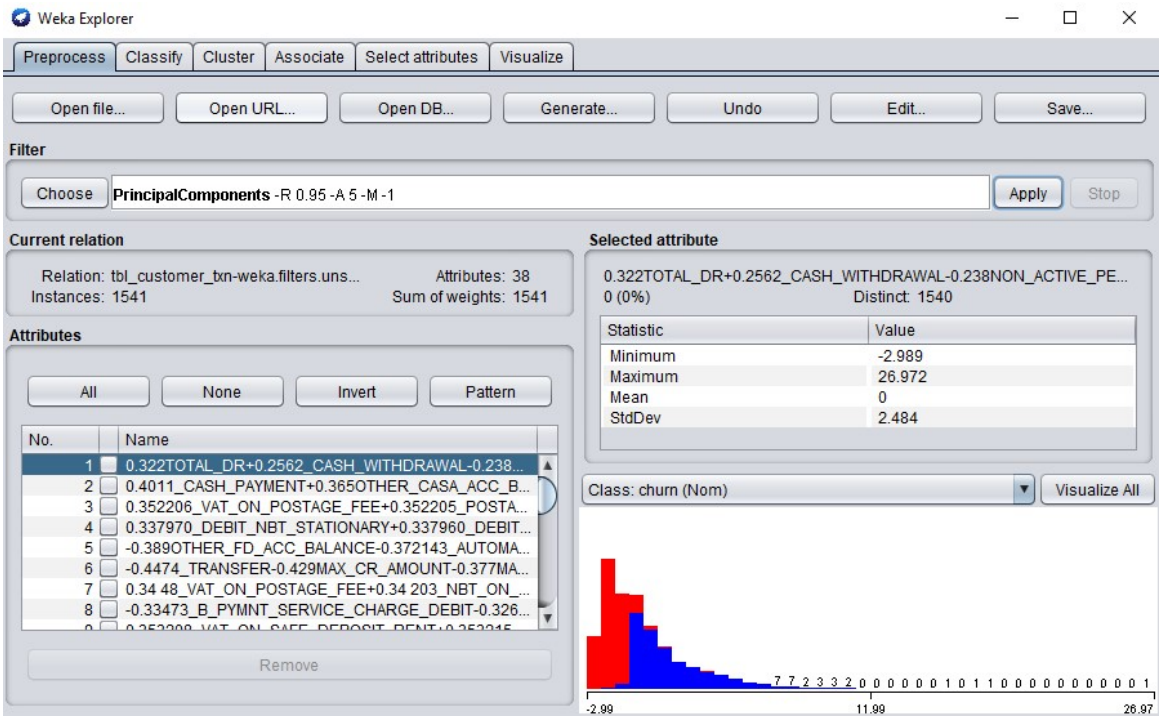


Figure 16 : PCA Resulting features

When look at these and compare the chosen attributes, the features are graded. The very top one is the best feature according to principle component analysis.That captures most of the variance and then lined up second most and so on.As the ranking reflects, we can see the variance of the new feature. The variance is useful. The first feature of tanked list is the maximumvariance. Then we can think about threshold where only keeps features that have say 50 percent of that variance to do that we take max variance

Threshold 0.5

Cut off variance = 2.484 / 2 = 1.242

That we only once keep feature variance above 1.242 these are new features which combined

Keep features whose variance > max feature variance X threshold

Here we calculate boundary feature variance against threshold. This mechanism is applied to reduce number of features, but for applying a classification model there should be test and evaluate what is the best threshold that is going to be algorithm perform well.

Threshold	Boundary Variance
0.5	1.242
0.25	0.621

Feature	StdDev
1	2.484
2	2.034
3	1.918
4	1.783
5	1.714
6	1.622
7	1.567
8	1.497
9	1.416
10	1.349
11	1.31
12	1.256
13	1.121
14	1.115
15	1.075
16	1.039
17	1.037
…….	

2.484/2=1.241 0.5 Threshold Value

Table 1 : PCA Feature Selection

After successful preprocessing the best variance feature set has been selected. The next step is applying the deferent kind of data mining algorism and find out what is the best performing for customer churn prediction domain based on customer transaction.

6.6 Selecting the Algorithms for customer churn prediction

According to the bank customer data set the prediction is two class classifications. Therefore should select best performing machine learning algorism that going to suite for two class prediction.In research done by Guoxun Wang for Predicting credit card holder churn in banks of China using data mining and MCDM[1] have used number of algorithms that includes BayesNet , Logistic , J48 , PART , NaiveBayes , RandomTree , IBK and DecisionTable. It shows the best applicability aligned with their dataset of performing well in two class similar kind of banking customer churn problem. In general, a binary classifier belongs to a normal state, and another class is called an abnormal state. In our banking customer prediction problem active customers (non-churn) are normal state closed account holders (churn) are abnormal state. These states are assigned to class label 0 and class label 1. For classification, this means that the model predicts an instance of class 1 or the probability of an abnormal situation. Logistic Regression, k-Nearest Neighbors, Decision Trees, Support Vector Machine, Naive Bayes can be considered as best performing machine learning algorithms for binary classification. Some of the algorithm has designed for only binary classifiers and Logistic Regression and Support Vector Machines are some of them[15].While going through these details BayesNet, Naïve Bayes, J48 and Support Vector have selected as machine learning algorithms for bank customer churn prediction using their transaction behavior. The training dataset is used to model by each algorithm and select the best model that is going to suite bank customer churn prediction.The collected data set is trained using different classification methods namely BayesNet, Naïve Bayes, J48, Support Vector by the help of WEKA tool.

The BayesNet, Naïve Bayes,J48, Support Vectormodelswere created using the training dataset by selecting the classifier.Summary results are saved for further critical comparison. Load the model which was saved and tested evaluate against test data set. Then check whether the accuracy level is high or not.

6.7 Summary

In previous chapter, we describe data processing, feature selection, and modeling. The feature set has been identified for dealing with a large number of dimensionality reduction techniques. The next chapter will perform a detailed evaluation of the selected machine learning model.

Chapter 7

Results and evaluation

7.1 Introduction

Chapter 6 describes the step by step implementation of all the identified elements of the proposed bank customer churn predictor.A number of models have been created. This chapter critically evaluates how selected data mining methods and data models fit into the proposed solution.

7.2Evaluation of classification techniques

We have built number of models using different classification techniques namely BayesNet ,Naïve Bayes, J48 and Support Vector. The preprocessed training dataset was used for training and the test data set was used for testing. WEKA concludes each classification and the report summary tells how it was done. While going through the summary, accuracy and confusion matrix we are going to compare and evaluate the classifiers quality. For that TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, ROC Area and PRC measurements are used. These measurements are described in below table.

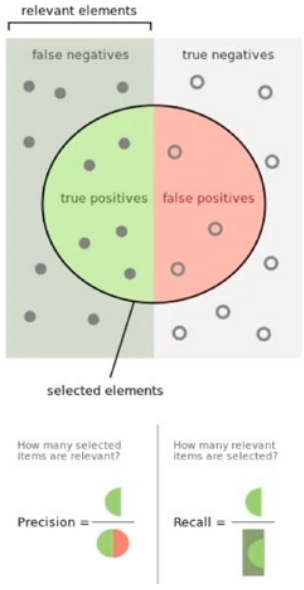


Figure 17 : Precision and recall

Measurement	Description
True positive rate (TP Rate)	The Percentage of items that correctly classified divided by total number of instance. TP is an outcome where the model correctly predicts the positive class.
False positive rate (FP Rate)	The Percentage of items that false classified divided by total number of instance. FP is an outcome where the model incorrectly predicts the positive class.
Precision	The percentage of positive identifications was actually correct. That is out of everything we classify what percentage of actually belonging there. TP / (TP + FP)
Recall	The percentage of How much of positive labeled instances belong to actually captured on predicted as positive. That is how many relevant items are selected. TP / (TP + FN)
Fmeasure	Combines precision and recall

Table 2 : Classification Evaluation Measurements

ROC Area (Receiver Operator Characteristic)tells the percentage of time correctly put into the actual class. Perfect test has the area above 0.5 usually best models have the higher value. Comparison of summary outcome of build model is listed in table.Critical evaluation of the confusion matrix calculates sensitivity and specificity, and a detailed evaluation is conducted to select the best model.

We got customer transaction and we apply machine learning methods to them to predict weather or not someone will churn or not. To do this, we have used BayesNet, Naïve Bayes, J48 and Support Vector. For deciding the work best with customer transaction data the model building was being divided into training and testing dataset. Later we have trained all the selected methods with training data and test each method on the testing set. The accuracy comparison is listed in table 8.2

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC
BayesNet	0.949	0.053	0.949	0.949	0.949	0.985
Naïve Bayes	0.74	0.228	0.81	0.74	0.73	0.95
J48	0.963	0.038	**0.963**	**0.963**	0.963	0.962
Support Vector SMO	0.963	0.037	**0.963**	**0.963**	0.963	0.963

Table 3 : Classification technique result comparison

Later on output model is evaluated on the testing data set. The evaluation results are listed in table 8.3

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC
BayesNet	0.946	0.069	0.947	0.946	0.946	0.974
Naïve Bayes	0.687	0.477	0.783	0.687	0.619	0.934
J48	0.964	0.046	0.964	0.964	0.964	0.977
Support Vector SMO	0.949	0.067	0.949	0.949	0.948	0.941

Table 4 : Evaluation on test set result comparison

According to the comparison,J48 and Support Vectorhave encountered the highest accuracy value than other selected classification techniques. Detailed summary results arelisted in appendixB.For further conclude, a confusion matrix study has been conducted. For selecting the best performing algorithm ensemble method is carried on together with model comparison via WEKA experimenter. Due to the selected J48 and SMO shows similar accuracy level Ensemble learning has been carried out

7.3 Ensemble learning

Because there was an equal accuracy result, there would be separate ensemble learning have been conducted. It often improves predictive performance by having a bunch of different machine learning algorithm method of producing classifier for the same problem allowing vote on unknown test instance. Even thoughproduce outputs are hard to analyze it results very good performance. This approach is to produce a single comprehensible structure. The methods called Bagging, Randomization and Boosting have been applied on transaction dataset then compared and summarized the output for selecting the best suited model.

There we have to produce several different decision structures. In the sense of using J48 decision trees, then had to produce slightly different decision trees. To achieve that having several different training set having same size, able to get those by sampling the original training set. In fact in bagging method samples the set with replacement which means of sometimes might get two samples to the same samples chosen in original sample. So produced several different training sets then built the model for each one for different machine learning algorithm then have combined the prediction on the different

training model by voting. This is very suitable for leaning schemas called unstable. In unstable learning schemas small change in the training data make big change in the model. On decision trees tiny little change on training data get a completely different kind of decision tree. In WEKA we get bagging classifier. Have chosen bagsize sets to 100 percent whose going to sample the training set to get another set with same size with replacement.That resulted in different sets in same size on each time we sample. But each set might contains repeats of the original training set. Then the classifier has been chosen which want to bag and number of bagging iteration. The results are collected for evaluation.

Randomize Forests here instead of randomizing the training data. How we randomize the algorithm depends on what the algorithm is. Randomize forests when using the decision trees. Attribute selection for J48 decision tree it didn't pick the best, have picked randomly from the k best options. Generally improves decision trees. The results are collected for evaluation.

Creating a model has then look at the instances that were misclassified for that model also called hard instances for the classifier. Then can put extra weight on those instances to make a training set for producing the next model in the iteration. This was a kind of encouraging new model to become an expert for instances misclassified by earlier models. A real life committee members should complement each other's expertise by focusing on different aspect of a problem. Then combined with uses voting but weights models according to their performance in boosting.For that there is a very good schema called AdaBoostM1.

Ensemble Learn	Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC
Bagging	J48	0.971	0.03	0.971	0.971	0.971	0.993
Bagging	SMO	0.961	0.04	0.961	0.961	0.961	0.978
AdaBoostM1	J48	0.971	0.03	0.971	0.971	0.971	0.993
AdaBoostM1	SMO	0.936	0.037	0.963	0.963	0.963	0.985
Random Forests		**0.977**	**0.024**	**0.977**	**0.977**	**0.977**	**0.977**

Table 5 : Ensemble learning result comparison

The comparison done in the result accuracy summary Random Forests has got higher precision and recall. It is noticed ensemble learning is improved the performance of the classifier.

7.4 Model Comparison

Model comparison was performed for check the performance of the classifier generated by multiple numbers of algorithms. To that end,the experimenter optionprovided by WEKA has used.Same data set has presented for evaluation comparison. For algorithms section we have pointed out the selected algorithms which want to compare, that is four algorithms plus ensemble learning schemas. In setup tab data set has been loaded, the experimenter was done under 10 folds cross validation.BayesNet, Naïve Bayes, J48, Support Vector and RandomForest have been loaded on algorithms section.Figure has listed in appendix A.The results of the experiment are analyzed as follows. The configure test option have set of parameters, testing width has selected to significant test paired T-Tester. To format test output, on row data set has selected. The row is going to be displayed as dataset. The column set to scheme. There are many evaluation measures in the comparison field how are going to evaluate performances of the classifiers F_Measure was set to as the option.The significance was to default 0.05 this means any different found among the classifiers generated by selected five algorithms will be 95 percent confident.Finally the test has been performed.

Simultaneously test output we have shown the details of how the performance was going to evaluate.The result has shown the dataset as tbl_customer_txn.Our first algorithm was bayesNetit was added to the selected number 1 because we can notify that it was originally selected in the Settings tab.It was our base algorithm, we were trying to compare other four algorithm performance against bayesNetclassifier.So the BayesNet has 0.95 score, NaiveBayes has 0.69 score, J48 has 0.97 score, SMO has 0.97 score and RandomForest has 0.98 score.While compered 98 percent was much higher than 95 percent in RandomForest. By statistically significant the results were compared, also that was indicating ‘v’ as victory.HerebayesNet performed 95 percent score NaiveBayes has 69 percent score it has indicated as ‘*’, that means performed 69 percent compared to 95

percent of bayesNet and that was statistically significant different as the poor result.Whereas it was compared bayesNet performance with RandomForest performance 95 percent and 98 percent respectively then we can see a ‘v’ sign that is RandomForest perform significantly better than the other classifiers.

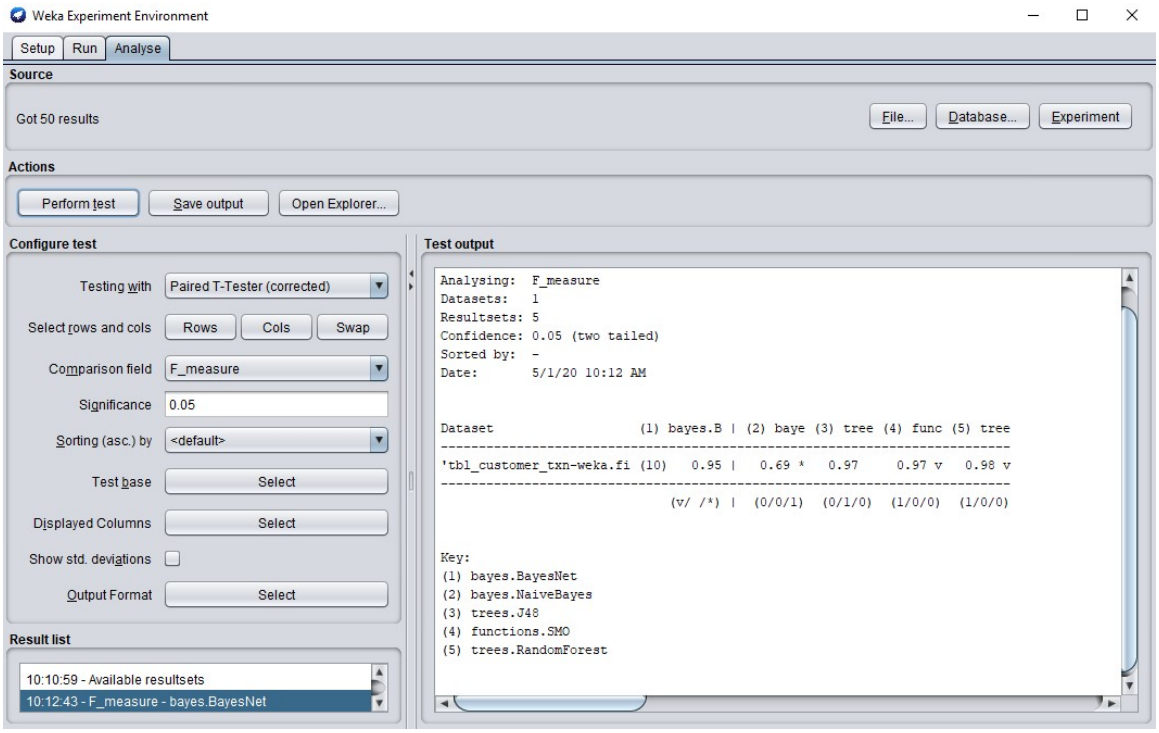


Figure 18 : WEKA experimenter test output

7.5 Discussion on model selection

While going through the model results the best performing machine learning model is going to select here. For that confusion matrix evaluation, sensitivity and specificity evaluation is considered.

7.5.1 Confusion Matrix evaluation

We have taken step by step approach to derive the classifiers into ensemble modeling,to check whether what will be the best suited algorithm. Now we are going to evaluate those results by means of confusion matrix. To verify the performance of each method, the result of modeling is summarized in the information listed in the appendix

B.Comparison was carried with the confusion matrix for each method. The rows in confusion matrix correspond to what the machine learning algorithm predicted and the columns correspond to the known truth.Since there are only two categories to choose from according to customer transaction behavior is “churn customer” or “non-churn customer”.These are that customers that have non-churned that were correctly identified by the algorithm. True negatives are in the bottom right-hand corner. These are the customers that did churn that were correctly identified by the algorithm. The bottom left-hand corner contains the false negatives. The top right-hand corner contains the false positives. False positives are customers that non-churn but algorithm says they are churned.

When we applied bagging to the testing set there were 606 true positive with non-customer churn were correctly classified. And 511 true negatives customers with churned that were correctly classified. However the algorithm misclassified 21 customers that did not-churn by saying that did churn and the algorithm misclassified 12 customers that did churn by saying did not churn. The numbers along the diagonal true positive to true negative tell us how many times the samples were correctly classified. The numbers not on the diagonal are samples the algorithm messed up. Here we can compare the Bagging confusion matrix to the confusion matrix we get when we use AdaBoostM1. In AdaBoostM1 there were 595 true positive with non-customer churn were correctly classified. And 512 true negatives customers with churned that were correctly classified. However the algorithm misclassified 20 customers that did not-churn by saying that did churn and the algorithm misclassified 23 customers that did churn by saying did not churn. While observing the numbersBagging was worse than the AdaBoostM1 at predicting churn customers. So next we are going to compare the RandomForest confusion matrix there were 609 true positive with non-customer churn were correctly classified. And 515 true negatives customers with churned that were correctly classified. However the algorithm misclassified 17 customers that did not-churn by saying that did churn and the algorithm misclassified 9 customers that did churn by saying did not churn. These two confusion matrixes are very similar and make it hast to choose which machine learning method is a better fit for this data.Furthermore by considering more sophisticated metrics by calculating the sensitivity and specificity the evaluated methods

are going to distinguish how they will be performed well into customer churn prediction for early identify churning customer in banking domain.

7.5.2 Sensitivity and Specificity evaluation

This section describes calculating and interpreting Sensitivity and Specificity of the best performed confusion matrix. In confusion matrix rows correspond to what was predicted the columns correspond to the known truth. There are only two categories to choose from this scenario the two choices were “churn customer” and “Does not churn customer”. The top left-hand corner contains the true positives. The true positives are customers that non-churn that were also predicted to non-churn customers. True negatives are bottom right-hand corner True negatives are customers that have churn and were predicted to churn customers. The bottom left-hand corner contains false negatives. False negatives are when a customer non-churn but the prediction said they churned. Lastly the top right-hand corner contains the false positives. False positives are customers that churned, but the prediction says that they don’t. Once we’ve filled out the confusion matrix, we can calculate two useful metrics Sensitivity and Specificity. In this case, Sensitivity tells us what percentage of customers with non-churned were correctly identified. That’s true positives divided by the sum of the true positives and the false negatives.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Specificity tells what percentage of churn-customers were correctly identified. That is true negatives divided by sum of the true negative and the false positives.

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

According the observation RandomForest and AdaBoostM1 has ranked into higher performance on the customer prediction dataset. Furthermore based on the confusion matrix now is discussing the Sensitivity and Specificity to finding out what classifier is best suited for customer churn prediction scenario.

Calculate Sensitivity and Specificity for RandomForest

```
=== Confusion Matrix ===

   a   b   <-- classified as
 609   9 |   a = no
  17 515 |   b = yes
```

$$\text{Sensitivity} = \frac{609}{609 + 17} = 97\%$$

$$\text{Specificity} = \frac{515}{515 + 9} = 98\%$$

Sensitivity tells that 97 % of the non-churn customers were correctly identified, and 98 % churn customers were correctly identified by the RandomForest model.

Calculate Sensitivity and Specificity for AdaBoostM1

```
=== Confusion Matrix ===

   a   b   <-- classified as
 606  12 |   a = no
  21 511 |   b = yes
```

$$\text{Sensitivity} = \frac{606}{606+21} = 96\ \%$$

$$\text{Specificity} = \frac{511}{511+12} = 97\ \%$$

Sensitivity tells that 96 % of the non-churn customers were correctly identified, and 97 % churn customers were correctly identified by the AdaBoostM1 model. Now we can compare the sensitivity and specificity values that we calculated for RandomForest to the values we calculated for the AdaBoostM1.Sensitivity tells us that the RandomForest is slightly better at identifying positives. Which is in the case non-churn

customers.Specificity tells us RandomForest is slightly better at identifying negatives. Which in this case are churned-customers. We would choose the RandomForest model if correctly identifying churned-customers was more important than correctly identifying non-churn customers.

7.6 Summary

In this chapter, we critically analyze model performance in different perspectives.It enables us to deduce the best performance model for predicting the customer churn of our chosen problem domain. In the next chapter we will discuss the limitations and direction of further work to improve the identified findings.

Chapter 8

Conclusion and Further Work

8.1 Introduction

Throughout this research,the problem of banking customer churn forecasts has been solved by using customer transaction behavior. We have successfully modeled the bank customer churn predictor. Using the Transaction Behavior Dataset, the best performed features are used for modeling mechanism. According to a series of evaluation steps, RandomForest is ranked as the most appropriate classification in this research domain.

8.2 Bank Churn Predictor Conclusion

In this research problem the main concern was to find out a mechanism to determine if the customer is going to leave the bank with the help of his transaction behavior.In order to that the main objectiveof this research was to develop a machine learning model forearly identification of potential bank churn customers.While going through the methodology roadmap we have divided the project into two parts. First one is feature selection and at this stage, the best features from the feature set are discovered, while extracting the dataset of customer behavior. It was the most critical stage because of prediction model performance and accuracy was depends on the selected feature set. With the PCA feature selection algorithm, new features are extracted from the feature combination. Subsequently, by defining a threshold value high-variability features are filtered. Next by series of machine learning algorithm RandomForest is used to model bank churn customers prediction model. In the conclusion we have demonstrated RandomForest has shown the quality and accuracy with selected bank customer transaction behavior dataset. The main contributions can be listed as follows.

- Data has been preprocessed, which have effectively contributed to model building.
- While comparison of machine learning techniques. Classification techniques has selected as best approach of this two class prediction domain of bank customer churn prediction.

- By evaluation of various selected classifiersRandomForest has shown best approach to solve build the model of bank customer churn predictor.
- The usage of abstract knowledge to retain organization customers.

9.2 Limitations

Although the human behavior is an unpredictable thing, by considering past transaction behavior we have archived the predict customer churn. This model was created using transaction behavior, but can cause a wide range of matters beyond the scope of the transaction due to the churning of customers. The new customer will churn the bank but churn model behaves using the historical data in the time of model was created.

9.2 Future Developments

As future work,we can integrate global economy, geographical conditions, cross bank transaction iterations in to the features there forcan increase the model beyond the transactional behavior. In addition, customer behavior of another banking instances, and the behavior that family members often deal with adds value. While going through with the big-data concepts the usage of real-time training data for dynamic model building and integrate with GUI based prediction visualization mechanism will be added value for banking sector as well.

9.4 Summary

This chapter concludes the thesis by describing a solution as a form of data mining for predicting bank customer churn and the advantages of using a predictive model to retain customers and early identify the customers who are going to churn form the bank.

References

[1].Guoxun Wang et al., “Predicting credit card holder churn in banks of China using data mining and MCDM” in International Conference on Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM 2010 , P. 216

[2].JussiAhola and EsaRinta-Runsala, “Data Mining Case Studies in Customer Profiling” VTT Information Technology Research Report TIE1-2001-29

[3]. Peters, Edward M.L.et al, “Understanding Service Quality and Customer Churn by Process Discovery for a Multi-National Banking Contact Center.” 2013 IEEE 13th International Conference on Data Mining Workshops

[4]. Shaoying Cui and Ning Ding “Customer Churn Prediction Using Improved FCM Algorithm” 3rd International Conference on Information Management 2017, p.112.

[5].ShrishaBharadwaj et al.,“Customer Churn Prediction in Mobile Networks using Logistic Regression and Multilayer Perceptron(MLP)” Second International Conference on Green Computing and Internet of Things (ICGCIoT) 2018 Available:https://ieeexplore.ieee.org/document/8752982

[6].Wojewnik P, Kaminski B, Zawisza M, et al. Social-Network Influence on Telecommunication Customer Attrition 2011

[7]. Zhao Jing and Dang Xing-hua, “Bank Customer Churn Prediction Based on Support Vector Machine” 2008 4th International Conference on Wireless Communications, Networking and Mobile Computing.

[8]. “Fintech used to be a back-office support function, now it's defining an industry” [Online]. Available:https://www.investopedia.com/the-future-of-fintech-4770491#:~:targetText=Over%20the%20last%20decade%2C%20private,place%20in%20the%20innovation%20economy. [Accessed: 20-Oct-2019].

[9].“Growth of Digital Banking Market Size & Share” [Online]. Available:https://www.zionmarketresearch.com/report/digital-banking-market [Accessed: 12-Aug-2019].

[10]. “Acquisition vs Retention: The Importance of Customer Lifetime Value” [Online]. Available:https://www.huify.com/blog/acquisition-vs-retention-customer-lifetime-value[Accessed: 05-Oct-2019].

[11]. “Sri Lankan Customers’ Behavioural Intention to Use Mobile Banking” Available:http://www.seu.ac.lk/jisit/publication/v2n2/paper1.pdf [Accessed: 14-Sep-2019].

[12]. “Guidelines on opening of new banks in Sri Lanka” Available:https://www.cbsl.gov.lk/sites/default/files/cbslweb_documents/laws/banks.pdf[Accessed: 03-Aug-2019]

[13]. “A Bank Clarity Report” Available: https://www.abrigo.com/blog/2018/08/01/why-do-people-switch-banks/ [Accessed: 12-Aug-2019]

[14]. Comparison of Data Mining Classification Algorithms Determining the Default Risk Available: https://www.hindawi.com/journals/sp/2019/8706505/ [Accessed 5/12/2019]

[15].Types of Classification Tasks in Machine Learning. Available: https://machinelearningmastery.com/types-of-classification-in-machine-learning/ [Accessed 03/01/2020]

[16]. How Neural Network Algorithms Works, Available: https://vinodsblog.com/2018/12/31/how-neural-network-algorithms-works-an-overview/]. [Accessed 03/01/2020]

[17]. Understanding Support Vector Machine, Available:https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ [Accessed 05/02/2020]

[18].Mining Tools, Available: https://www.softwaretestinghelp.com/data-mining-tools/ [Accessed 05/02/2020]

[19].Orange Data Mining, Available: https://orange.biolab.si/ [Accessed 05/02/2020]

Appendix - A

```
=== Summary ===

Correctly Classified Instances        1091               94.8696 %
Incorrectly Classified Instances        59                5.1304 %
Kappa statistic                          0.8967
Mean absolute error                      0.0544
Root mean squared error                  0.2162
Relative absolute error                 10.9446 %
Root relative squared error             43.366  %
Total Number of Instances             1150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.960    0.064    0.946      0.960    0.953      0.897    0.985     0.986     no
                 0.936    0.040    0.952      0.936    0.944      0.897    0.985     0.986     yes
Weighted Avg.    0.949    0.053    0.949      0.949    0.949      0.897    0.985     0.986

=== Confusion Matrix ===

   a   b   <-- classified as
 593  25 |   a = no
  34 498 |   b = yes
```

Appendix A :1 BayesNet Model Accuracy

```
=== Summary ===

Correctly Classified Instances         369               94.6154 %
Incorrectly Classified Instances        21                5.3846 %
Kappa statistic                          0.8862
Mean absolute error                      0.0541
Root mean squared error                  0.2198
Total Number of Instances              390

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.903    0.025    0.959      0.903    0.930      0.887    0.974     0.970     no
                 0.975    0.097    0.939      0.975    0.956      0.887    0.974     0.979     yes
Weighted Avg.    0.946    0.069    0.947      0.946    0.946      0.887    0.974     0.976

=== Confusion Matrix ===

   a   b   <-- classified as
 139  15 |   a = no
   6 230 |   b = yes
```

Appendix A :2 BayesNet Model Evaluation on test set

```
=== Summary ===

Correctly Classified Instances         851               74      %
Incorrectly Classified Instances       299               26      %
Kappa statistic                          0.4946
Mean absolute error                      0.2604
Root mean squared error                  0.4976
Relative absolute error                 52.3648 %
Root relative squared error             99.7999 %
Total Number of Instances             1150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.544    0.032    0.952      0.544    0.692      0.553    0.950     0.939     no
                 0.968    0.456    0.646      0.968    0.775      0.553    0.949     0.956     yes
Weighted Avg.    0.740    0.228    0.810      0.740    0.730      0.553    0.950     0.947

=== Confusion Matrix ===

   a   b   <-- classified as
 336 282 |   a = no
  17 515 |   b = yes
```

Appendix A :3 Naïve Bayes Model Accuracy

```
=== Summary ===

Correctly Classified Instances         268               68.7179 %
Incorrectly Classified Instances       122               31.2821 %
Kappa statistic                          0.2429
Mean absolute error                      0.3078
Root mean squared error                  0.5491
Relative absolute error                 60.6019 %
Root relative squared error            107.8397 %
Total Number of Instances              390

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.214    0.004    0.971      0.214    0.351      0.364    0.934     0.913     no
                 0.996    0.786    0.660      0.996    0.794      0.364    0.934     0.949     yes
Weighted Avg.    0.687    0.477    0.783      0.687    0.619      0.364    0.934     0.935

=== Confusion Matrix ===

   a   b   <-- classified as
  33 121 |   a = no
   1 235 |   b = yes
```

Appendix A : 4 Naïve Bayes Model Evaluation on test set

```
=== Summary ===

Correctly Classified Instances        1107               96.2609 %
Incorrectly Classified Instances        43                3.7391 %
Kappa statistic                          0.9248
Mean absolute error                      0.0429
Root mean squared error                  0.191
Relative absolute error                  8.6229 %
Root relative squared error             38.3008 %
Total Number of Instances             1150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.964    0.039    0.966      0.964    0.965      0.925    0.962     0.966     no
                 0.961    0.036    0.959      0.961    0.960      0.925    0.962     0.915     yes
Weighted Avg.    0.963    0.038    0.963      0.963    0.963      0.925    0.962     0.942

=== Confusion Matrix ===

   a   b   <-- classified as
 596  22 |   a = no
  21 511 |   b = yes
```

Appendix A : 5 J48 Model Accuracy

```
=== Summary ===

Correctly Classified Instances         376               96.4103 %
Incorrectly Classified Instances        14                3.5897 %
Kappa statistic                          0.9244
Mean absolute error                      0.0483
Root mean squared error                  0.1852
Relative absolute error                  9.5026 %
Root relative squared error             36.371  %
Total Number of Instances              390

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.935    0.017    0.973      0.935    0.954      0.925    0.977     0.965     no
                 0.983    0.065    0.959      0.983    0.971      0.925    0.977     0.973     yes
Weighted Avg.    0.964    0.046    0.964      0.964    0.964      0.925    0.977     0.970

=== Confusion Matrix ===

   a   b   <-- classified as
 144  10 |   a = no
   4 232 |   b = yes
```

Appendix A : 6 J48Model Test Evaluation on test set

```
=== Summary ===

Correctly Classified Instances        1108               96.3478 %
Incorrectly Classified Instances        42                3.6522 %
Kappa statistic                          0.9266
Mean absolute error                      0.0365
Root mean squared error                  0.1911
Relative absolute error                  7.3453 %
Root relative squared error             38.3284 %
Total Number of Instances             1150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.964    0.038    0.968      0.964    0.966      0.927    0.963     0.952     no
                 0.962    0.036    0.959      0.962    0.961      0.927    0.963     0.940     yes
Weighted Avg.    0.963    0.037    0.963      0.963    0.963      0.927    0.963     0.947

=== Confusion Matrix ===

   a   b   <-- classified as
 596  22 |   a = no
  20 512 |   b = yes
```

Appendix A : 7 Support Vector SMO Model Accuracy

```
=== Summary ===

Correctly Classified Instances         370               94.8718 %
Incorrectly Classified Instances        20                5.1282 %
Kappa statistic                          0.8915
Mean absolute error                      0.0513
Root mean squared error                  0.2265
Relative absolute error                 10.0979 %
Root relative squared error             44.4765 %
Total Number of Instances              390

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.903    0.021    0.965      0.903    0.933      0.893    0.941     0.910     no
                 0.979    0.097    0.939      0.979    0.959      0.893    0.941     0.932     yes
Weighted Avg.    0.949    0.067    0.949      0.949    0.948      0.893    0.941     0.923

=== Confusion Matrix ===

   a   b   <-- classified as
 139  15 |   a = no
   5 231 |   b = yes
```

Appendix A : 8 Support Vector SMO Evaluation on test set

```
=== Summary ===

Correctly Classified Instances        1124               97.7391 %
Incorrectly Classified Instances        26                2.2609 %
Kappa statistic                          0.9545
Mean absolute error                      0.0463
Root mean squared error                  0.1363
Relative absolute error                  9.3198 %
Root relative squared error             27.3279 %
Total Number of Instances             1150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.985    0.032    0.973      0.985    0.979      0.955    0.997     0.998     no
                 0.968    0.015    0.983      0.968    0.975      0.955    0.997     0.997     yes
Weighted Avg.    0.977    0.024    0.977      0.977    0.977      0.955    0.997     0.997

=== Confusion Matrix ===

   a   b   <-- classified as
 609   9 |   a = no
  17 515 |   b = yes
```

Appendix A :9 RandomForest Model Accuracy

```
=== Summary ===

Correctly Classified Instances         380               97.4359 %
Incorrectly Classified Instances        10                2.5641 %
Kappa statistic                          0.9461
Mean absolute error                      0.0475
Root mean squared error                  0.1551
Relative absolute error                  9.3608 %
Root relative squared error             30.4552 %
Total Number of Instances              390

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.955    0.013    0.980      0.955    0.967      0.946    0.990     0.988     no
                 0.987    0.045    0.971      0.987    0.979      0.946    0.990     0.990     yes
Weighted Avg.    0.974    0.033    0.974      0.974    0.974      0.946    0.990     0.989

=== Confusion Matrix ===

   a   b   <-- classified as
 147   7 |   a = no
   3 233 |   b = yes
```

Appendix A :10 RandomForest Model Evaluation on test set

```
=== Summary ===

Correctly Classified Instances        1110               96.5217 %
Incorrectly Classified Instances        40                3.4783 %
Kappa statistic                          0.93
Mean absolute error                      0.0559
Root mean squared error                  0.1628
Relative absolute error                 11.2334 %
Root relative squared error             32.6496 %
Total Number of Instances             1150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.974    0.045    0.962      0.974    0.968      0.930    0.990     0.988     no
                 0.955    0.026    0.969      0.955    0.962      0.930    0.990     0.988     yes
Weighted Avg.    0.965    0.036    0.965      0.965    0.965      0.930    0.990     0.988

=== Confusion Matrix ===

   a   b   <-- classified as
 602  16 |   a = no
  24 508 |   b = yes
```

Appendix A :11 Bagging Model Accuracy

```
=== Summary ===

Correctly Classified Instances         375               96.1538 %
Incorrectly Classified Instances        15                3.8462 %
Kappa statistic                          0.9187
Mean absolute error                      0.0573
Root mean squared error                  0.1723
Relative absolute error                 11.276  %
Root relative squared error             33.8427 %
Total Number of Instances              390

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.922    0.013    0.979      0.922    0.950      0.920    0.977     0.975     no
                 0.987    0.078    0.951      0.987    0.969      0.920    0.977     0.971     yes
Weighted Avg.    0.962    0.052    0.962      0.962    0.961      0.920    0.977     0.973

=== Confusion Matrix ===

   a   b   <-- classified as
 142  12 |   a = no
   3 233 |   b = yes
```

Appendix A :12 Bagging Model Evaluation on test set

Appendix - B

```
=== Summary ===

Correctly Classified Instances        1117               97.1304 %
Incorrectly Classified Instances        33                2.8696 %
Kappa statistic                          0.9422
Mean absolute error                      0.043
Root mean squared error                  0.1443
Relative absolute error                  8.652  %
Root relative squared error             28.9504 %
Total Number of Instances             1150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.981    0.039    0.967      0.981    0.973      0.942    0.993     0.991     no
                 0.961    0.019    0.977      0.961    0.969      0.942    0.993     0.992     yes
Weighted Avg.    0.971    0.030    0.971      0.971    0.971      0.942    0.993     0.992

=== Confusion Matrix ===

   a   b   <-- classified as
 606  12 |   a = no
  21 511 |   b = yes
```

Appendix B :1 Bagging J48

```
=== Summary ===

Correctly Classified Instances        1105               96.087  %
Incorrectly Classified Instances        45                3.913  %
Kappa statistic                          0.9213
Mean absolute error                      0.041
Root mean squared error                  0.1795
Relative absolute error                  8.2547 %
Root relative squared error             35.9941 %
Total Number of Instances             1150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.964    0.043    0.963      0.964    0.964      0.921    0.978     0.975     no
                 0.957    0.036    0.959      0.957    0.958      0.921    0.978     0.961     yes
Weighted Avg.    0.961    0.040    0.961      0.961    0.961      0.921    0.978     0.969

=== Confusion Matrix ===

   a   b   <-- classified as
 596  22 |   a = no
  23 509 |   b = yes
```

Appendix B :2 Bagging SMO

```
=== Summary ===

Correctly Classified Instances        1124               97.7391 %
Incorrectly Classified Instances        26                2.2609 %
Kappa statistic                          0.9545
Mean absolute error                      0.0463
Root mean squared error                  0.1363
Relative absolute error                  9.3198 %
Root relative squared error             27.3279 %
Total Number of Instances             1150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.985    0.032    0.973      0.985    0.979      0.955    0.997     0.998     no
                 0.968    0.015    0.983      0.968    0.975      0.955    0.997     0.997     yes
Weighted Avg.    0.977    0.024    0.977      0.977    0.977      0.955    0.997     0.997

=== Confusion Matrix ===

   a   b   <-- classified as
 609   9 |   a = no
  17 515 |   b = yes
```

Appendix B :3 Randomization

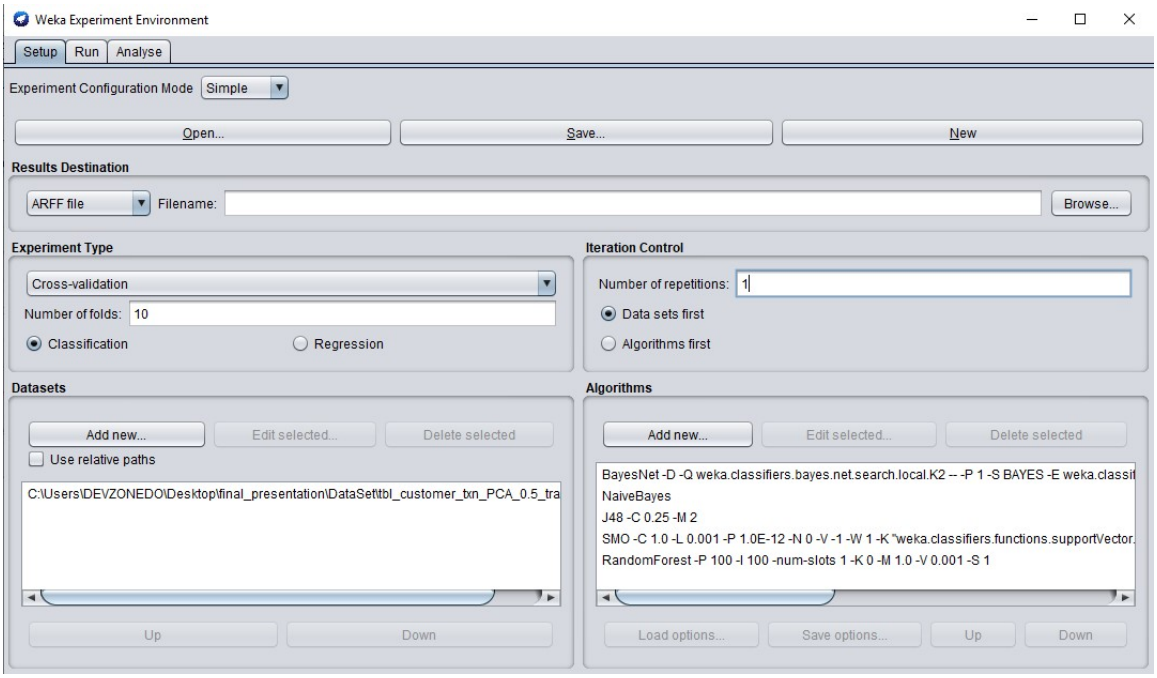
```
=== Summary ===

Correctly Classified Instances        1117               97.1304 %
Incorrectly Classified Instances        33                2.8696 %
Kappa statistic                          0.9422
Mean absolute error                      0.043
Root mean squared error                  0.1443
Relative absolute error                  8.652  %
Root relative squared error             28.9504 %
Total Number of Instances             1150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.981    0.039    0.967      0.981    0.973      0.942    0.993     0.991     no
                 0.961    0.019    0.977      0.961    0.969      0.942    0.993     0.992     yes
Weighted Avg.    0.971    0.030    0.971      0.971    0.971      0.942    0.993     0.992

=== Confusion Matrix ===

   a   b   <-- classified as
 606  12 |   a = no
  21 511 |   b = yes
```

Appendix B :4 AdaBoostM1 for classifier J48

```
=== Summary ===

Correctly Classified Instances        1107               96.2609 %
Incorrectly Classified Instances        43                3.7391 %
Kappa statistic                          0.9248
Mean absolute error                      0.0498
Root mean squared error                  0.1849
Relative absolute error                 10.0231 %
Root relative squared error             37.0756 %
Total Number of Instances             1150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.963    0.038    0.967      0.963    0.965      0.925    0.985     0.986     no
                 0.962    0.037    0.957      0.962    0.960      0.925    0.985     0.980     yes
Weighted Avg.    0.963    0.037    0.963      0.963    0.963      0.925    0.985     0.983

=== Confusion Matrix ===

   a   b   <-- classified as
 595  23 |   a = no
  20 512 |   b = yes
```

Appendix B : 5 AdaBoostM1 for classifier SMO

Appendix - C



Appendix B :6 WEKA experimenter