2020

Customer Churn Prediction

Deepshikha Wadikar
Technological University Dublin

Recommended Citation

Wadikar, D. (2020). *Customer churn prediction*. Masters Dissertation. Technological University Dublin. DOI:10.21427/kpsz-x829

Customer Churn Prediction

Deepshikha Wadikar

D17128916

A dissertation submitted in partial fulfilment of the requirements of
Technological University Dublin for the degree of
M.Sc. in Computer Science (Data Analytics)

Jan 2020

DECLARATION

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Technological University Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: **Deepshikha Wadikar**

Date: **05 January 2020**

ABSTRACT

Churned customers identification plays an essential role for the functioning and growth of any business. Identification of churned customers can help the business to know the reasons for the churn and they can plan their market strategies accordingly to enhance the growth of a business. This research is aimed at developing a machine learning model that can precisely predict the churned customers from the total customers of a Credit Union financial institution.

A quantitative and deductive research strategies are employed to build a supervised machine learning model that addresses the class imbalance problem handled feature selection and efficiently predict the customer churn. The overall accuracy of the model, Receiver Operating Characteristic curve and Area Under the Receiver Operating Characteristic Curve is used as the evaluation metrics for this research to identify the best classifier.

A comparative study on the most popular supervised machine learning methods – Logistic Regression, Random Forest, Support Vector Machine (SVM) and Neural Network were applied to customer churning prediction in a CU context. In the first phase of our experiments, the various feature selection techniques were studied. In the second phase of our study, all models were applied on the imbalance dataset and results were evaluated. SMOTE technique is used to balance the data and then the same models were applied on the balanced dataset and results were evaluated and compared. The best over-all classifier was Random Forest with accuracy almost 97%, precision 91% and recall as 98%.

Key words: *Credit Union, Churn Prediction, Supervised Machine Learning, Classification, Sampling, Feature Selection.*

ACKNOWLEDGEMENTS

I would first like to express my sincere thanks to my supervisor **Prof. Vincent McGrady** for providing me with the data for my research. His immense knowledge, continuous support, guidance and advice throughout the project helped me and encouraged me a lot to do better in my thesis writing. You are an amazing mentor and without your support, this thesis would not have been possible.

I would also like to thank **DIT** and **Prof. Luca Longo**, M.Sc. thesis coordinator, for providing me with the opportunity to work on this thesis.

Finally, I would like to thank all my friends and family for all their encouragement, support and motivation during my studies. Special gratitude to my parents **Pradeep** and **Chetna**, and my husband **Nitin** for their love, support and encouragement throughout my studies. This accomplishment would not have been possible without them.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

LIST OF ACRONYMS

CU Credit Union

CRISP-DM Cross Industry Standard Process for Data Mining

BOI Bank Of Ireland

AIB Allied Irish Bank

ILCU Irish League of Credit Union

SVM Support Vector Machine

CRM Customer Relationship Management

SMOTE Synthetic Minority Oversampling technique

AUC Area Under Curve

ROC Receiver Operating Curve

ANN Artificial Neural Network

SOM Self Organizing Map

DT Decision Tree

MLP Multi Layer Perceptron

TDL Top Decile Lift

EDA Exploratory Data Analysis

RBF Radial Basis Function

RELU Rectified Linear Unit

TP True Positive

FP False Positive

TN True Negative

FN False Negative

TPR True Positive Rate

FPR False Positive Rate

1. INTRODUCTION

1.1 Background

A Credit Union (CU) is a non-profit organisation which exists to serve their members in Ireland since 1958. They have more than 3.6 million members in Ireland. CUs functions the same as banks, they accept deposits, provide loans at a reasonable rate of interest and offer a wide variety of financial services. A CU is a group of people connected by a 'common bond' based on the area they live in, the occupation, or the employer they work for, who can save together and lend to each other at a fair and reasonable rate of interest.[1] There is CU present based on geographical areas. In every area, there is one CU present for its members.

CU is different from the banks (BOI, AIB, Ulster bank) in many ways –

1) CU is a not-for-profit democratic financial institute owned by its members whereas banks are profit-earning financial institutes.

2) Any surplus income is either distributed amongst members in the form of dividends or is used to develop new and existing services.

3) No hidden administration or transaction fees for the members.

4) Loans and Savings are insured at no direct cost.

5) Flexibilities are offered to the members regarding loan repayments.

6) CU is committed to their local communities and provides support to local youth initiatives, charities, sporting clubs, and cultural events.

7) Banks are ahead of CU in terms of the number of employees working. The number of employees in BOI is 11,086; in AIB 10,500 whereas in CU there are 3,500 employees.

[1] https://www.creditunion.ie/about-credit-unions/what-is-a-credit-union/

The Irish League of Credit Union (ILCU) describes CU as “a group of people who save together and lend to each other at a fair and reasonable rate of interest”. CU offers their members the chance to have control over their finances. Regular savings form a common pool of money, which provides many benefits for members.

With advancements and competition amongst financial institutions, there is a need to retain their old customers. Customer retention is crucial in a variety of businesses as acquiring new customers is often more costly than keeping the current ones (Kaya, Dong, Suhara, Balsicoy & Bozkaya, 2018). Customer Churn has become a major problem in all industries including the banking industry and banks have always tried to track customer interaction so that they can detect the customers who are likely to leave the bank. Customer Churn modeling is mainly focusing on those customers who are likely to leave and so that they can take the necessary steps to prevent churn (Oyeniyi & Adeyemo 2015). For CUs customer churn is important as getting new members is expensive. Moreover, to join the CU the member must satisfy the common bond criteria, a common bond of either within a community (geographical), or industrial (employment).

The ILCU has an affiliated membership of 351 CUs – 259 in the Republic of Ireland and 92 in Northern Ireland. In this research, we are using the member/customer data of one of these CUs to predict customer churn.

1.2 Research Project

Supervised machine learning techniques have been used in customer churn prediction problems in the past with SVM-POLY using AdaBoost as the best overall model (Vafeiadis, Diamantaras, Chatzisavvas & Sarigiannidis, 2015). The most common techniques applied for predicting customer churn are Decision tree, Multilayer perceptron, and SVM.

In existing research of customer churn prediction problem in telecommunication industry the researcher Guo-en, X., have used SVM model as it can solve the nonlinearity, high dimension, and local minimization problems. The model prediction depends on the data structure and condition.

Techniques that are most commonly used to predict customer churn are neural networks, support vector machines and logistic regression models. Data mining research literature suggests that machine learning techniques, such as neural networks should be used for non-parametric datasets because they often outperform traditional statistical techniques such as linear and quadratic discriminant analysis approaches (Zoric, 2016).

Logistic Regression is a type of probability statistical classification model mainly used for classification problems (Nie, Rowe, Zhang, Tian & Shi, 2011). The technique can work well with a different combination of variables and can help in predicting the customer churn with higher accuracy.

Random Forest is an ensemble learning method for classification, regression problems and uses the bagging technique to generate the results. The default hyperparameters of Random Forest gives good results and it is great at avoiding overfitting (Pretorius, Bierman & Steel, 2016).

Based on the previous literature in this area and for reasons mentioned further on in this section, four supervised machine learning techniques will be compared when aiming to predict customer churn, the four techniques are logistic regression, random forest, SVM and neural network.

Currently, the customer churn is not predicted using any of the machine learning algorithm techniques for CU members' data. The Logistic regression model is selected and in the previous research, it has been observed that SVM and random forest outperformed logistic regression when predicting customer churn.

The research question is framed as:

"Which supervised machine learning: Logistic regression, Random forest, SVM or Neural network; can best predict the customer churn of CU with the best accuracy, specificity, precision, and recall?"

1.3 Research Objectives

The key objective of the research is to identify whether the Supervised Machine Learning will help to predict the customer churn rate on CU customer data precisely. Currently, no specific method has been adopted by CU to identify the customer churn rate. This research help identify the customers which are more likely to churn and then

in turn the customers can focus more on those customers and thus can retain their old customers which leads to the growth.

The research objectives are as follows–

1) To collect the required customer data from the business for the research.
2) Understanding the data, identifying any data issues and then rectifying those to apply machine learning algorithms.
3) Preparing the data using sampling, encoding, feature selection and splitting the data.
4) Building the supervised machine learning models – Support Vector Machine, Logistic Regression, Random Forest and Neural Network to see the performance on training data set.
5) Validating the models on the Validation data set and based on evaluation metrics identifying the best model among all for predicting the customer churn.
6) Then testing the best performance model amongst all supervised models on the Test data set and then evaluating the results.
7) Identify the limitation and future research propose areas.

1.4 Research Methodologies

The Research can be classified based on different ways –

1.4.1 Based on type: Primary Vs. Secondary Research

Primary Research is also known as field research. The research is done in this to collect the original data that does not already exist. Secondary research is also known as desk research which involves the summary, collation and/or synthesis of existing research.

Here in this research of customer churn prediction of CU, this is a primary type of research as the research has been done to collect the original data from the financial

institute. This research is unique as no such work has been performed on the CU member dataset.

1.4.2 Based on objective: Qualitative Vs. Quantitative Research

Qualitative research is the non-statistical research to gain a qualitative understanding of the underlying reasons and motivations. It usually requires a smaller but focused dataset. It describes the research broadly and develops a deeper understanding of a topic. Quantitative research means the systematic study of the research and analysing the data statistically. It deals with investigating the quantitative data and recommending a final output of the research.

	Qualitative Research	Quantitative Research
Objective	To gain a qualitative understanding of the underlying reasons and motivations	To quantify the data and generalize the results from the sample to the population of interest
Sample	Small number of non-representative cases	Large number of representative cases
Data Collection	unstructured	structured
Data Analysis	Non-statistical	Statistical
Outcome	Develop an initial understanding	Recommend a final course of action

Table 1.1: Qualitative Vs. Quantitative Research (Source: Malhotra, 2007)

The current research is Quantitative research which uses data mining, involves the systematic investigation of customer data and is aimed at developing models, then verifying the results and then either the hypothesis is accepted or rejected based on the customer churn precision (Borrego, Douglas & Amelink, 2013).

1.4.3 Based on form: Exploratory Vs. Constructive Vs. Empirical

In Exploratory research, the research is being carried out for a problem that has not been clearly defined. It helps to determine the best research design, data collection method. Constructive research referred to a new contribution. A completely new

approach or new model or new theory is formulated in this research. It often involves the proper validation of the research via analytical comparison with predefined research, benchmark tests. Empirical research refers to the way of gaining knowledge through direct observation or experience. It involves the process of defining the hypothesis and then the predictions which can be tested with a suitable experiment.

This research is an empirical form of research because it involves defining the hypothesis and predicting the precision of customer churn by performing suitable experiments and then collating the results and then based on the results the hypothesis is accepted or rejected.

1.4.4 Based on reasoning: Deductive Vs. Inductive Research

A deductive approach is a top-down approach which is from the more general to more specific in which based on the pre-defined theory the hypothesis is defined and then the conclusion is drawn based on the research.

In Inductive research also known as a bottom-up approach which goes from specific observation to broader generalizations of theories.

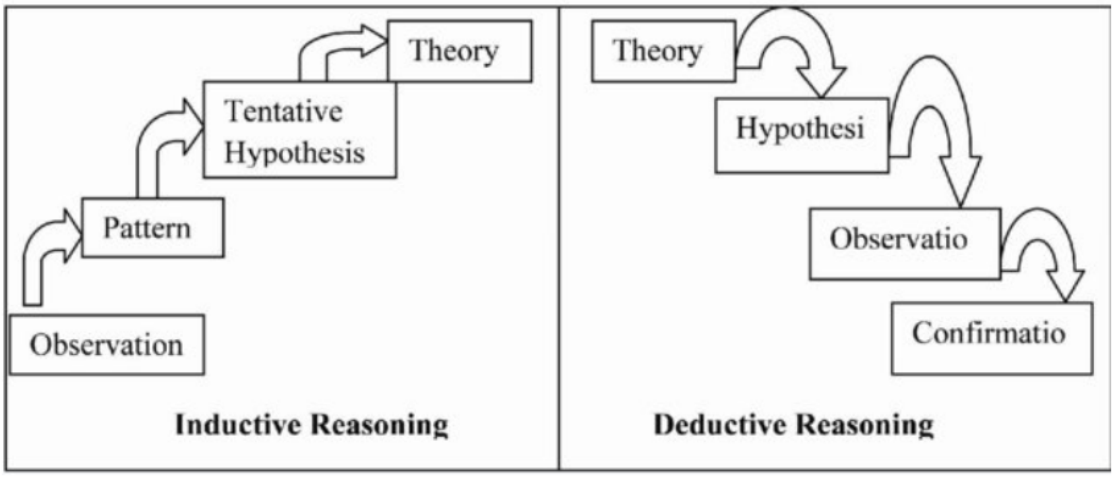


Figure : 1.1 Inductive Vs. Deductive Reasoning
(Source: Aliyu, Kasim & Martin, 2011)

A Deductive reasoning is employed in this research also known as the top-down approach (Saunders, Lewis, & Thornbill, 2009) as in this research firstly based on the theory the hypothesis is created and to study them the experiments were performed,

and the Supervised Machine learning models were built on the CU customer data to predict the churned customers. Then the champion model is selected based on the accuracy of the model.

Python programming language is used for statistical exploration of data, data cleaning, data preparation, building supervised machine learning models and evaluation of those models.

1.5 Scope and Limitations

The scope of this research is to develop a machine learning model using the CU's customer data to predict the customer churn.

The main limitation of the research is that the customer data is obtained from one CU only so it cannot be the representative of the other CU financial institutions. The customer base would be different for different CU institutions.

The other limitation of the research is that there are so many DateTime data type variables present which are not considered for building the classifiers. Also, the data imbalance is another limitation to overcome as the churned customers were less common, so less data was provided to the classifiers to study the features of churned variables.

1.6 Document Outline

This section outlines the thesis document:

This thesis report starts with defining and explaining the research problem and providing the importance of the research problem with the methodologies adopted, exploiting the problem and purpose of the problem with the proper research question.

Chapter 2 (Literature Review) discusses the literature related to customer churn prediction. This chapter reviews and compares the previous work done in this area using supervised machine learning techniques to predict the customer churn. It describes the use of SVM, Logistic Regression, Random Forest and Neural Network in previous work, their specifications and discusses most valuable work.

Chapter 3 (Design and Methodology) describes the design and methodology adopted to solve the research problem in detail. It follows the CRISP-DM methodology and each step is carried out and explained in detail in this chapter.

Chapter 4 (Implementation and Results) presents the implementation details and the results of the implementation. It describes in detail which models are chosen and which models have performed with proper justification. The hypothesis of the research is considered, and results are compared, and the hypothesis is evaluated.

Chapter 5 (Evaluation and Discussion) discusses the evaluation criteria of the supervised machine learning models. The champion model is determined, and the results were discussed. The research problem is discussed with the results and the hypothesis is evaluated. Also, the strength and limitations of the thesis are discussed.

Chapter 6 (Conclusion) discusses the research problem with the result obtained and evaluation. It summarises the research, discusses the contribution of the research towards the research question. Also, it recommends some future research work in a similar area.

2. LITERATURE REVIEW

This chapter provides a review of the literature available on CUs, Customer Churn prediction methods, various approaches adopted to solve the problem and evaluation metrics used for evaluating the models. The chapter concludes with the gaps in the existing research and forms the objective for the research.

2.1 Background

Customer Churn Prediction is important in all businesses because it helps to gain a better understanding of your customers and of future expected revenue. It can also help your business identify and improve upon areas where customer service is lacking. A lot of work has been done on this and still, and there are a lot of industries customer data to explore. The results differ for the data of different industries.

2.2 Customer Churn Prediction

The term Customer Attrition refers to the customer leaving one business service to another. Customer Churn Prediction is used to identify the possible churners in advance before they leave the company. This step helps the company to plan some required retention policies to attract the likely churners and then to retain them which in turn reduces the financial loss of the company (Umayaparvathi & Iyakutti, 2012).

Customer churn is a concern for several industries, and it is particularly acute in the strongly competitive industries. Losing customers leads to financial loss because of reduced sales and leads to an increasing need for attracting new customers (Guo-en & Wei-dong, 2008).

Customer retention is crucial in a variety of businesses as acquiring new customers is often more costly than keeping the current ones. Due to the unpredictable nature of customers, it is quite a daunting task to predict whether the customer will quit the company or not. For financial institutes, it is even more complex to identify the

customer churn due to the sparsity of the data as compared to another domain. This requires longer investigation periods for churn prediction (Kaya, et.al., 2018).

The economic value of customer retention is widely recognized (Poel & Lariviere, 2004):

(1) Successful customer retention allows organizations to focus more on the needs of their existing customers instead of seeking new and potentially risky ones.

(2) Long term customers would be more beneficial and, if satisfied, may provide new referrals.

(3) Long term customers tend to be less sensitive towards a competitive market.

(4) Long term customers become less expensive to serve due to the bank's knowledge

(5) Losing customers leads to reduced sales, and increased sales to attract new customers.

Customer Churn has become a major problem in all industries including the banking industry and banks have always tried to track customer interaction so that they can detect the customers who are likely to leave the bank. Customer Churn modeling is mainly focusing on those customers who are likely to leave and so that they can take steps to prevent churn (Oyeniyi & Adeyemo, 2015).

In an era of the competitive world, more and more companies do realize that their most precious asset is the existing customer base and their data. We mainly investigate the predictors of churn incidence as part of customer relationship management (CRM). Churn Management is an important task to retain valuable customers.

Business organizations, such as banks, insurance companies, and other service providers are changing their employees to be more customer-and service oriented and they are setting strategies to ensure customer retention (Nashwan & Hassan, 2017). The best core marketing strategy for the future is to retain existing customers and avoiding customer churn (Kim, Park & Jeong, 2004)

Previous research indicates that there were two types of targeted approaches to managing customer churn: reactive and proactive. In a reactive approach, the company

waits until the customer asks to cancel their service. In a proactive approach, the company tries to identify customers who are likely to churn. The company then tries to retain those customers by providing incentives. If churn predictions are inaccurate then companies will waste their money on customer churn so the customer churn should be accurate (Tsai & Lu, 2009).

2.3 Data Exploration and Pre-processing

Data Exploration is required to gain further understanding of the data and business problem. The CRISP-DM methodology is widely accepted for the Data mining model. It is mainly for conducting a data mining process, whose life cycle consists of six phases as shown in the below figure.

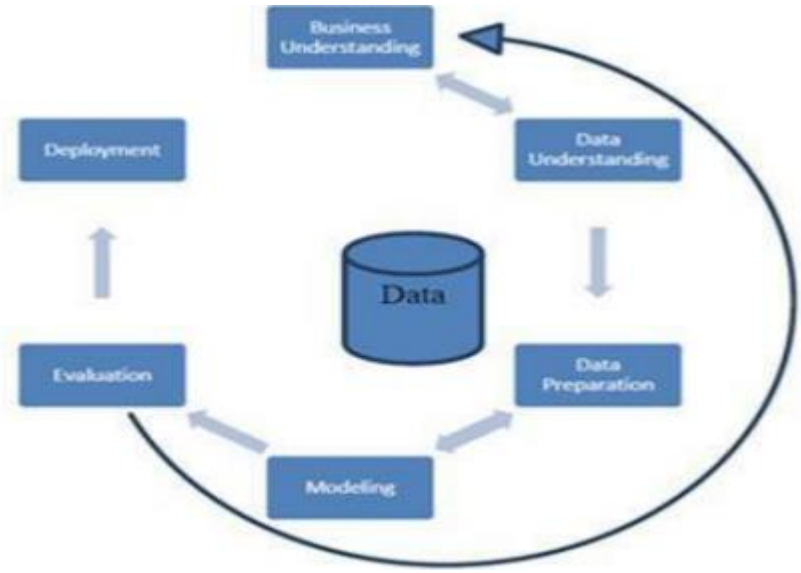


Figure 2.1: The phases of the CRISP-DM data mining model[2]

The most important stage of Data Analysis is the Data Preparation. In general, the data cleaning and pre-processing take approximately 80% of the time. The data preparation is more challenging and time-consuming part.

The Real-world data can be noisy, incomplete and inconsistent. The data preparation stage deals with – incomplete data where some attribute values were missing, where certain important attributes were missing. In the data preparation stage the outliers and errors in data were also handled, even the data discrepancies were handled in the data preparation. Data preparation generates a smaller dataset than the original one. This

[2] https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html

task includes selecting relevant data, attribute selection, removing anomalies, eliminating duplicate records. This stage also deals with filling the missing values, reducing ambiguity and removing outliers (Zhang, Zhang & Yang, 2003).

This stage is of high importance due to the following:

(1) the real data is impure;

(2) high-performance mining requires quality data;

(3) quality data yields high-quality patterns

2.3.1 Class Imbalance

As seen in research by Guo-en & Wei-dong (2008) class imbalance has become a common problem within datasets in data mining. This is a common problem in customer churn prediction area. In such a problem, almost all examples are labelled as the not churned class, while fewer examples are labelled as a churned class, the most important class.

One remedy to deal with the problem of class imbalance was using Under-Sampling, Random Sampling and Over-Sampling suggested by Bin, Peiji & Juan (2007). In their study on Customer Churn Prediction on Personal Handyphone System Service, the proportion of nonchurn and churn was set as 5:1. Models were trained using three different sample methods and better performance was observed using a random sampling method. The below figure depicted the performance.

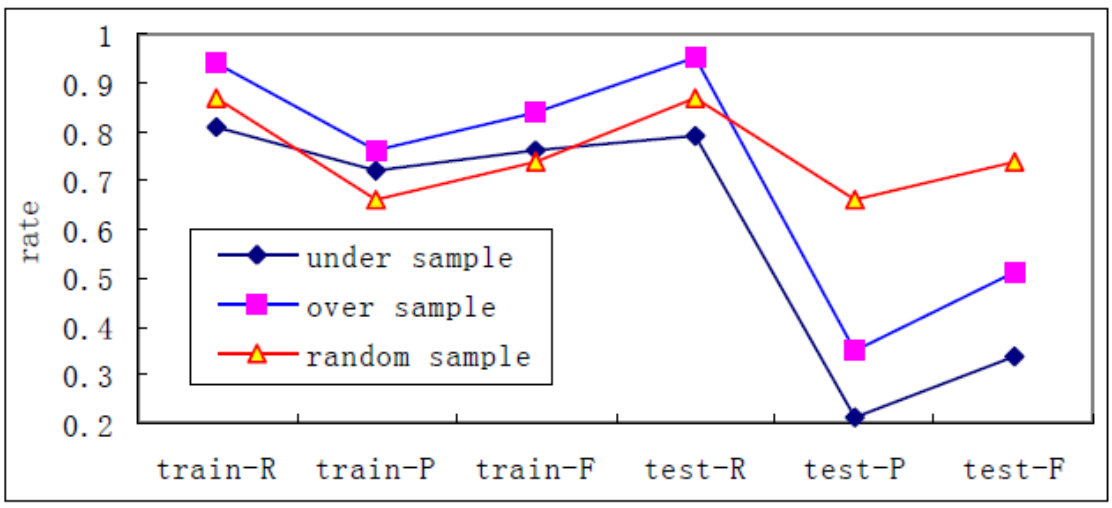


Figure 2.2: Effects of Sample methods

(Source: Bin, Peiji & Juan, 2007)

From the above figure, it was concluded that random sampling yielded the best results for a Decision Tree model.

It has been observed in the previous research by Maheshwari, Jain & Jadon (2017) detailed description of various approaches was discussed to handle class imbalance. The various approaches for handling class imbalance were – Data Level approach, Algorithm Level approach and Cost-sensitive approach. In the data level approach, various methods were – Under sampling, oversampling and hybrid sampling. The under-sampling leads to losing potentially useful data while oversampling leads to overfitting and increases learning process time if data is large. Another data level approach for handling class imbalance was to use SMOTE technique which yielded better accuracy as compared to other methods. Algorithm level approach for handling class imbalance are – bagging and boosting methods. For bagging method, the algorithms were Decision tree (C4.5) and Random forest, for boosting method AdaBoost and SMOTEBOOST algorithms were used. Bagging algorithms may lead to overfitting and boosting ignores the overall performance of the classifier. The cost-sensitive method incorporates both data and algorithm level approach. As per the research, it was evaluated that the Data Level approaches were the best approach to handle the class imbalance.

In research by Kaya, et. al. (2018) it was observed that the SMOTE technique yielded the best results for SVM in predicting customer churn. SMOTE generates minority classes by interpolating instead of replication and avoided over-fitting problem hence provided better results.

Imbalance problem can be solved at the algorithm level also. Cost-Sensitive Learning is a type of learning that considered misclassification costs. A cost-sensitive learner assigned a greater cost to the false negatives compared to the false positives. However, it was not a very feasible approach, as the cost information was dependent on many other factors (Ganganwar, 2012). Another algorithm-based approach was one class learning which follows the separate-and-conquer approach in which the classifier was modelled only on minority class. This approach was useful for highly unbalanced data sets composed of a high dimensional noisy feature space (Kotsiantis, Kanellopoulos, & Pintelas, 2006).

2.3.2 Feature Selection

Feature Selection is the process of identifying the fields which are the best for prediction as a critical process (Hadden, Tiwari, Roy & Ruta, 2005). This step is important in customer churn prediction. Feature selection is a process of selecting a subset of original features is an important and frequently used dimensionality reduction technique for data mining.

In one of the researches done by Khan, Manoj, Singh & Bluemenstock (2015) t-test was performed separately for each feature, which indicated the extent to which a single feature can accurately differentiate between people who have churned or not. A Tree-based method was used for feature selection. This method was useful in producing a list of correlated predictors.

The feature selection was categorized into two categories based on Label Information and Search Strategy. The below diagram will detail the division.

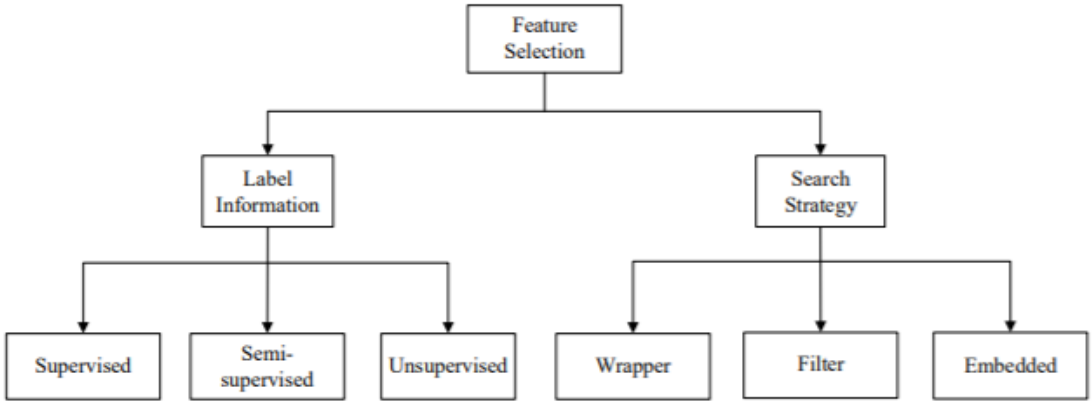


Figure 2.3: Feature Selection Category

(Source: Miao & Niu, 2016)

The training data can be labelled, unlabelled or partial labelled which leads to the development of Supervised, Unsupervised and Semi-Supervised feature selection algorithms. Supervised feature selection determined the feature relevance by evaluating the feature's correlation. Unsupervised feature selection exploited data variance and separability to evaluate feature relevance. Semi-supervised feature selection algorithm used both labelled and semi-labelled data and improved the feature selection of unlabelled data. Based on search strategy three categories of feature selection are filter, wrapper and embedded models. The filter model evaluated features without involving any learning algorithm which relies on the general characteristics of data. The wrapper model required a predetermined learning algorithm and used its

performance as an evaluation criterion to select features. Algorithms with an embedded model, e.g., C4.5 and LARS, were the examples of wrapper models which incorporate variable selection as a part of the training process, and feature relevance was obtained analytically from the objective of the learning model (Miao & Niu, 2016).

According to researchers Cai, Luo, Wang & Yang (2018) Supervised feature selection for classification problem using the correlation between the feature and the class label as its fundamental principle. The correlation between the features were determined and compared to the threshold to decide if a feature was redundant or not. This method was an optimal feature selection method which maximized the classifiers accuracy.

2.4 Machine Learning

Machine Learning is a method of data analysis which assists in analytical model building. It is a branch of Artificial Intelligence (AI). The machine learning models learn from the data, identify general patterns in it and construct decision with minimal human intervention.

Machine Learning is mainly used when we have a complex problem or task involving a huge amount of data. It is a good option for more complex data and deliver faster, more accurate results. It helps an organization of identifying profitable opportunities or any unknown risks (Sayed, Fattah & Kholief, 2018).

Machine learning mainly uses two types of learning techniques:

1) Supervised Machine Learning
2) Unsupervised Machine Learning

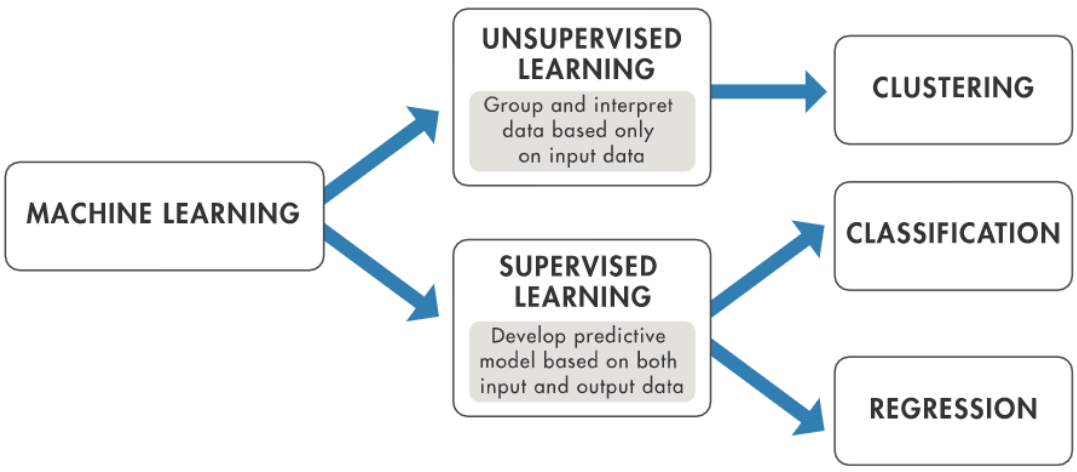


Figure 2.4: Machine Learning Techniques – Unsupervised and Supervised Learning[3]

2.4.1 Supervised Machine Learning

Supervised Machine Learning is the computational task of learning correlations between variables in training dataset and then utilising this information for creating a predictive model capable of inferring annotations for new data (Fabris, Magalhaes & Freitas, 2017). In Supervised Machine learning, we have an input variable (X) and an output variable (Y) and we use an algorithm to learn the mapping from the input to the output.

$$Y = f(X)$$

The goal is to approximate the mapping function so well that when the new input data (X) is introduced the model predicts the output variable (Y) for that data.

The learning is called as Supervised learning when instances are given with known labels. The features can be continuous, categorical or binary (Kotsiantis, Kanellopoulos & Pintelas, 2006).

The supervised learning problems can be grouped into regression and classification –

1) **Classification –** When the output variable is categorical, such as "red" or "blue" and "yes" or "no" then it is considered as Classification problems.
2) **Regression –** When the output variable is a real value, then such problems are considered as Regression problems.

[3] https://vitalflux.com/dummies-notes-supervised-vs-unsupervised-learning/

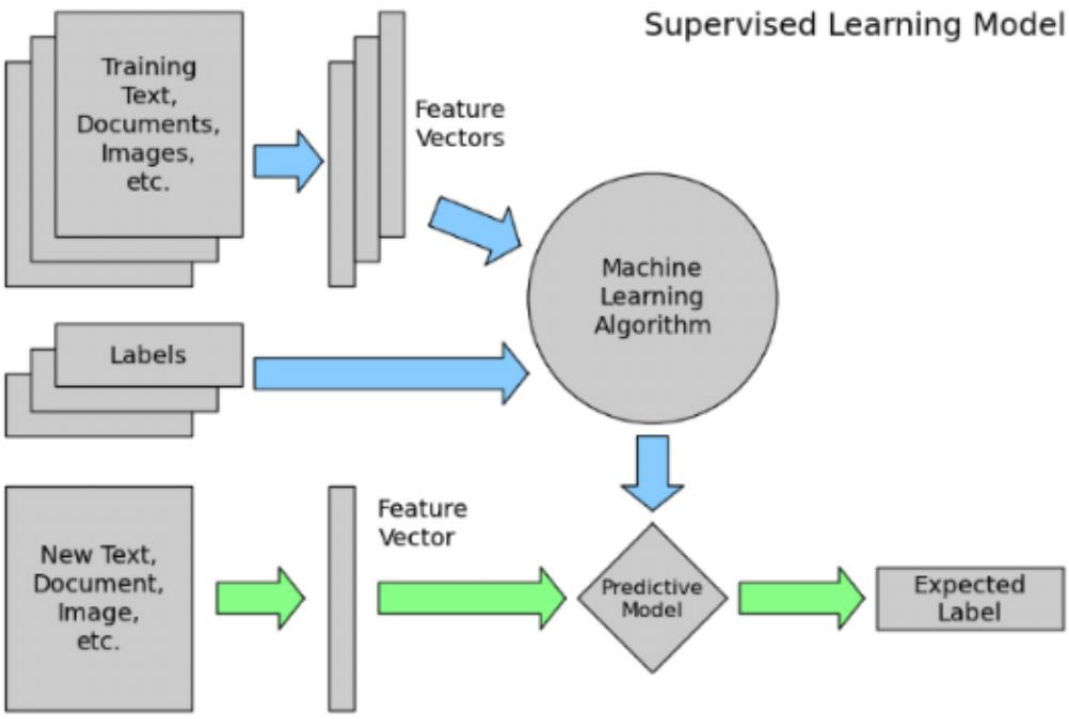


Figure 2.5: Supervised Machine Learning Model

(Source: Vladimir, 2017)

2.5 Machine Learning Techniques

Several machine learning techniques have previously been used in similar customer churn prediction problems.

2.5.1 Logistic Regression

Logistic Regression was very widely used statistical model used for Customer Churn and has been proven a powerful algorithm.

The formula in figure 7 below represents logistic regression where pi is the probability and xi is the independent variables which predicted the outcome pi.

$$p_i = \frac{e^{x_i\beta}}{1 + e^{x_i\beta'}}$$

Figure 2.6: Logistic Regression Formula

(Source: Nie, 2011)

In a study on churn prediction of credit card in China's banking industry Logistic Regression and Decision Tree model were built. It was observed that Logistic Regression has performed better than Decision Tree (Nie, 2011). There were 135 variables and, in this research, instead of selecting all 135 variables certain variables were selected and models were built based on correlation and in the study, it has been observed that Logistic Regression model has performed better than Decision tree algorithm.

Researchers have implemented the binary and ordinal logistic regression models for customer churn prediction using SAS 9.2 with the Logistic regression procedure and Cox regression models (Ali & Ariturk, 2014).

In another study of comparison of models performed on predicting Customer Churn on Telecom dataset, it was observed that the Logistic Regression model has outperformed Decision Tree. Logistic Regression uses maximum likelihood estimation for transforming the dependent variable into a logistic variable. The proposed system provided a statistical survival analysis tool to predict customer churn. The confusion matrix was used for evaluation purpose (Khandge, Deomore, Bankar & Kanade, 2016).

2.5.2 Random Forest

Random Forest is an ensemble learning method model for a classification or regression problem. A decision tree is the building block of random forest. The multitude of decision tree makes the random forest and the output is the mode of the class in classification and mean prediction for the regression problem. Random Forest will not overfit if enough number of trees are there in the classifier. It can handle missing values and is suitable for categorical variables also.

In one of the previous researches on financial customer churn, the researchers have used Random forest classification technique (Kaya, et. al., 2018).

2.5.3 Support Vector Machine

Support Vector Machine model is a supervised machine learning model which can be used for classification as well as regression problems. SVM is mostly used in a

classification problem as it can separate two classes using a hyperplane. The objective of SVM is to find a hyperplane that can distinctly classify the data. Hyperplanes are decision boundaries that help classify the data points. Support Vectors are data points that are closer to hyperplane and influence the position and orientation of the hyperplane.

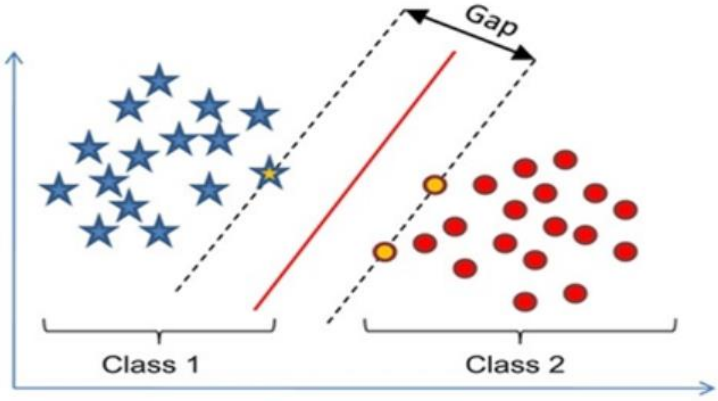


Figure 2.7: Support Vector Machine

(Source: Ali, 2018)

Several researchers have implemented mainly two methods for customer churn prediction. The first method was the traditional classification method using supervised learning mainly for quantitative data and the other was artificial intelligence method for large scale, high dimensionality, nonlinearity and time-series data (Guo-en & Wei-dong, 2008). In existing research of customer churn prediction problem in telecommunication industry the researcher Guo-en, X., have used SVM model as it can solve the nonlinearity, high dimension, and local minimization problems. The model prediction depended on the data structure and condition.

2.5.4 Neural Network

Neural Networks are a set of algorithms, that are designed to recognize patterns. The basic building blocks of neural network is neurons. The output depends on the activation function of the neuron.

The researcher Zoric, 2016 have used neural network model within the software package Alyuda NeuroInteligence for his research on customer churn prediction in Banking industry because neural network worked well for pattern recognition, image processing, optimization problems etc.

Another group of researchers Huang, Kechadi, Buckley, Keirnan, Keogh & Rashid, 2010 have proposed the comparison between the popular modelling technique – Multilayer Perceptron Neural Networks and Decision Tree with the innovative modelling technique – SVM (Huang, et.al. 2010) for customer churn prediction in the telecom industry. MLP and SVM were more efficient than Decision Tree.

2.6 Model Evaluation

According to researchers, the evaluation strategies were of two types – filter and wrapper. In wrapper evaluation method, evaluation was performed on a subset of features using a learning algorithm whereas in filter evaluation, the evaluation was done on a subset features external to the classification design (Hadden, Tiwari, Roy & Ruta, 2005).

Customer churn prediction problem is a classification problem and to evaluate the performances of the supervised machine learning models, precision, recall, accuracy and F-measure were calculated using Confusion Matrix (Vafeiadis, Diamantaras, Sarigiannidis & Chatzisavvas, 2015).

Apart from confusion matrix, many researchers have used the AUC (Area Under Curve) also for evaluation of the model. The AUC is the area under the Receiver Operating Curve (ROC) curve. ROC is a plot between true positive rate versus false positive rate. Another evaluation metrics was TDL (top-decile lift) which focuses on the customer most likely to churn (Ali & Ariturk, 2014)

2.7 Historic Customer Churn Prediction

In any organisation, the Customer Relationship Management is a prominent field in the business analysis field. It deals with retaining existing customers, then identifying, expanding and attracting the potential customers. CRM has two aspects one is the technical aspect and the other is the operational aspect. The technical aspect of the CRM also known as Customer Analytics (Senanayake, Muthugama, Mendis & Madushanka, 2015). Customer Analytics can be broken into two categories:

(1) Descriptive Analytics – In which the customer identification was done

(2) Predictive Analytics – In this, the retention of customers was focused.

The Predictive Analytics is the customer churn analysis which mainly focuses on retaining the customers.

According to the researchers Senanayake, Muthugama, Mendis, & Madushanka, (2015) the typical approach of identifying the customer without machine learning was to analyse the data of those customers who have already churned and identifying customer attrition from the existing customers based on observation and customer behaviour.

2.8 Customer Churn Prediction using Machine Learning

Now as the time passes the data increases and due to the volume of data is immense it becomes a daunting task for the data analysts to analyse such huge data. So, then the customer churn prediction using machine learning and data mining techniques played a significant role.

Customer churn prediction using machine learning models follow a set of steps. The data is collected, next, the selected data was pre-processed and transformed into a suitable form for building a machine learning model. After modelling the testing was performed and then finally the model was deployed (Kim, Shin & Park, 2005). The machine learning investigated the data and detects the underlying data patterns for the customer churn analysis (Kim, Shin & Park, 2005). Using machine learning the prediction of customer churn was more accurate than the traditional approach.

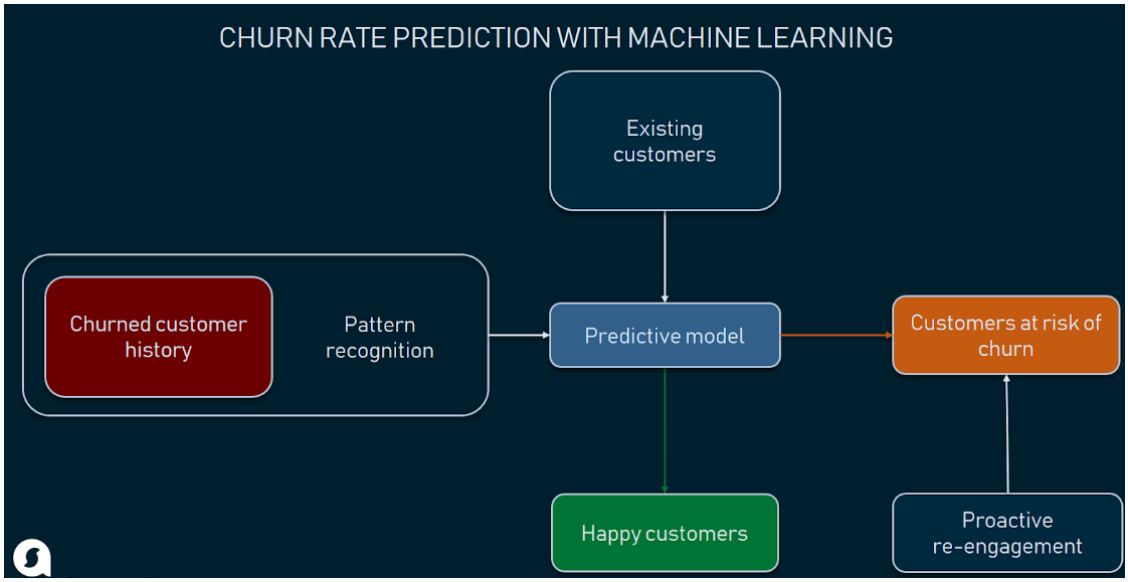


Figure 2.7: Churn Rate Prediction using Machine Learning

(Source: Beker, 2019)

Several features were involved as variables in customer churn analysis. The various category of variables was customer variables of recency, frequency and monetary value (RFM), demographic features like the geographical details, cultural information and age (Senanayake, Muthugama, Mendis & Madushanka, 2015).

2.9 Approaches to solve the problem

Many researchers have worked on the prediction of customer churn. Most of the research was based on applying machine learning algorithms on customer data and predicting the customer churn rate. A few of the studies are discussed in this section.

Researchers Guo-en, & Wei-dong, (2008) have applied the machine learning method SVM on structural risk minimization to predict the customer churn on telecom industry customer data set. They have analysed the results of the SVM model with an artificial neural network, decision tree, logistic regression, and naïve Bayesian classifiers. In the experiment it was found that the SVM has outperformed with best accuracy rate, hit rate, covering rate and lift coefficient. There were two datasets used in the research and for SVM model the kernel function was selected using MATLAB 6.5. For the first dataset the SVM has acquired good results using kernel function as radial basis function and for the other dataset Cauchy kernel function was used. The SVM model

accuracy was calculated as 90% and 59% for dataset 1 and dataset 2 respectively. Decision Tree C4.5 had the least performance for both the datasets with accuracy as 83% and 52% respectively.

Another study on European financial bank customer data was conducted by Poel & Lariviere, (2004) using the Cox proportional hazard method to investigate customer attrition. The focus was on churn incidence. The SAS enterprise miner was used in this research. They performed the research by combining several different types of predictors into one comprehensive proportional hazard model. By analysing this bank customer dataset two critical customer churn periods were identified – firstly the early years after becoming the customer and a second period is after some 20 years. Demographic and environmental changes were of major concern and have a great impact on customer retention. In this research, four retention predictor categories were used it would have been more advantageous if the data obtained were merged and would have incorporated in a single retention model instead of four different models.

Hybrid neural networks were built, and the performance was compared with the baseline ANN model by the researchers Tsai & Lu (2009). The customer churn was predicted on the American telecom company data. In this research, they have built one baseline ANN model and two hybrid models by combining the clustering and classification methods to improve the performance of the single clustering or classification techniques. It comprised of two learning stages, the first one, was used for pre-processing the data and the second one for the final output prediction. The two hybrid models built were ANN+ANN (Artificial Neural Network) and SOM (Self Organizing Maps) +ANN. These models were evaluated based on the Type I and Type II error rates and the accuracy of the models. In statistical hypothesis testing a type I error was the rejection of a true null hypothesis, while type II error was the non-rejection of a false null hypothesis. The actual results showed that the ANN+ANN model performed better than both the ANN and SOM+ANN models in terms of Type I error rates. Also, the prediction accuracy for ANN+ANN hybrid model was better than that of ANN and SOM+ANN models. Thus, in this research paper hybrid techniques were performed. The hybrid model with two ANN has performed better when compared to SOM+ANN hybrid model. Feature selection was not considered in this research.

In one of the research papers on customer churn in the financial industry by researchers (Kaya, et. al., 2018) they have emphasized more impact on Spatio-temporal features. They have adopted Random Forest as the classification model for their study and trained the model with 500 trees and maximum of 2 features per tree. Stratified 8-fold cross-validation was adopted for evaluation. In this research, Spatio-temporal and choice features were found more superior than demographic features in financial churn decision prediction. In this research, it was observed that young people were more likely to leave the bank. The results of this research suggested that based on mobility, temporal and choice entropy patterns which can be extracted from customer behaviour data we can predict the customer churn rate. The evaluation was performed using AUC ROC evaluation metrics.

Researchers Oyeniyi & Adeyemo (2015) have predicted the customer churn problem on one of the Nigerian bank datasets and they have used WEKA tool for knowledge analysis. K-means clustering algorithm was used for clustering phase followed by a JRip algorithm rule generation phase.

Customer Churn prediction was performed on Personal Handy Phone System Service by researchers Bin, Peiji & Juan, (2007). They have built a Decision tree and three experiments were conducted to build an effective and accurate customer churn model. In this research 180 days data was randomly sampled and utilized in the research for churn prediction. In the first experiment sub-periods for training data sets were changed, in the second experiment, the misclassification cost was changed in churn model and then in the third experiment being conducted sample methods were changed in the training data sets. In this study in first experiment, the number of sub-periods were considered as 18, 9, 6 and 3 which means the 180 days call record data is divided into 18, 9, 6 and 3 parts. In second experiment, the misclassification cost means setting the proportion of nonchurn and churn customers in training dataset. In third experiment, various sampling techniques were adopted to balance the dataset. This research helped in churn prediction and in improving the performance of churn prediction models. In this study, it has been observed that the performance of the model was superior when sub-period was set as 18. In the case of misclassification cost when it was set as 1:2, 1:3 and 1:5 the result was superior and finally in case of sample method random sample method has yielded the best results in the research.

A comparative study on customer churn prediction was performed by Vafeiadis, et.al. (2016) on telecom data set. The performance comparison of multi-layer perceptron, Decision Tree, SVM, Naïve Bayes and Logistic regression were compared. All the models were built and evaluated using cross-validation. Monte Carlo simulations were used and SVM has outperformed other models with an accuracy of 97% and F-measure of 84%.

In one of the previous researches, on Customer churn prediction, the researchers have used the traditional method supervised machine learning algorithms – Decision Tree, Regression Analysis for prediction and also the Soft Computing methodologies such as fuzzy logic, neural networks and genetic algorithms (Hadden, Tiwari, Roy & Ruta, 2005).

Sharma & Panigrahi (2011) have performed the customer churn prediction on telecom dataset using Neural Network. The neural network has yielded better result with accuracy of 92%. The researcher has focused on changing the number of neurons and increasing the hidden layers in the neural network model. Feature selection and class imbalance problem were not considered in the research.

In one of the comparison research paper by Xie, Li, Ngai & Ying (2009) it has been observed that balanced Random Forest has outperformed the other classifiers ANN, SVM and DT based on precision and recall.

2.10 Summary, Limitations and Gaps in the Literature Survey

A detailed study of state-of-the-art approaches in predicting customer churn has been performed for this research. It has been observed that there is a need to focus more on data pre-processing stage. Most of the research has not handled the feature selection and class imbalance problem.

Most of the research (Xie, et.al., 2009, Sharma, 2011, Vafeiadis, et.al., 2016) were performed on telecom customer dataset and few research (Oyeniyi & Adeyemo, 2015), Kaya, et. al., 2018) were performed on financial dataset. Customer churn prediction on Personal Handy Phone System Service by researchers Bin, Peiji & Juan, L. (2007) was performed. No research has been focused on the customer churn prediction on CU financial institute.

Currently, the vital and active areas of research in Customer Churn prediction was using feature selection for data mining purposes (Guo-en & Wei-dong, 2008). Also, while implementing SVM how to select fitting kernel function and parameter, how to weigh customer samples (Guo-en & Wei-dong, 2008). For further research, it would be a challenge to incorporate customer behaviour, customer perceptions, customer demographics and macroenvironment into one comprehensive retention model (Poel & Lariviere, 2004). More focus should be emphasized on the pre-processing stage for better performance, the dimensionality reduction or feature selection. Also, other domain data sets for churn prediction can be used for further comparisons (Tsai & Lu, 2009). Research should be aligned towards improving the predictive ability of churn model by using other data mining techniques, for example, neural net, logistic regression, self-organizing map, support vector machine and so on (Bin, Peiji & Juan, 2007).

Most of the studies were done using archived data. In the existing research not, much guidance was provided on how to analyse the real-world application dataset. To address the limitations and research gaps presented in this section, the research was focused on covering the data pre-processing steps of feature selection by using correlation technique and extra tree classifier method, handling class imbalance using SMOTE technique. Further, secondary research was also conducted focusing on a comparative study on churn prediction of Banking domain dataset (Kumar & Vadlamani, 2008) with the current research prediction results.

3. DESIGN AND METHODOLOGY

In this chapter, the design of the research and the methodology will be explained in detail to answer the research question. The experiment design followed the CRISP-DM process in the research lifecycle. Python programming was used to carry out the experiments of the research.

This research aimed at building and comparing the supervised machine learning techniques using a CU customer dataset to predict the customer churn rate. The Logistic Regression, Random Forest, SVM and Neural Network supervised machine learning models were built, and the results were compared. The secondary research focused on a comparative study of research results with the existing research paper results on banking domain (Kumar & Vadlamani, 2008).

The overall workflow of the research is as shown below.

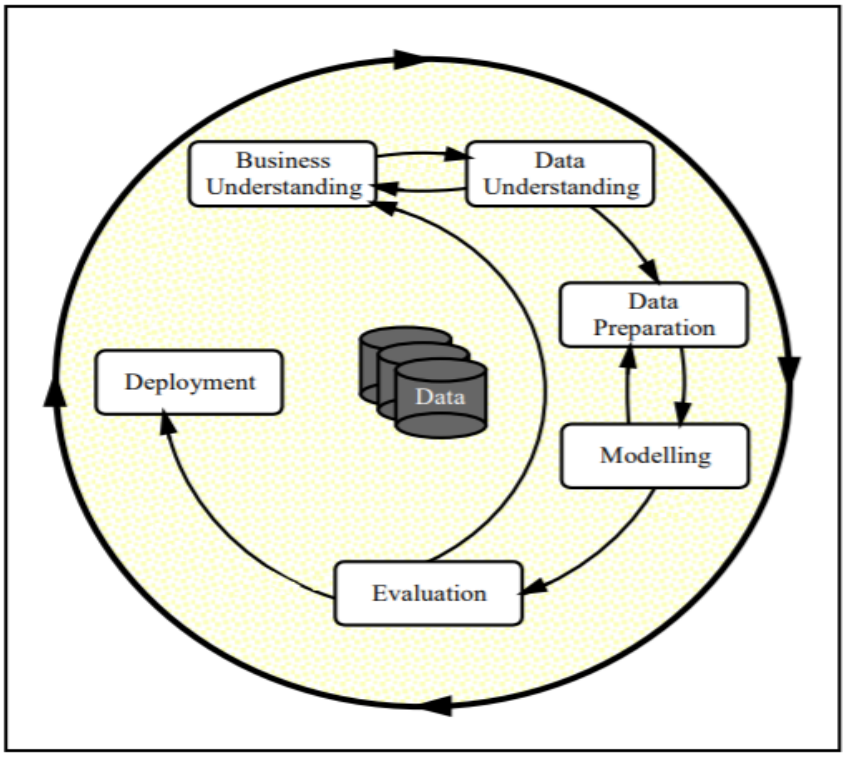


Figure 3.1: CRISP_DM Process

(Source: Wirth & Hipp, 2000)

The thesis followed the CRISP-DM methodology, and each of the phases are described in detail below.

3.1 Business Understanding

This is the initial phase of CRISP-DM and focuses on understanding the business problem. In this phase, the problem was studied based on a business perspective. The business in this research was CU financial institution which have 3.6 million members throughout Ireland and here in this research 96968 members data were taken into consideration for the study for a single CU for the years 1911 – 2019.

CUs are not-for-profit, member-owned, financial cooperatives, whose earnings are paid back to members in the form of higher savings rate and lower loan rates. Members of CU are both customers and owners. Banks, on the other hand, are primarily to generate profit that returns to its owner.

Machine learning is not currently used in CUs for customer churn prediction.

In this research the main aim was to identify the members which were more likely to close their membership and to identify this, the supervised machine learning models (Logistic Regression, Random Forest, Support Vector Machine and Neural Network) were built and the results were compared and based on accuracy, precision metrics the best model was selected. This research will be helpful for the CU to identify the members who are likely to close their membership and then they can focus on those members to retain them. They can communicate with those customers and can understand their needs and in turn, can reduce the customer churn which in turn leads to the growth of the institute.

The thesis of this research is to demonstrate the following hypothesis:

H0: *A random forest supervised model build using the CU customer data, will not achieve high accuracy than the other supervised machine learning algorithms like Logistic Regression, Support Vector Machine and Neural Network, to predict the customer churn.*

HA: *A random forest supervised model build using the CU customer data, will achieve high accuracy than the other supervised machine learning algorithms like Logistic Regression, Support Vector Machine and Neural Network, to predict the customer churn.*

3.2 Data Understanding

The Data Understanding phase deals with the collection of data and data exploration to get basic insight into the type of data. Some understanding of data was gained in this phase.

The dataset used in this research was the customer data of the financial institute called CU. The dataset was completely original, and no statistical research has been done on this dataset. It consists of the data of all customers who have joined the CU from 1911 to 2019. The dataset has 96967 records of distinct members with 48 features. The customer churn was defined as the total number of customers who have closed their accounts. In this research, the customers who are not deceased and whose accounts were either closed or dormant were considered as churned from CU.

The data was loaded using the pandas library of python. The number of records were explored, using the info() function the datatype of each independent variables were identified.

The basic quantitative analysis of the data was carried out. The measures of central tendency, range, standard deviation, mean, max, min of the variables was measured here using Descriptive Statistics. Also, the skew and kurtosis of the variables were measured to check the normality of the variables. Exploratory Data Analysis (EDA) was performed. The data visualisation was performed using matplotlib and seaborn python libraries and histogram, box-plot was created to view the data distribution for checking the normality and to identify the outliers in the variables.

The correlation matrix was built using the Spearman method to identify the correlation between the dependent and independent variables and to identify the correlation between dependent variables to avoid multicollinearity.

3.3 Data Preparation

In the data preparation phase, all activities were performed to convert the raw data into the final dataset which we can feed into the modelling algorithms and build models. Various tasks like data cleaning, removing outliers, imputing missing values,

construction of new attributes, feature selection and transformation of data all tasks were performed in this phase.

3.3.1 Handling Missing Values

It is very important to handle missing values as many machine learning algorithms do not support data with missing values.

The given dataset comprised of many missing values. This may be caused due to a number of various factors. One of the reasons may be that the data was not collected. Variables with more than 60% of values missing can be removed from the final dataset (Kelleher, Mac Namee, & D'Arcy,2015).

For continuous variables which can take any values between its minimum and maximum values, with missing values from 2% till 30% the values were imputed by mean values. Mean is a reasonable estimate for randomly selected observations from a normal distribution. Missing values may be caused by several different factors. Missing data generated various problems. Missing data reduced the statistical power which can lead to the wrong evaluation of the hypothesis. It could also reduce the representativeness of the sample. It could complicate the analysis of data. Few algorithms do not work with missing data. The missing values in categorical variables which contain labels were imputed using the maximum likelihood and last observation carried forward techniques were the most common techniques for imputing the missing values (Kang, 2013). In maximum likelihood the missing values were imputed with the values which occurred most of the time. In the last observation carried forward technique the previous observation was imputed in the missing value.

3.3.2 Normalizing Data

Normalization is a technique applied as a part of data pre-processing for building machine learning models. The goal of normalization is to change the values to a common scale. The skew and kurtosis were measured for each numerical column and if the skew and kurtosis values were outside the range of +/-2 then the variable was said to be skewed data. Also, the histogram can be used to depict the normal distribution of the data.

These attributes were normalized using the sklearn MixMaxScaler() function.

3.3.3 Feature Selection

Feature selection is applied to the dataset as there is a large number of dependent variables in the given dataset. The feature selection is used to find relevant features for the model construction. The Correlation Matrix with heatmap method was used to do the feature selection in this research. A correlation was determined between the dependent and independent variables and between the dependent variables. Correlation was a measure of how strongly one variable depends on another. If the correlation goes beyond the threshold of correlation greater than 0.5 the variables will not be considered as it will affect the model accuracy (Mukaka, 2012). The size of correlation with interpretation is displayed in the below table.

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

Table 3.1: Correlation Table

(Source: Mukaka, 2012)

Feature selection improves the accuracy of the model. It trains the model faster and reduced the complexity of the model. Another method Tree based classifier was also implemented to find the most predicting features for feature selection based on a literature review. It is an ensemble learning method and used to predict the best features in predicting target.

3.3.4 Encoding

The dataset contains continuous and categorical variables. There are few machine learning algorithms like SVM and Logistic Regression which accepts only numeric data. For this reason, the categorical data is converted into 0 and 1 using label

encoding. In this dataset a total of 21 variables were categorical variables with True and False values or some nominal values. These values were transformed into numerical using sklearn's[4] LabelEncoder function.

3.3.5 Data Sampling

In many real-world application class imbalances is the most common data issue. In such problems, most of the examples are labelled as one class, while fewer examples are labelled as the other class, usually the important ones. This problem is known as a class imbalance. Class imbalance problem exists in lots of application domains (Guo, 2016).

Before undertaking an experiment, a decision must be made on class imbalance problem as the minority class was of prime importance in this research. Here in this research, the class imbalance ratio of approximately 18:5 was found which means against 18 non churned customers 5 will be churned. The non-churned members were 75% of total members whereas the churned members were 25% of the total CU members.

The course of action can be taken in the data pre-processing phase of the project was either the random undersampling or random oversampling which were the data level methods to handle class imbalance problem. As observed in the previous research by Maheshwari, Jain & Jadon (2017) both the undersampling and oversampling have advantages as well as disadvantages. Oversampling can lead to overfitting and lead to more computation work for large datasets whereas undersampling can lead to the removal of some significant data records. Here in this research, SMOTE technique was used to handle class imbalance problem.

3.3.5.1 SMOTE Technique

SMOTE (Synthetic Minority Over-sampling Technique) generates synthetic samples from the minority class. It was used to balance the training dataset synthetically which was then used to train the classifier. It created synthetic samples from the minor class instead of creating copies of data from minority class. It selected the similar records

[4] https://scikit-learn.org/stable/about.html%22%20%5Cl%20%22citing-scikit-learn

from minority class and altered that record one column at a time by a random amount to balance the data. It simply added the records only in minority class records (Ali & Ariturk, 2014).

The final dataset will then be splited into train and test datasets. The data split was done as 80% in training and remaining 20% for testing based on the previous research (Shaaban, Helmy, Khedr & Nasr, 2012). The data was divided using a for loop so that every time random data was chosen for building the model as well as evaluating the model in each iteration.

3.4 Modelling

In this phase, the pre-processed data obtained was used to build the machine learning model to predict customer churn. The main objective of the research was to build supervised machine learning models to do a comparative study and to find the best model for prediction. Logistic Regression, Random Forest, SVM and Neural Network algorithms were used to predict the customer churn. The four models in detail are –

3.4.1 Logistic Regression

Firstly, the Logistic Regression model was built. It is the most preferred algorithm for modelling binary dependent variables. It is a type of probability statistical classification model mainly used for classification problems (Korkmaz, Guney & Yighiter, 2012). The technique can work well with a different combination of variables and can help in predicting the customer churn with higher accuracy. The predictive power of the variables can be calculated.

It is a statistical model in which the curve is fitted to the dataset. This technique is useful when the target variable is dichotomous. It is a predictive analysis algorithm based on the concept of probability. (Singh, Thakur & Sharma, 2016).

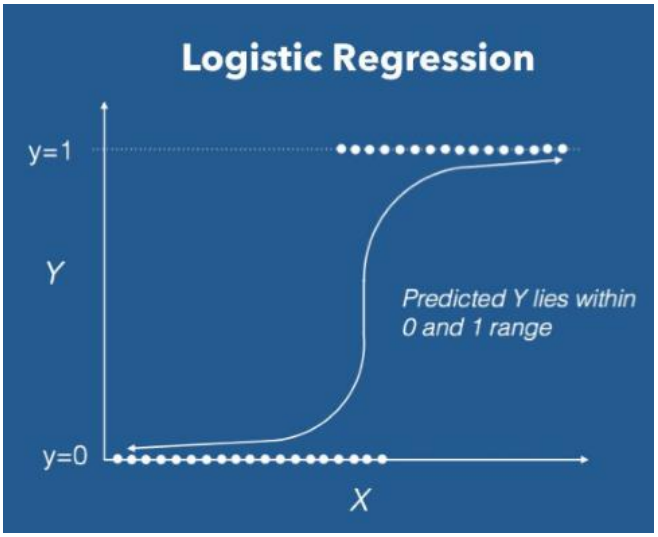


Figure 3.2: Logistic Regression

The Logistic Regression uses a complex cost function which is defined as 'Sigmoid Function' or also known as the 'logistic function'. The sigmoid function was used to map the predictions to probabilities. It limits the output and returned a probability score between 0 and 1. Logistic Regression was one of the popular algorithms for classification problem.

3.4.2 Random Forest

The Random Forest machine learning technique was chosen for predicting customer churn. It is a combination of multiple decision trees. It is an ensemble (a group of Decision Trees) learning method for classification, regression problems and used the bagging technique to generate the results. In ensemble learning, a group of weak learners come together to form a strong learner. Bagging also known as Bootstrap Aggregation was used to reduce the variance of the Decision Tree. It is an ensemble method in machine learning which is used to combine the predictions from multiple machine learning algorithms together to generate accurate results. The default hyperparameters of Random Forest gives good result and it is great at avoiding overfitting (Pretorious, Bierman & Steel, 2016). In Random forest, the most common output in all the decision trees was selected as the predicted class as the result.

It operates by constructing a multitude of decision trees at training time and outputting the class as the mode of the class or mean prediction of the individual trees. Random

Forest was used for customer churn prediction in this research as it was quite fast and can deal with unbalanced and missing data as well (Idris, Rizwan & Khan, 2012).

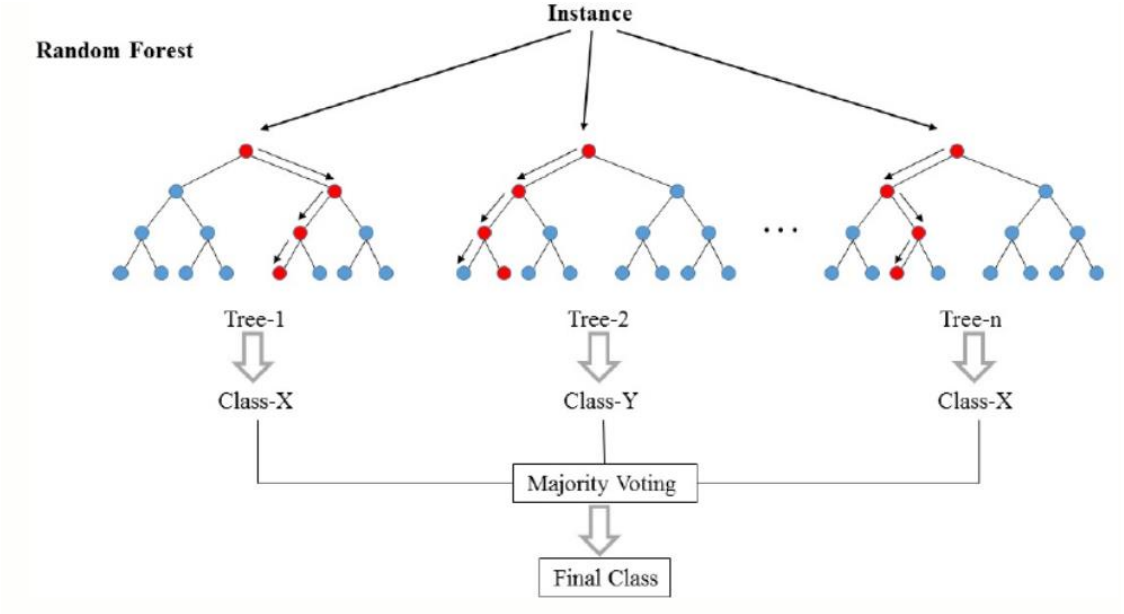


Figure 3.3: Random Forest

3.4.3 Support Vector Machine

Another machine learning technique that will be evaluated is SVM. SVM is mostly used for classification problems and it builds the hyperplane margin between two classes. This algorithm uses a set of mathematical function called a kernel which transforms the input data into the required form. Popular kernel choices for SVM were linear, polynomial and Radial Basis Function (RBF) (Tian, Shi & Liu, 2012). Here in this research linear kernel was used based on the previous researches on customer churn for linearly separable data (Guo-en & Wei-dong, 2008).

Support Vector Machine (SVM) is a supervised machine learning models with associated learning algorithms that analyze the data for classification or regression problems. This algorithm works on the kernel function. The data is transformed based on the kernel function and an optimal boundary is set between the possible outputs (Guo-en & Wei-dong, 2008).

Guo-en, & Wei-dong (2008) have used SVM to determine the customer churn prediction in telecommunication customer data. SVM solved the nonlinearity, high dimension, and local minimization problems in customer churn prediction. As per the

existing research SVM can work well with financial customer dataset also (Guo-en & Wei-dong, 2008).

3.4.4 Neural Network

Finally, the Neural Network was evaluated. Neural Network is a nonlinear predictive model which learns through training and the structure is like a biological neural network. Neurons acts as the basic building blocks of the network. The output depends on the activation function of the neuron. Here in this research relu activation function was used based on the previous research. The relu (Rectified Linear Unit) activation function was computationally less expensive. It takes input and then each input was multiplied by a weight. Then all the weighted inputs were summed up together with a bias. Finally, the sum was passed through an activation function. The most common activation function used for Neural Network was 'Sigmoid' function. This function is useful for binary classification as it outputs in the range of 0 and 1 (Zoric, 2016).

In this research, the techniques selected were Logistic Regression, Random Forest, SVM and Neural Network. All these techniques were used in previous researches on different datasets for identifying the customer churn problem. This research is adding value to the previous research.

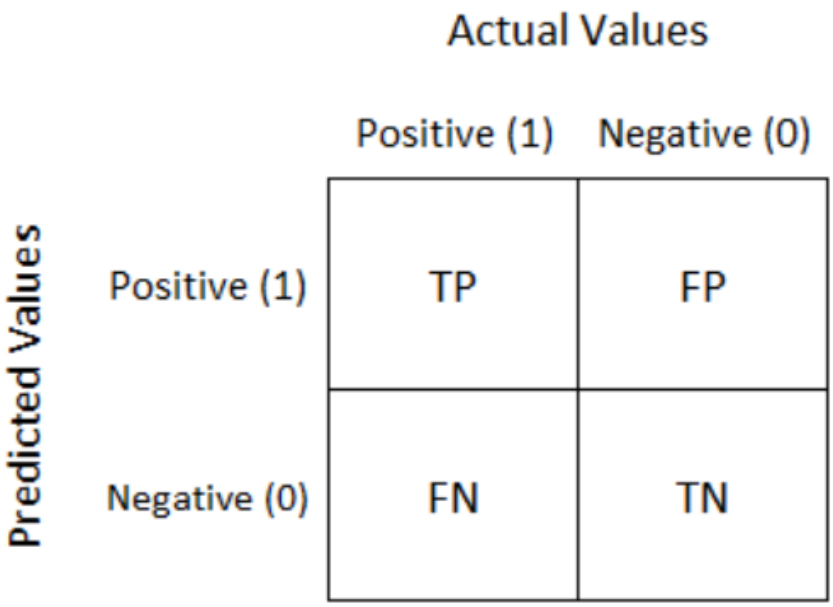
Here in this research for loop was used to divide the data randomly in train and test datasets. Each model was fitted every time on the same dataset. The accuracy score was appended and stored in a list every time the for loop was executed and thus different accuracy was found each time, we run the model. Finally, the average accuracy of each classifier was obtained and compared to identify the champion model. This approach was used so that each model was fitted to the same split of data which ensured that the model results can be compared. The for loop ensures that the results were generalised and that the split of data does not have an impact on the model performance.

3.5 Evaluation

Several measures will be used to evaluate the models created in this research. A confusion matrix was created for each model and then accuracy, precision, recall and specificity were calculated. The number of true positive (TP), true negative (TN), false positive (FP), false negative (FN) are obtained (Wieringa, Maiden, Mead & Rolland, 2006)

The following metrics can be calculated using a confusion matrix:

Confusion Matrix:



Where,

TP (True Positive): In TP, both predicted and actual values are True (1)

TN (True Negative): In TN, both predicted and actual values are False (0)

FP (False Positive): In FP, the actual value is false (0), but it is predicted as true (1)

FN (False Negative): In FN, the actual value is true (1), but it is predicted as false (0)

Accuracy: (TP+TN)/(TP+FP+TN+FN). The total number of correct predictions as churned or not churned customer, out of all the customers.

Specificity: TN/(TN+FP). The total number of customers correctly predicted as true negative, i.e., retained, out of all the customers.

Sensitivity/Recall: TP/(TP+FN). The total number of customers correctly predicted as true positive, i.e., churned, out of all the customers.

F1 score: F-measure also called as F1 score is the harmonic mean of precision and recall.

F1 score = (2 * Precision * recall) / (Precision + Recall)

ROC curve can also help in evaluating the model performance by graphical representation. It is a plot between True Positive Rate (TPR) and False Positive Rate (FPR).

3.6 Strengths and Limitation

In this section, the strength and limitation of the Design and Methodology are discussed in brief.

The feature selection was used which eliminates the irrelevant features and thus it helped in improving the performance of the model. This also helped to reduce the training time and avoid overfitting. Another strength was that in this research the customer's age, gender and area were also considered and they were the prominent predictors of identifying the customer churn. The main strength was that in this research CU members data were used to determine Customer Churn prediction and there was not much research performed in this area.

The main limitation of the research was that the data was very imbalanced and due to that the classifiers were more likely to be biased towards the majority class. SMOTE sampling technique was used to overcome this issue. Also, there were so many Date time datatype variables present in the dataset which were not taken into consideration in this research. A time-series model can be built for utilizing the Datetime data type variables. In this research, a single snap of data was used so it was difficult to build a time-series model for prediction. For time-series modelling different sets of data with the proper date was required.

4. IMPLEMENTATION AND RESULTS

This chapter outlines how the experiment was performed, based on the steps mentioned in chapter three. It includes all the steps of data pre-processing, how the machine learning models were created. This chapter also covers the details of evaluation measures of each supervised machine learning model.

The research was carried out using the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. It provided a uniform framework and guidelines for data miners. This methodology consists of six phases or stages – Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. The Business Understanding phase was covered in chapter three – Design and Methodology section.

4.1 Data Understanding

Proper planning and a deep understanding of data were required to execute machine learning research. This is the second phase of the CRISP-DM process which focuses on data collection, exploring data and checking the quality of data and to get insight from the data to generate a hypothesis.

The original customer data was collected from one of the CU in Ireland. The complete data analysis was done using statistical and visual analysis of the data using python.

4.1.1 Dataset

The CU members dataset contains 49 variables; which contained information of 96,967 members of CU who have joined the CU from the year 1911 to 2019. The target variable was ‘Closed’ which was a boolean variable that denotes whether a customer/member was retained or churned. Few of the variables are discussed in detail:

a) Age

Age is a numerical variable which denoted the age of the members of CU. Its range was from 0 to 119 with a mean of 45.23 and standard deviation of 21.35. This variable

was found to be normal as the skew and kurtosis were in the range of +/-2. Also, by looking into the histogram of Age variable it depicted the normal distribution curve.

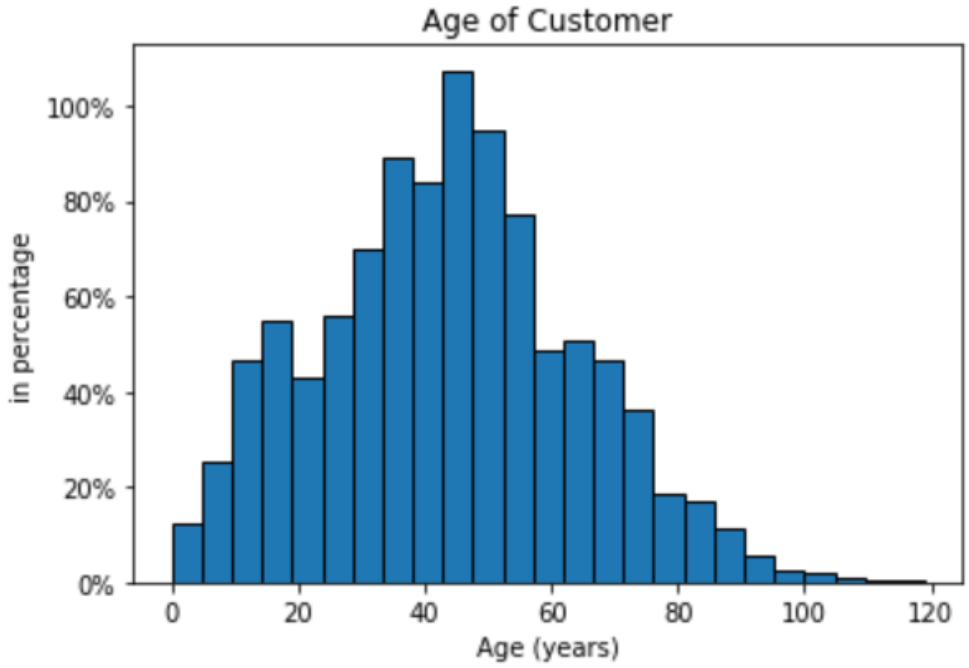


Figure 4.1: Age variable Histogram

b) AgeAtJoining

AgeAtJoining is a numerical variable which denotes the age of the members of CU while joining the institute. Its range was from 0 to 109 with a mean of 27.28 and a standard deviation of 17.85. This variable was normal as the skew and kurtosis were in the range of +/-2. Also, by looking into the histogram of AgeAtJoining variable it was slightly skewed, but we can assume the variable as normal because the skew and kurtosis were in the range of +/-2.

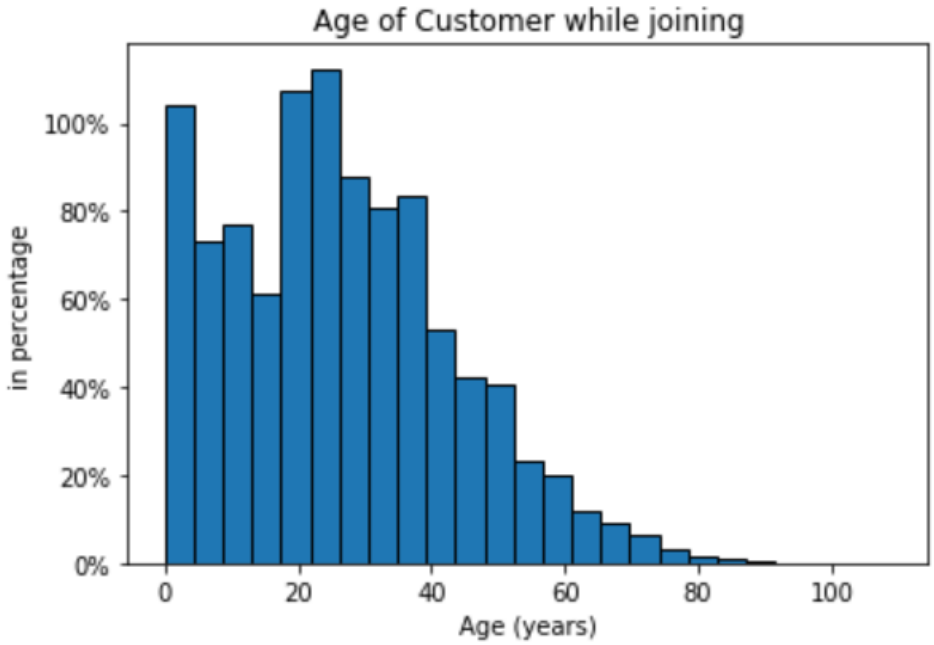


Figure 4.2: AgeAtJoining variable Histogram

c) TotalSavings

TotalSavings variable is again a numerical variable which denotes the total savings of the members of the CU. It ranged from 0 to 74,518.48 with a standard deviation of 4655.26 and a mean of 1764.33. By looking into the below histogram of the variable it was observed that it was left-skewed and the skew and kurtosis of the variable were not in the statistical range of between +/- 2. So, the variable was not considered as normally distributed and hence the standardization was performed on the variable to make it normal.

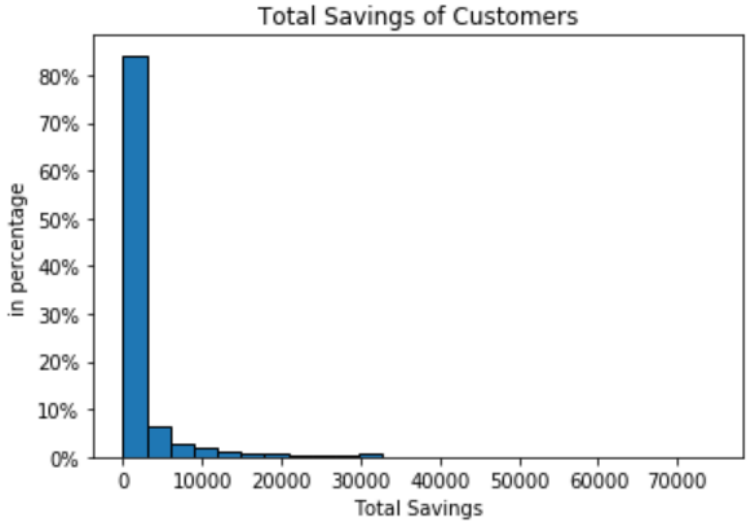


Figure 4.3: TotalSavings variable Histogram

d) TotalLoans

The TotalLoans is also a numerical variable which is highly skewed. It ranged from 0 to 282598.30 with mean 690.66 and with a standard deviation of 3783.11. The variable TotalLoans was highly skewed. By looking into the below histogram of the variable it was observed that it was left-skewed and the skew and kurtosis of the variable were not in the statistical range of between +/- 2. So, the variable was not considered as normally distributed and hence the standardization was performed on the variable to make it normal.

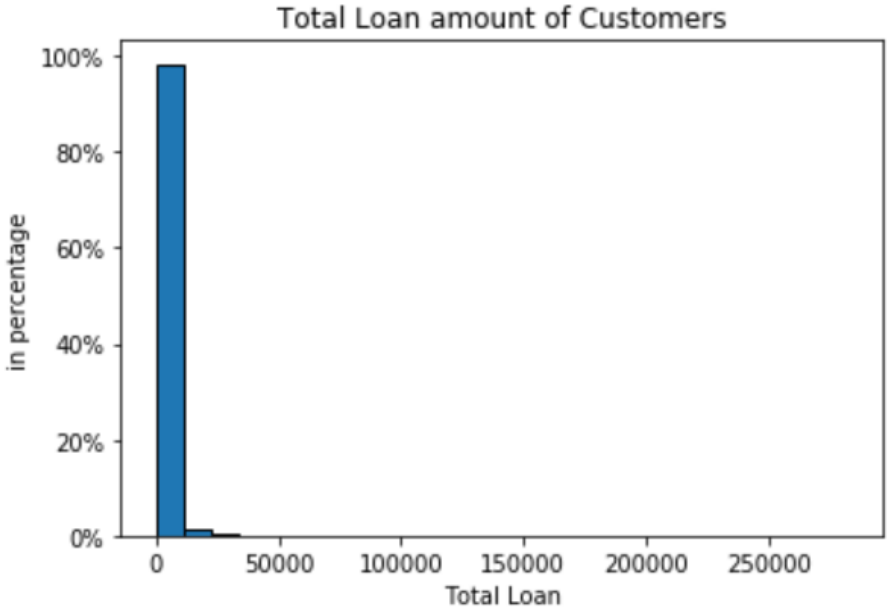


Figure 4.4: TotalLoans variable Histogram

e) Closed

The ‘Closed’ is the target variable with values as ‘True’ for those who have closed their membership with CU and ‘False’ for those which are still the members of CU. A closed account member were the members who were not deceased or dormant and they were not the member of CU. The data distribution is as shown below:

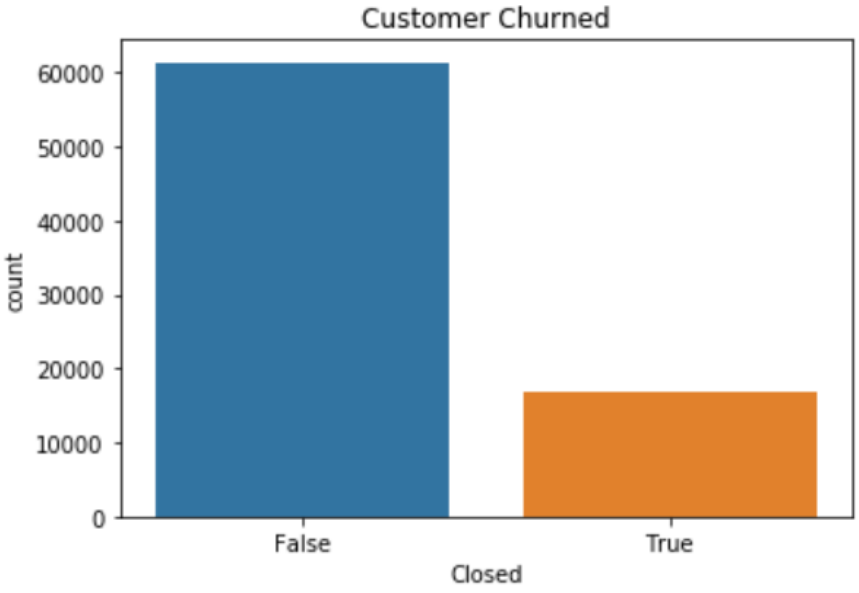


Figure 4.5: Closed Variable distribution

f) Gender

Gender is a categorical variable with two categories – male and female type of members in CU. The below graph depicted the distribution of gender in Churned and retained members.

From the below graph it can be seen that the females were more likely to leave the CU as compared to the males. Also, the institute has more females than males as the members of the CU financial institute.

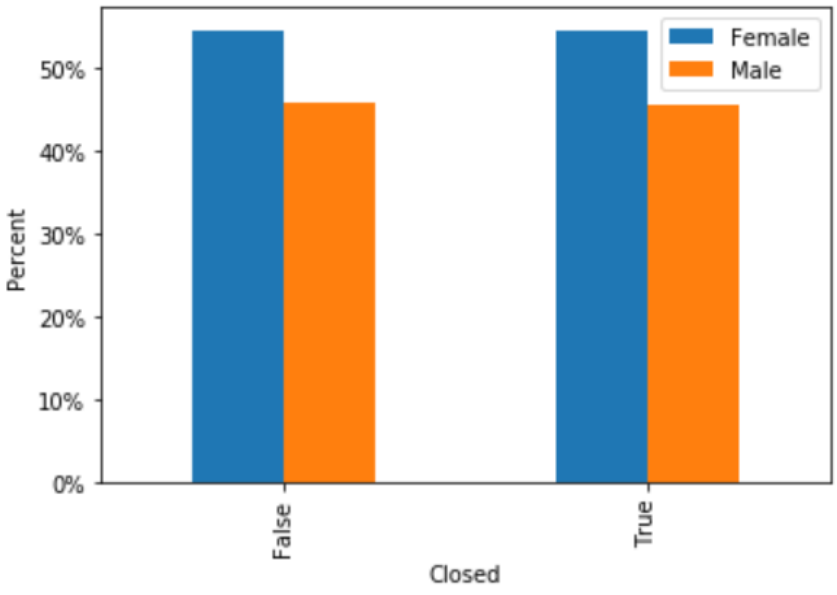


Figure 4.6: Gender Variable distribution

g) MaritalStatus

MaritalStatus is a categorical variable with multinomial categories. The below graph depicted the distribution of MaritalStatus in Churned and retained members.

From the below graph it could be seen that the married couple and singles were more likely to leave the CU as compared to the other category members. Also, the CU has more married couples as the members.

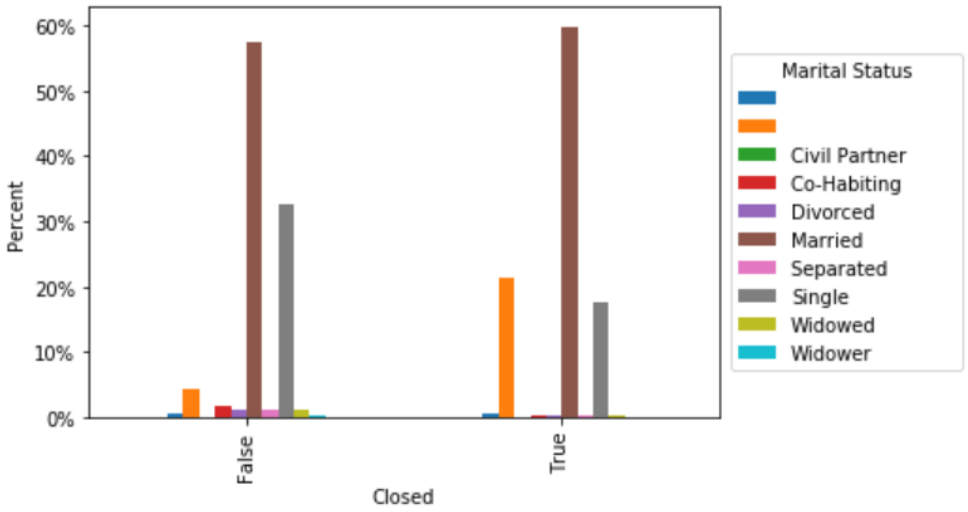


Figure 4.7: MaritalStatus Variable distribution

h) AccomodationType

AccomodationType is a categorical variable with multinomial categories. The below graph depicted the distribution of AccomodationType in Churned and retained members.

From the below graph it could be seen that the Home Owners were more likely to leave the CU as compared to the other category members. People who had mobile home accommodation type and people who had to rent accommodation type had an equal probability of leaving the CU. Also, the institution had more Home Owners accommodation type of people as the members of the CU.

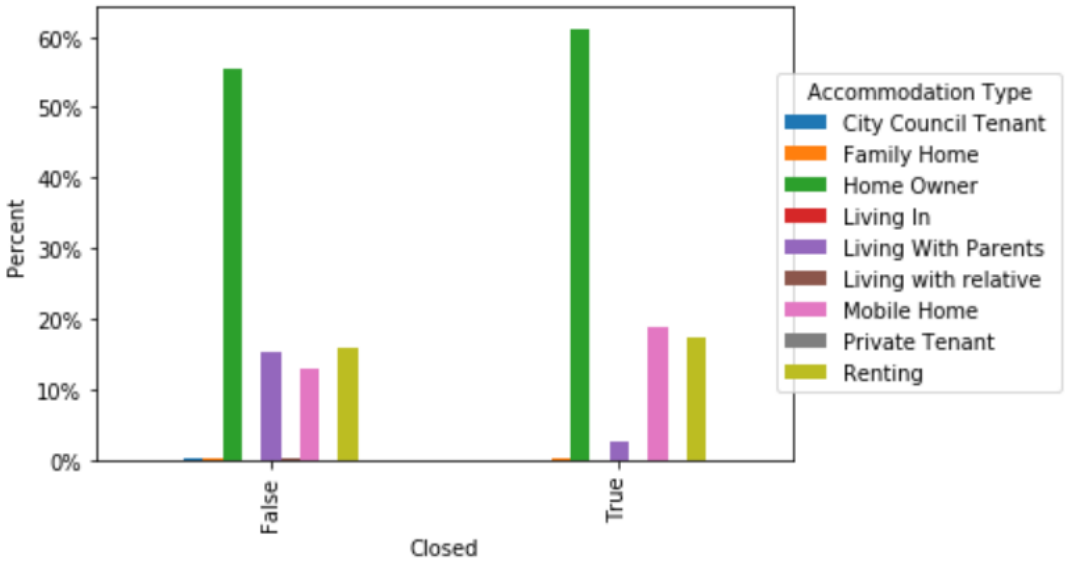


Figure 4.8: AccomodationType Variable distribution

i) PaymentMethod

PaymentMethod is a categorical variable with different categories of payment method options members prefer to do. The below graph depicted the distribution of PaymentMethod in Churned and retained members.

From the below graph it could be seen that the members who use Teller payment method were more likely to leave the CU as compared to the other category members. Also, the institute has more members who prefered Teller payment method rather than any other payment methods.

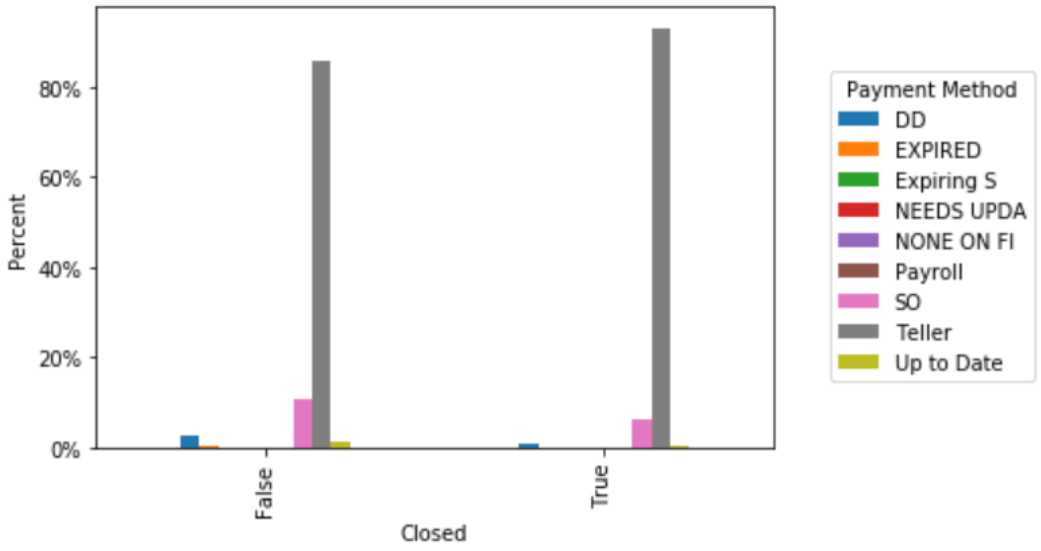


Figure 4.9: PaymentMethod Variable distribution

j) Dormant

Dormant is a categorical variable with two values – False and True. If Dormant is set as ‘True’ then that means that the customer is not active member of CU and if Dormant is set as ‘False’ that means the customer is an active member of CU. The below graph depicted the distribution of Dormant in Churned and retained members.

From the below graph it could be seen that nearly 10% of dormant accounts have closed their accounts from CU.

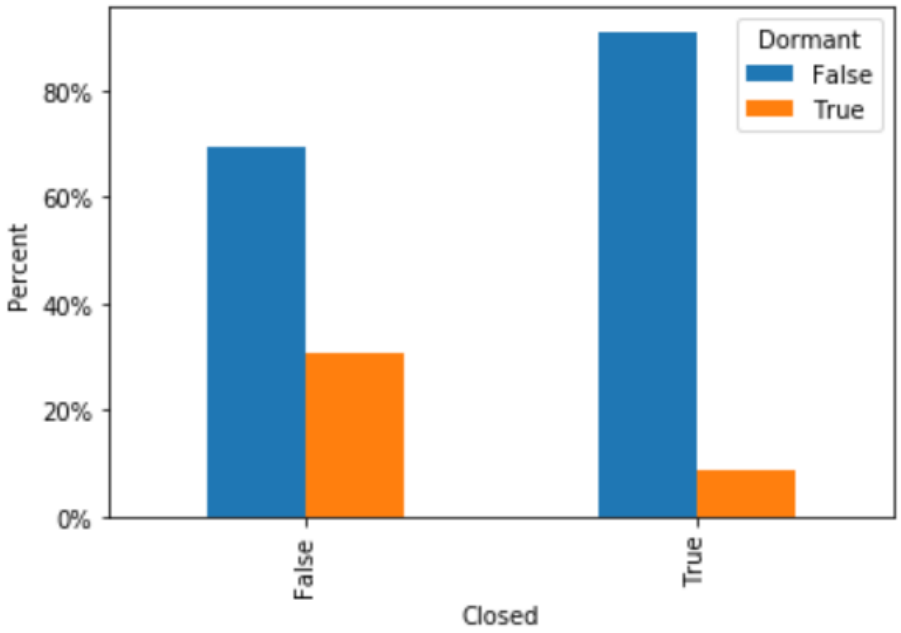


Figure 4.10: Dormant Variable distribution

To obtain a better overview of data, the statistical metrics like count, mean, standard deviation and measure of central tendency were calculated and shown in below table 4.1:

	Member_ID	Age	SecondACHolderAge	AgeAtJoining	TotalSavings	TotalLoans	NumberofLoansTaken	TotalValueofLoansTaken	SavingsAt
count	96967.000000	96967.000000	2492.000000	96966.000000	96967.000000	96967.000000	96967.000000	96967.000000	
mean	50031.469479	45.232553	47.846308	27.286276	1764.329845	690.660396	2.683614	7795.204333	
std	28831.067432	21.354460	15.575436	17.850147	4655.262112	3783.112886	5.760311	17481.672295	
min	3.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	25079.500000	30.000000	38.000000	14.000000	0.000000	0.000000	0.000000	0.000000	
50%	50130.000000	45.000000	45.000000	26.000000	12.850000	0.000000	0.000000	0.000000	
75%	74999.000000	59.000000	57.000000	38.000000	959.620000	0.000000	3.000000	7189.315000	
max	99998.000000	119.000000	119.000000	109.000000	74518.480000	282598.300000	250.000000	390245.860000	

Table 4.1: Descriptive Statistics of Customer data

From the above table, it was observed that the count is less than 96967 so this means there were missing values present. It could be seen that the mean for most of the data features have deviated from the standard deviation so standardization of data was needed before modeling.

4.1.2 Correlation Analysis

Correlation Analysis of the data was done to analyse the correlation between the independent and dependent variable and also to identify multicollinearity between the independent features. The 'Spearman' correlation method was used to identify correlation as the dissertation consists of both continuous and categorical variables. The correlation heatmap was generated as shown in the below figure and also the correlation matrix listed in table 4.2.

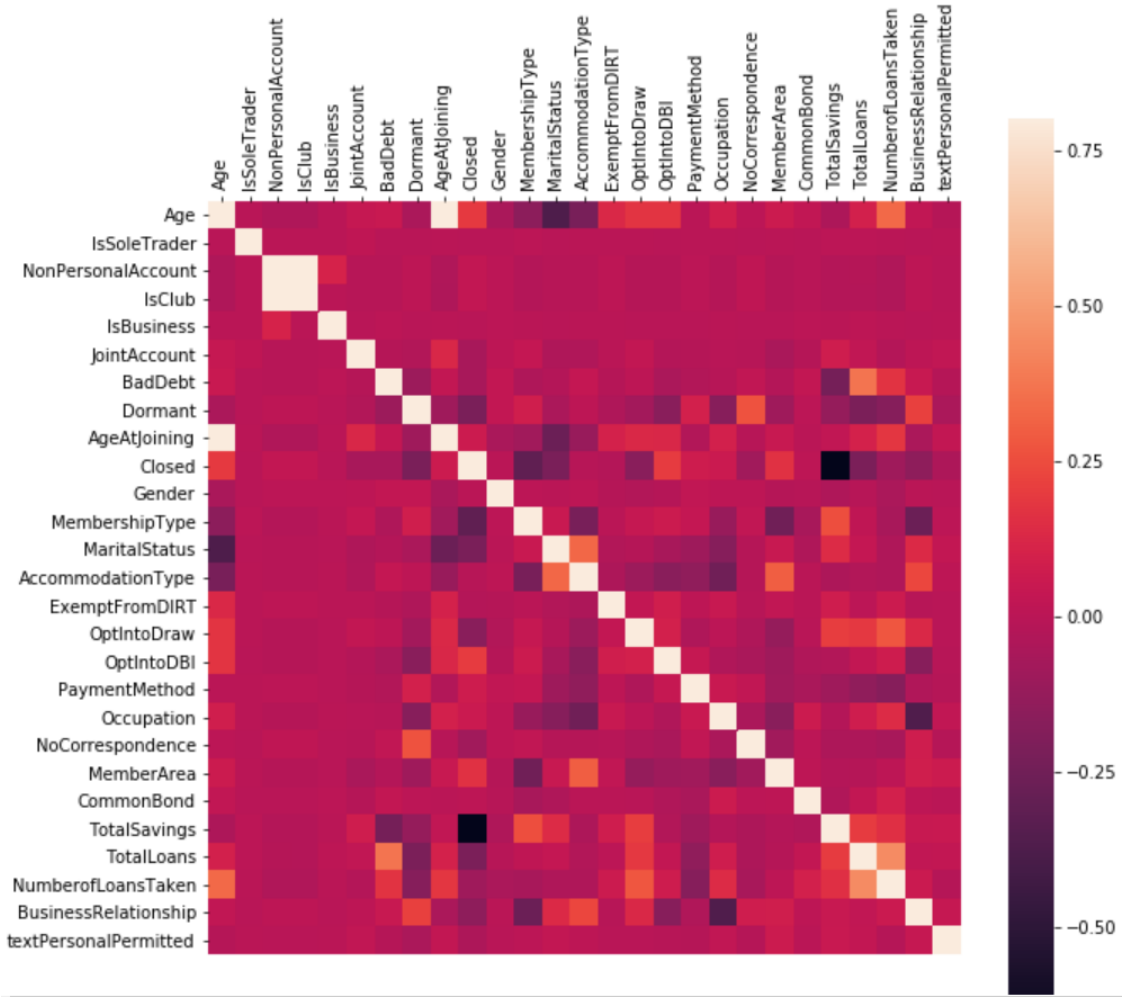


Figure 4.11: Correlation heatmap of the variables

	Age	IsSoleTrader	NonPersonalAccount	IsClub	IsBusiness	JointAccount	BadDebt	Dormant	AgeAtJoining	Closed	Gender
Age	1.000000	-0.003323	-0.040504	-0.041395	0.006081	0.042200	0.048434	-0.046481	0.832988	0.201814	-0.057887
IsSoleTrader	-0.003323	1.000000	-0.000472	-0.000470	-0.000051	0.021949	-0.001625	-0.005085	0.002061	-0.005064	-0.002999
NonPersonalAccount	-0.040504	-0.000472	1.000000	0.994220	0.107059	-0.008888	-0.008785	0.011192	-0.026039	0.030924	0.006235
IsClub	-0.041395	-0.000470	0.994220	1.000000	-0.000307	-0.008837	-0.009880	0.011616	-0.027186	0.031461	0.005591
IsBusiness	0.006081	-0.000051	0.107059	-0.000307	1.000000	-0.000952	0.009665	-0.003329	0.009228	-0.003315	0.006289
JointAccount	0.042200	0.021949	-0.008888	-0.008837	-0.000952	1.000000	-0.016769	-0.027015	0.122429	-0.062148	0.011288
BadDebt	0.048434	-0.001625	-0.008785	-0.009880	0.009665	-0.016769	1.000000	-0.102541	0.020125	-0.061707	0.030289
Dormant	-0.046481	-0.005085	0.011192	0.011616	-0.003329	-0.027015	-0.102541	1.000000	-0.094380	-0.215325	0.032406
AgeAtJoining	0.832988	0.002061	-0.026039	-0.027186	0.009228	0.122429	0.020125	-0.094380	1.000000	0.067109	-0.047426
Closed	0.201814	-0.005064	0.030924	0.031461	-0.003315	-0.062148	-0.061707	-0.215325	0.067109	1.000000	-0.000605
Gender	-0.057887	-0.002999	0.006235	0.005591	0.006289	0.011288	0.030289	0.032406	-0.047426	-0.000605	1.000000
MembershipType	-0.114607	0.004665	-0.001196	-0.001334	0.001211	0.040088	-0.010940	0.055932	-0.047901	-0.257144	0.007135
MaritalStatus	-0.278249	-0.001146	-0.011208	-0.011192	-0.000750	-0.010896	-0.004163	0.001489	-0.188057	-0.250235	0.000809
AccommodationType	-0.148706	-0.000304	-0.004812	-0.004894	0.000505	-0.025792	0.045284	0.022165	-0.070459	0.012977	0.004971
ExemptFromDIRT	0.145118	-0.000825	0.014880	0.015024	-0.000540	0.002170	-0.015360	-0.038138	0.106604	-0.020604	-0.017250
OptIntoDraw	0.161466	0.001406	-0.015394	-0.015281	-0.001873	0.033075	0.013033	-0.077970	0.120607	-0.165318	-0.026919
OptIntoDBI	0.178607	-0.002581	-0.015784	-0.015693	-0.001690	-0.020668	-0.051099	-0.167531	0.129033	0.199711	-0.012694
PaymentMethod	0.015015	0.000781	0.003836	0.005119	-0.011675	-0.021133	-0.018678	0.064480	-0.021746	0.068469	0.009523

Table 4.2: Correlation Matrix of the variables

It could be seen from the above correlation heatmap that the variables ‘TotalSavings’ was the most correlated variable with the target variable ‘Closed’. It is negatively correlated with the target variable. The other variables ‘OptIntoDBI’, ‘Age’, ‘MemberArea’ also had a strong correlation with the target variable ‘Closed’.

The variables ‘Age’ and ‘AgeAtJoining’ were positively correlated with each other. So AgeAtJoining variable was removed before modelling in order to avoid multicollinearity.

Thus finding the correlation also helped in feature selection.

4.1.3 Outlier Analysis

The boxplot for only ‘Age’ variable was plotted to find the outliers present in the data. As the data was highly imbalance, an analysis was done to identify if the outliers fall into minority class. Any outlier presented in minority class cannot be removed, as it could lead to information loss. The below boxplot and scatterplot created with respect to the target variable will give information on outliers.

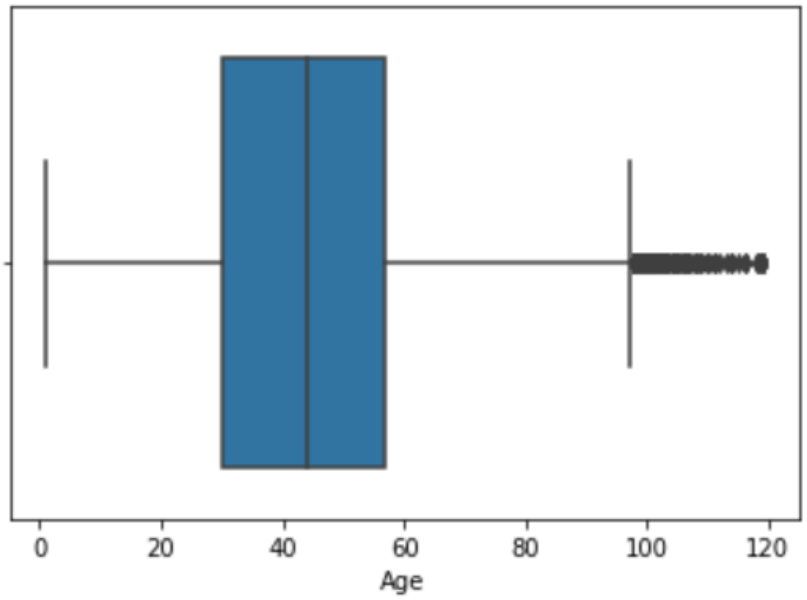


Figure 4.12: Boxplot of Age variable

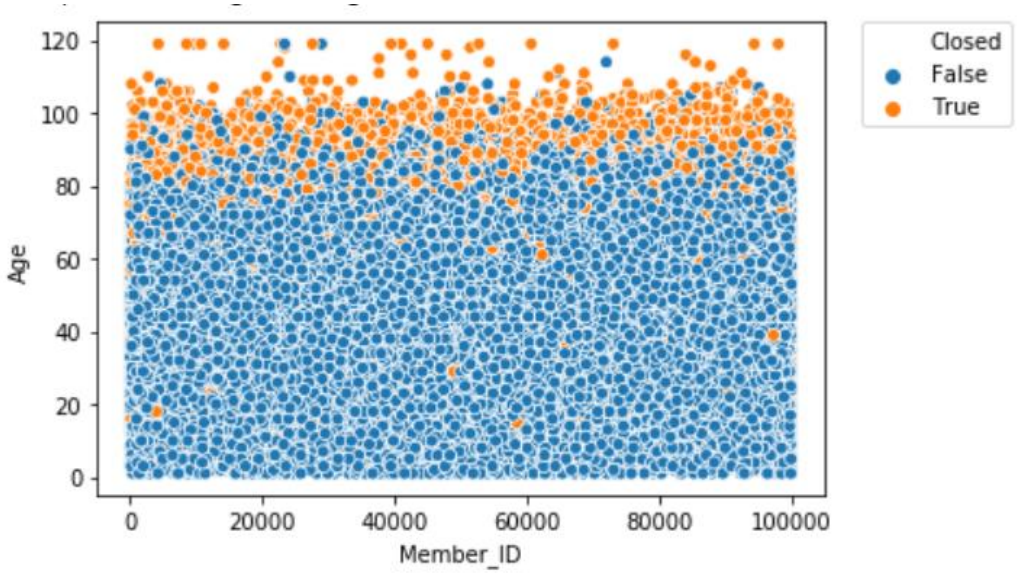


Figure 4.13: Scatterplot of Age variable with respect to the target variable

As seen in the above graphs it was observed that the outliers belong to the minority class, i.e., Closed = 'True'. So, in this case, the removal of outliers will lead to information loss hence the outliers were not removed from the data in this research.

4.2 Data Pre-processing

To prepare the final dataset for the experiment, several data pre-processing steps were carried out based on the understanding of data from the previous phase. The Data Pre-processing phase included handling missing values, Encoding, Normalizing the data, feature selection, feature extraction, standardizing the data and finally data splitting.

In CU customer dataset it has been observed that there were 3,138 records present with 'Deceased' value set as 'True', so all those records were dropped from the dataset using python drop command.

4.2.1 Handling Missing Values

In the Descriptive Analysis of Customer dataset, it has been observed that there were missing values present in the dataset.

In this research there were few variables – 'SecondACHolderAge','SecondACHolderGender','SavingsAttachedAsMember', 'TextMarketingUpdateDate' and 'textPersonalPermittedDate' with more than 60% of values missing and these were removed from the final dataset (Kelleher, Mac Namee, & D'Arcy, 2015).

The missing values in categorical variables were imputed using the maximum likelihood and last observation carried forward techniques were the most common techniques for imputing the missing values (Kang, 2013). For imputing the missing values in categorical variables – 'PaymentMethod', 'CommonBond', 'AccommodationType' the techniques like bfill, ffill were used. The bfill is the backward fill and ffill is the forward fill methods used in python for imputing missing values in the dataset. In ffill propagated the last observed non-null value forward and bfill propagated the first observed non-null value backwards. For 'MemberArea' variable the missing values were imputed by the most frequently occurred value.

For continuous variables with missing values from 2% till 30% the values were imputed by mean values. Mean was a reasonable estimate for randomly selected observations from a normal distribution (Kang, 2013). The records with missing 'Gender' variable values were dropped from the final dataset as it was difficult to impute the gender missing values.

4.2.2 Normalizing the Data

The skew and kurtosis of the continuous variables were determined. If the value of skew and kurtosis lied outside the range of +/-2 then the variable was not normally distributed. Even the histogram plot can determine whether the variables were normally distributed or not. In the CU dataset, it has been observed that the variable 'Age', 'AgeAtJoining' were normally distributed whereas the variables 'TotalSavings', 'TotalLoans' were highly skewed.

The data can be normalized by applying sklearn MixMaxScaler() function on the particular variables of the dataset to normalize the data.

4.2.3 Feature Selection

In machine learning and statistics, feature selection is the process of selecting relevant features for predicting the output and useful in model construction. The feature selection is also known as variable selection, attribute selection or variable subset selection. Correlation method was used here in this research to identify the best predicting features to predict the customer churn.

A correlation matrix was used to find the correlation between the independent and dependent variables and the multicollinearity issue was detected as well as the most relevant feature in predicting the output was also detected.

Another method based on previous research (Khan, et.al., 2015) was used to find the top predicting features were also identified using sklearn.ensemble ExtraTreeClassifier function to identify the top predicting features. This method was used to know the feature importance ranking for predicting the target. A graph of feature importance was plotted as shown below:

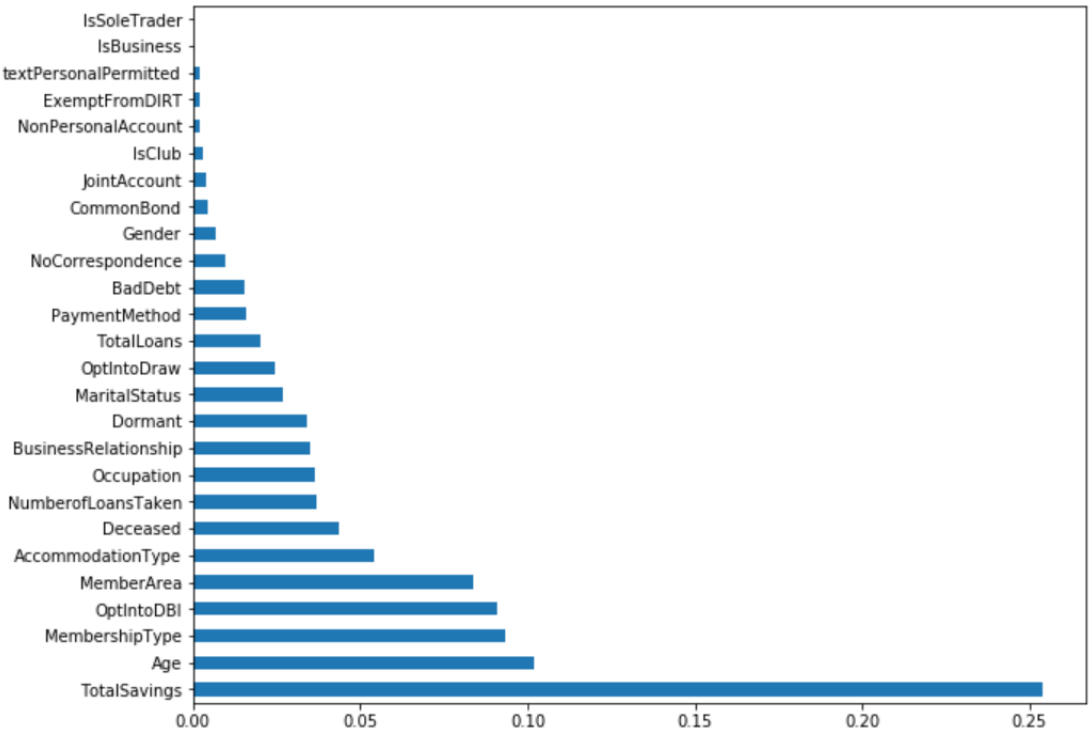


Figure 4.14: Feature Importance graph with respect to the target variable

As seen from the above graph the most important feature in predicting the output was 'TotalSavings' followed by 'Age', 'MembershipType' and so on.

4.2.4 Encoding

Encoding means converting categorical data into numeric data. A categorical datatype was the most common type of data present in the dataset. These variables were normally stored as text values. Machine learning algorithms work only on numeric data as they are based on mathematical equations. So, it was impossible to keep the categorical variables as it was and there was a need to convert these categorical variables into numeric form.

In this research the Label Encoding method was used to encode the categorical variables. In this method, each category was mapped to a number. Using this method will not increase the number of columns and thus will not slow down the learning process. In this research 23 categorical variables were encoded using label encoder. The variables 'Gender', 'Closed', 'IsSoleTrader', 'Dormant', 'NonPersonalAccount', 'IsClub', 'IsBusiness', 'JointAccount', 'BadDebt', 'Deceased', 'MembershipType', 'MaritalStatus', 'AccommodationType', 'ExemptFromDIRT', 'OptIntoDraw',

'OptIntoDBI', 'PaymentMethod', 'Occupation', 'NoCorrespondence', 'MemberArea', 'CommonBond', 'BusinessRelationship', 'textPersonalPermitted' were encoded into numeric form using Label Encoder.

4.2.5 Sampling

The most common data problem is Class Imbalance. The class imbalance can be seen in the below figure and the distribution table of 'Closed' variable.

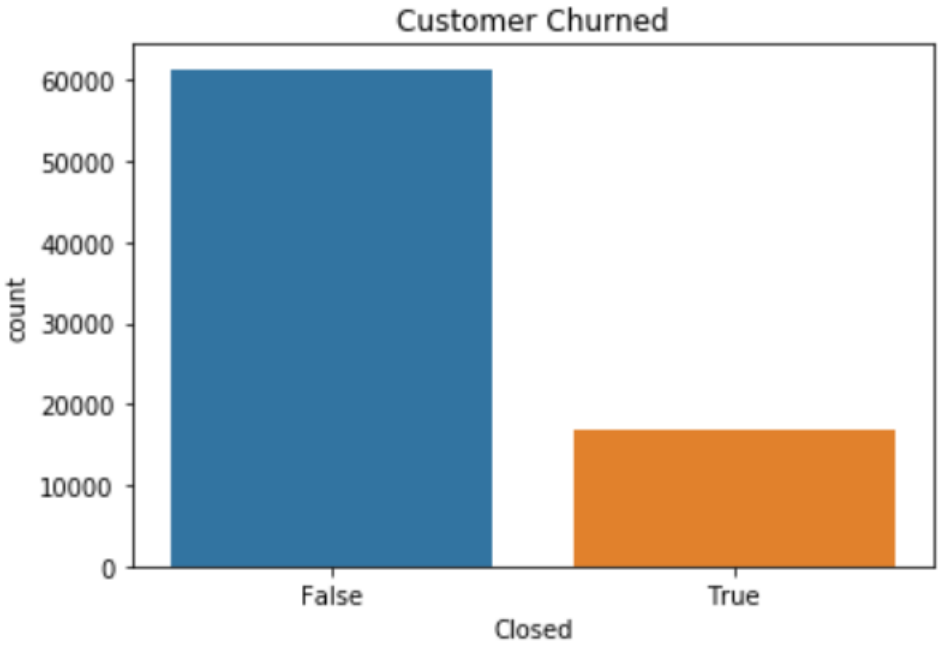


Figure 4.15: Target variable distribution

Closed	
FALSE	TRUE
61391	16769

Table 4.3: Target Variable Counts

The Synthetic Minority Over-Sampling Technique (SMOTE) was used in this research to resolve the class imbalance problem based on the previous research (Ali & Ariturk, 2014). In the SMOTE technique, it synthesised new minority instances between the existing minority instances. Using this function, the data will be balanced. The dataset

was imbalanced with the class imbalance ratio of approximately 18:5 which means against 18 non churned customers 5 will be churned.

The SMOTE() function is imported from imblearn.over_sampling.

4.2.6 Data Splitting

The final dataset thus obtained was then split into train and test dataset. The final dataset obtained after handling missing values, normalizing dataset, data sampling is shown below in the below table.

Variable	Description	Type
Age	Age of the member	Float
IsSoleTrader	Member is Sole Trader	Integer
NonPersonalAccount	Not Personal Account	Integer
IsClub	Is Club	Integer
IsBusiness	Member has business	Integer
JointAccount	Whether it is joint account	Integer
BadDebt	Has Bad Debt	Integer
Dormant	Account is dormant	Integer
Closed	Account is closed	Integer
Deceased	Account is deceased	Integer
Gender	Gendeer of member	Integer
MembershipType	Type of membership	Integer
MaritalStatus	Marital Status of member	Integer
AccommodationType	Accomodation Type of member	Integer
ExemptFromDIRT	Exempt from DIRT	Integer
OptIntoDraw	Opt into Draw	Integer
OptIntoDBI	Opt Into DBI	Integer
PaymentMethod	Payment Method used by member	Integer
Occupation	Occupation of member	Integer
NoCorrespondence	No Correspondence	Integer
MemberArea	Area of member	Integer
CommonBond	Common Bond	Integer
TotalSavings	Total Savings of member	Float
TotalLoans	Total loan amount of member	Float
NumberofLoansTaken	Number of loans taken by member	Integer
BusinessRelationship	Business Relationship	Integer
textPersonalPermitted	Personal text permission	Integer

Table 4.4: Final Dataset Description

In this research, the final dataset was splited into 80% training and 20% test data sets based on the existing researches (Shaaban, Helmy, Khedr & Nasr, 2012). Using for loop along with sklearn train_test_split function the data was splited into train and test

datasets 40 times. Each time the loop runs the data splited randomly and models were created and different accuracies were stored in the list. The supervised machine learning models were compared based on their accuracy score and the champion model was chosen to predict the customer churn. The data was splited, and the same split was used for all the models to fit the model and to calculate the score thus it doesn't affect the accuracy result of the models.

The experiment was performed with an imbalanced dataset and balanced dataset (Sampling using SMOTE technique) for better result comparison.

4.3 Modelling

The four proposed supervised machine learning models were created to test which best predicted the customer churn.

4.3.1 Logistic Regression

The Logistic Regression model was built using LogisticRegression function imported using sklearn.linear_model class in python. Sklearn or Scikit Learn is an open-source Machine Learning library for python. It provides many supervised and unsupervised learning algorithms.

Only significant variables were included and all insignificant parameters were excluded from the model and the model was trained and tested using 27 parameters. The same set of features with default parameter settings were used in all the supervised machine learning algorithms.

The following results were obtained from the logistic regression.

Evaluation Measures	Values
Accuracy	85%
Precision	61%
Recall	87%
Specificity	85%

Table 4.5: Logistic Regression Results for a balanced dataset

Evaluation Measures	Values

Accuracy	87%
Precision	79%
Recall	60%
Specificity	95%

Table 4.6: Logistic Regression Results for imbalance dataset

4.3.2 Random Forest

The Random Forest was built in python using sklearn.ensemble class. Random Forest algorithm is based on ensemble learning. Ensemble Learning uses multiple machine learning models to make better predictions on a dataset.

In this research the 'gini' criterion was selected which was the by default function to measure the quality of split, n_estimators was chosen as 100 that means 100 random decision trees were ensembled together to build the Random Forest. The max_depth was selected as 10 which means the tree can expand till the maximum depth of 10. Finally, the class_weight was set as 'balanced' which depicted the weight associated with the classes. In Random Forest by default, the weight is inversely proportional to the frequency the class appears in the data.

The following results were obtained from the Random Forest.

Evaluation Measures	Values
Accuracy	96%
Precision	88%
Recall	98%
Specificity	96%

Table 4.7: Random Forest Results for a balanced dataset

Evaluation Measures	Values
Accuracy	97%
Precision	91%
Recall	98%
Specificity	97%

Table 4.8: Random Forest Results for imbalance dataset

4.3.3 Support Vector Machine

The SVM model was built using svm function imported from sklearn in python. In this research for SVM model 'linear' kernel was selected based on the previous customer churn research (Cao & Shao, 2008). In the Support Vector Machine model, kernel function is the most important parameter. In this research as the kernel function was set as linear and it will separate the class linearly using a single line. It is useful when the dataset is large. The main advantage of the linear kernel was fast processing.

The following results were obtained from the Support Vector Machine.

Evaluation Measures	Values
Accuracy	86%
Precision	62%
Recall	91%
Specificity	84%

Table 4.9: Support Vector Machine Results for a balanced dataset

Evaluation Measures	Values
Accuracy	89%
Precision	81%
Recall	65%
Specificity	95%

Table 4.10: Support Vector Machine Results for imbalance dataset

4.3.4 Neural Network

Here in this research, keras[5] sequential model was used to build the neural network in Python. Keras is a user-friendly neural network library in Python. The Keras sequential model is a linear stack of layers. Here two hidden layers were added in this model. The 'Dense' function implements the following operation:

output = activation (dot(input, kernel)+ bias)

[5] https://keras.io/getting-started/faq/%22%20%5Cl%20%22how-should-i-cite-keras

where activation is the activation function chosen in the hidden layer. Here ‘relu’ activation function was used which means Rectified Linear Unit. It was less computationally expensive as it involved a simpler mathematical calculation. Mathematically, it is represented as:

$$y = \max(0,x)$$

In the output layer, ‘sigmoid’ activation function was used as by default settings.

The following results were obtained from the Neural Network.

Evaluation Measures	Values
Accuracy	91%
Precision	72%
Recall	98%
Specificity	89%

Table 4.11: Neural Network Results for a balanced dataset

Evaluation Measures	Values
Accuracy	91%
Precision	72%
Recall	98%
Specificity	89%

Table 4.12: Neural Network Results for imbalance dataset

The confusion matrix plot was made for all the four models using sklearn.metrics confusion_matrix function in python. The confusion matrix was often used to determine the performance of the model. It was used to calculate the accuracy, precision, specificity and Recall measures of the model in the classification problem.

The below graph represents the confusion matrix of all the four models.

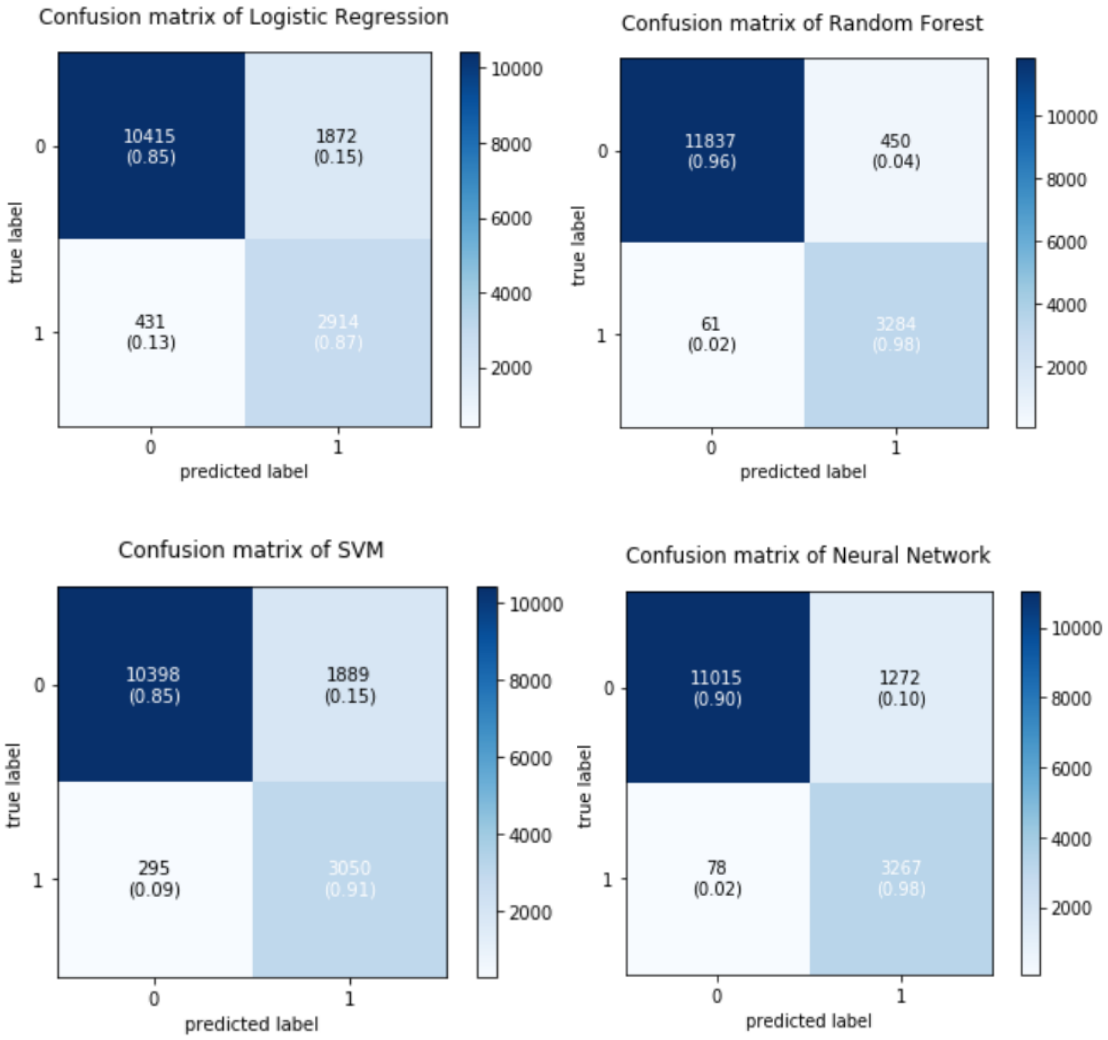


Figure 4.16: Confusion Matrix

4.4 Results

The following results in Table 4.13 were obtained from each supervised machine learning technique used.

Evaluation Metric	Logistic Regression	Random Forest	SVM	Neural Network
Accuracy	85%	96%	86%	91%
Precision	61%	88%	62%	72%
Recall	87%	98%	91%	98%
Specificity	85%	96%	84%	89%

Table 4.13: Results of Supervised Machine Learning Models

The result in Table 4.13 was obtained from the models built and tested on the dataset in which the SMOTE technique was used to balance the data. This technique was the data

level algorithm to balance the dataset. The CU customers/members dataset was highly imbalanced, so SMOTE technique was used to balance the dataset before building the models. The same experiment was repeated with an imbalanced dataset. For building the models the train and test dataset both were imbalanced, and the result obtained are tabulated in below Table 4.14.

Evaluation Metric	Logistic Regression	Random Forest	SVM	Neural Network
Accuracy	87%	97%	89%	92%
Precision	79%	91%	81%	90%
Recall	60%	98%	65%	79%
Specificity	95%	97%	95%	97%

Table 4.14: Results of Supervised Machine Learning Models with imbalanced dataset

The results with imbalanced dataset were better than the results after applying the SMOTE sampling technique. This is contrary to what was suggested in research by Guo, X. (2008).

Hence based on this comparison, the result in table 4.14 will be the focus of the analysis, evaluation, discussion and conclusion that follows.

4.5 Secondary Research

The above result obtained in this research was compared with the results of the existing research paper by Kumar & Vadlamani (2008) on Credit card customer churn of a Latin American bank. The secondary research was performed to compare the results with the other domain datasets and to comprehend how the results vary from dataset to dataset and based on different methodologies.

In the existing research on credit card, customer churn on Latin American bank data set the data was splited mainly using two techniques – Hold out method and Tenfold cross- validation method. The customer data was highly imbalanced. So for each type split the models were built. The models were built for Original data, SMOTE data, Undersampled data, Oversampled data and Combination of undersampling and oversampling data for hold-out and Tenfold cross-validation techniques and results were recorded. Multilayer Perceptron, Logistic Regression, Decision Tree (J48), Random Forest, RBF Network and SVM were built and majority voting ensemble

system was used to determine the best model. It was observed that among various methods tested, the results show that tenfold cross-validation method on SMOTED data has yielded excellent results with 92.37% sensitivity, 91.40% specificity and 91.90% overall accuracy. In the existing research, the SMOTED dataset accuracy was better than the original dataset whereas in this research the unbalanced dataset has produced better accuracy results when compared to SMOTED dataset.

Also, Random Forest has produced excellent results in the existing research which was alike to the results of this research where the random forest has yielded better results when compared to other algorithms such as Logistic Regression, SVM and Neural Network.

5. EVALUATION AND DISCUSSION

This section analyses the results obtained from the experiments done in the research. This chapter evaluates the predictive power of the supervised machine learning models built in chapter 4. Each model was compared by its accuracy. A comparison was also performed with the results obtained with an imbalanced dataset and sampled data using SMOTE technique.

5.1 Evaluation of the Results

An experiment was performed to find the best supervised machine learning model in predicting the customer churn for the CU customer churn dataset. In this experiment, the same set of experiment was performed twice one with the imbalanced dataset and the other with balanced dataset using SMOTE sampling technique.

It has been observed that the supervised machine learning algorithms performed better with imbalance dataset in terms of accuracy, precision and specificity as the evaluation metrics. In terms of recall as the evaluation metrics, better results were produced with the sampled dataset as compared to an imbalanced dataset. Imbalance dataset leads to high accuracy as most of the data belongs to one class. The results were biased towards the majority class. Once the data was balanced using sampling techniques the accuracy will be slightly reduced and the recall percentage will be increased as it will balance the data with both the class values.

In both the experiments, performed in terms of all the evaluation metrics the Random Forest machine learning algorithm has outperformed the Logistic Regression, SVM and Neural Network models.

The below graph was a comparison graph based on the accuracy measure of the supervised machine learning algorithms used in this research – Logistic Regression, Random Forest, SVM and Neural Network.

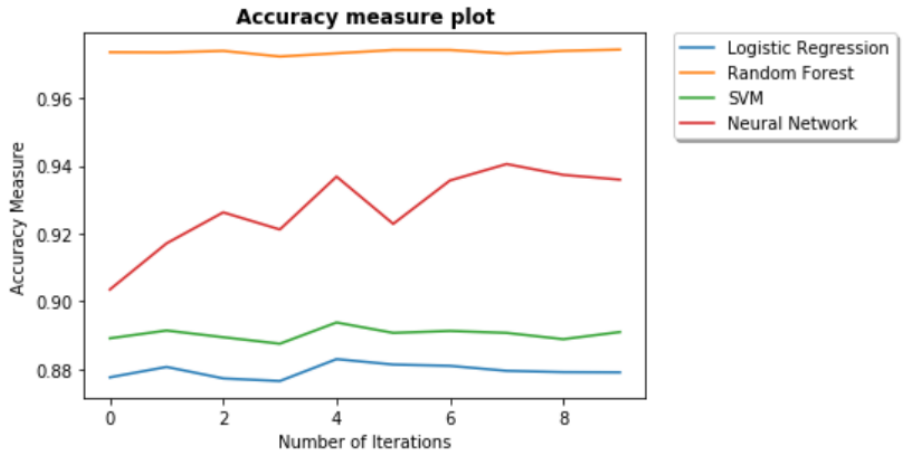


Figure 5.1: Accuracy Comparison graph

The Receiver Operating Characteristic (ROC) curve was also plotted for the best supervised model Random Forest in predicting the customer churn. The Area Under Curve was calculated as 0.93 which is quite close to 1.

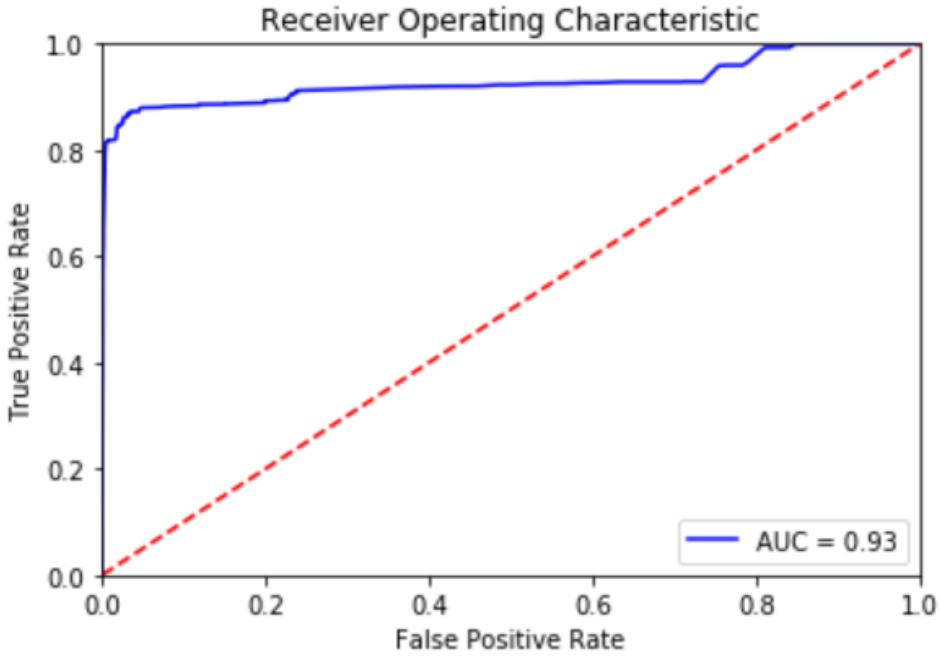


Figure 5.2: ROC graph

The ROC curve is a plot between true positive rate (Sensitivity) and false positive rate (Specificity) for different cut-off points of a parameter. The area under the ROC curve (AUC) is a measure of how well parameters can distinguish between churned and not churned groups. The AUC is better if it is close to 1.

Here in this above plot, it could be seen that AUC was equal to 0.93 for the Random Forest model which implies that this model is quite good in predicting the customer churn for CU customer dataset.

5.2 Hypothesis Evaluation

As seen in the above chapter Implementation and Results Random Forest model has the highest accuracy of 97%, recall as 98%, precision as 91% and specificity 97% so this model was chosen as the best model for predicting customer churn in this dataset.

In much of the literature reviewed for this research, in few of the papers, it was suggested that Support Vector Machine has provided better results. In one of the research Vafeiadis, T (2016) Support Vector Machine has performed better than the Logistic Regression and Neural Network.

In another research focused (Xie, Li, Nagi & Ying, 2008) on bank real dataset, the Random Forest has outperformed Support Vector Machine. This research was focused on a similar kind of dataset with class imbalance.

In a few of the financial bank-related dataset, only Neural Network was used to determine customer churn prediction.

Thus the hypothesis, *a random forest supervised machine learning model build using the CU customer data, will achieve high accuracy (97%) than the other supervised machine learning algorithms like Logistic Regression, Support Vector Machine and Neural Network, to predict the customer churn* is accepted.

In both the experiments performed the Random Forest was the best model because of the best parameters chosen for the model building. 'class_weight' as balanced was the most important feature in building the random forest and due to that the Random Forest was the best classifier for predicting customer churn.

5.3 Strengths of the Research

The main strength of the research was its ability to precisely identify the customer churn. The results suggest that the Random Forest model was the best predictor of CU customer data when compared to Logistic Regression, SVM and Neural Network. Another strength in this research was that the customer's age, gender and area were

also considered, and they were the prominent predictors of identifying the customer churn. The main strength of the research was CU members data was used to determine Customer Churn prediction and there was not much research performed in this area. Feature selection was used in this research which has increased the accuracy of the models.

5.4 Limitations of the Research

The main limitation of the research was that the data was very imbalanced and due to that the classifiers were more likely to be biased towards the majority class. Supervised machine learning models have performed well with an imbalanced dataset as compared to the balanced dataset. Also, there were so many Date time datatype variables present in the dataset which were not taken into consideration in this research. Time series model was not supported in this dataset. Another limitation was that the customers' data of only one CU was used for this research so it cannot be the representative of the other CU financial institutions. The customer base would be different for different CU institutions.

6. CONCLUSION

This chapter gives an overview of the research carried out. It summarises the results of the experiment performed in predicting customer churn. The chapter summarises the outcome of our research and derives proper interpretation from them. It summarises the finding with respect to the research question which was set at the beginning of the research: *"Which supervised machine learning: Logistic regression, Random forest, SVM or Neural network; can best predict the customer churn of CU with the best accuracy, specificity, precision and recall?"*

6.1 Research Overview

The goal of this research was to examine the predictive power of Supervised Machine learning algorithms on CU customer dataset in predicting customer churn. CU is a financial institution which is owned by its members and it is growing because of its reasonable rate of interest, as discussed in the literature review. The four Supervised machine learning algorithms were examined, Logistic Regression, Random Forest, Support Vector Machine and Neural Network to predict whether the member will be churned or retained with the institute. These four models were chosen for this research based on previous research.

The supervised machine learning models aimed to predict whether the customer of CU will churn or not. Many previous papers were dealt with the customer churn problem of financial institutions like the bank, telecom industry. The papers reviewed for this project did not cover CU customer data for churn prediction.

The main objective was to identify the supervised machine learning model with the best accuracy in predicting the customer churn. Chapter two described previous research carried out in this area, the various techniques and approaches applied to solve the problem. Chapter three detailed the method and design approach adopted in the current research to solve the problem. Chapter four outlined the implementation of the models. Chapter five outlined the result analysis of the four supervised machine learning models and compare their performance. The accuracy measure was considered as the evaluation metrics to get the best model. It was found that the Random Forest

technique outperformed the other algorithms for both the experiments performed one with imbalance dataset and others with balanced dataset using SMOTE technique. In this research, it was found that all the models have performed better in the imbalanced dataset on the contrary to the previous research on imbalance dataset. Therefore, the alternative hypothesis was accepted that *a random forest supervised machine learning model build using the CU customer data, will achieve high accuracy (97%) than the other supervised machine learning algorithms like Logistic Regression, Support Vector Machine and Neural Network, to predict the customer churn.*

6.2 Problem Definition

Customer Churn calculation and monitoring are very important in all sectors of an industry because it is far cheaper to retain old customers than to acquire new ones. CU is a financial institution owned by its members so churn prediction will be helpful for them to try to retain their existing members.

The literature review confirmed that many supervised machine learning techniques have been evaluated in the research area to predict customer churn. The SVM and Random Forest techniques were seen to be performed with good results for customer churn prediction in previous research.

This research aimed to determine which supervised machine learning algorithm would be best in predicting the customer churn on CU member dataset.

Currently, the CU has not adopted any techniques to identify the members who were likely to leave the institution. Adopting machine learning technique in building and evaluating the supervised machine learning techniques for CU member dataset has contributed to gain further insight into the members and helps the CUs to know the churn prediction.

6.3 Design, Evaluation and Results

The design of this project mostly followed the CRISP-DM methodology. As mentioned previously, the data was provided by one of the CU financial institutions. It contained information about their customer base.

The first step was data exploration, data cleaning. Variables with missing values more than 60% were not considered in the final dataset. Variables with 2% to 30% missing values were imputed with mean values for continuous variables and mode value for categorical variables. Feature Selection was also performed using a correlation matrix and using Extra tree classifier algorithm. The normal distribution of the continuous variables was also identified using the histogram and by measuring skew and kurtosis. Label encoding was performed as for machine learning models string data type was not accepted.

Finally, the data was divided into train and test data sets with 80% of data as training and the remaining 20% as test datasets. Four supervised machine learning models Logistic Regression, Random Forest, SVM and Neural Network were built on the training data set. All the models were trained on same train dataset picked up randomly and iterated for 40 times using for loop. They were tested on the test dataset. These models were evaluated based on accuracy as the evaluation metrics. A ROC curve was also plotted for the best model. In this research, Random Forest was the best in predicting the customer churn for CU dataset.

This experiment was performed twice one for imbalanced dataset and the other for a balanced dataset using SMOTE sampling technique.

The Random Forest model provided the highest accuracy of 97% in imbalance dataset and 96% accuracy in a balanced dataset. Even the precision and recall percentage was better for the random forest as compared to other models. All models performed with better accuracy on imbalanced dataset instead of a balanced dataset in this research. However, this was not true in the previous research papers.

6.4 Contributions and Impact

Most of the customer churn prediction literature were performed on telecom or bank or app dataset. In this research, the CU dataset was used which is unique. There is not much research done on CU dataset to identify the customer churn prediction. The research about CU churn prediction contributes literature for future research.

From the business point of view, it could be helpful for the institution to know the most likely to leave members. They can increase customer retention. Moreover, retaining an

old customer is more beneficial financially than getting new customers. Also, in CU as the members own the institute so it is hard for the institute to get the trustworthy members, so it is important for the CU to retain their old customers.

6.5 Future Work and Recommendations

Some future work identified throughout the project, which may be carried out. Here in this research only one branch of the dataset was explored and analysed. In future, another branch of CU dataset can be explored. Further research is needed to handle datetime data type variables.

The four machine learning techniques were used in this project on the CU dataset. Further other techniques can be explored as well. Different machine learning algorithms can be explored, and data can be analysed.

Further research can be done to build the time-series model to predict customer churn.

Also, there is a scope of using clustering unsupervised machine learning technique to examine the data. The similarities in data or some patterns can be determined using this technique.

BIBLIOGRAPHY

Ahmed, A., Maheshware, D. (2017). Churn prediction on huge telecom data using hybrid firefly based classification. *Egyptian Informatics Journal,18(3)* , 215-220. doi.org/10.1016/j.eij.2017.02.002

Ali, O., Ariturk,U. (2014). Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. *Expert Systems with Applications, 41(17).*7889-7903. doi.org/10.1016/j.eswa.2014.06.018

Aliyu, A., Kasim, R., Martin, D. (2011). Impact of Violent Ethno-Religious Conflicts on Residential Property Value Determination in Jos Metropolis of Northern Nigeria: Theoretical Perspectives and Empirical Findings. *Modern Applied Science, 5(5),* 171-183. doi:10.5539/mas.v5n5p171

Alwis, P., Kumara, B., Hapuarachchi, H. (2018). Customer Churn Analysis and Prediction in Telecommunication for Decision Making. *International Conference on Business Innovation.* 40-45. doi.org/10.1016/0305-0548(93)90063-O

Amin, A., Obeidat,F.,Shah,B., Adnan,A., Loo, J., Anwar,S. (2019).Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research,94*. 290-301. doi.org/10.1016/j.jbusres.2018.03.003

Bin, L., Peiji, S., & Juan, L. (2007). Customer Churn Prediction Based on the Decision Tree in Personal Handyphone System Service. *2007 International Conference On Service Systems And Service Management.* doi: 10.1109/icsssm.2007.4280145

Bin,L., Peiji,S., Juan,L. (2007).Customer Churn Prediction Based on the Decision Tree in Personal Handyphone System Service. *International Conference on Service Systems and Service Management,* 687- 696. DOI: 10.1109/ICSSSM.2007.4280145

Borrego, M., Douglas, E., & Amelink, C. (2009). Quantitative, Qualitative, and Mixed Research Methods in Engineering Education. *Journal Of Engineering Education, 98(1),* 53-66. doi: 10.1002/j.2168-9830.2009.tb01005.x

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing, 300,* 70–79. doi.org/10.1016/j.neucom.2017.11.077

Dalvi, P., Khandge, S., Deomore, A., Bankar, A., & Kanade, V. (2016). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. doi: 10.1109/cdan.2016.7570883

doi.org/10.1016/j.ejor.2011.09.031

F.Y, O., J.E.T, A., O, A., J. O, H., O, O., & J, A. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal Of Computer Trends And Technology, 48(3),* 128-138. doi: 10.14445/22312803/ijctt-v48p126*2016 Symposium On Colossal Data Analysis And Networking (CDAN).*

Fabris, F., Magalhães, J., & Freitas, A. (2017). A review of supervised machine learning applied to ageing research. *Biogerontology, 18(2),* 171-188. doi: 10.1007/s10522-017-9683-y

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering, 2(4).* 42-47.

Gordini,N., Veglio, V.(2017). Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. *Industrial Marketing Management, 62,*100-107. doi.org/10.1016/j.indmarman.2016.08.003

Guo-en, X., Wei-dong, J.(2008). Model of Customer Churn Prediction on Support Vector Machine. *SETP Journal Title, 28(1), 7*1-77. doi.org/10.1016/S1874-8651(09)60003-X

Hadden, J., Tiwari, A., Roy, R., Ruta, D.(2005).Computer assisted customer churn management:State-of-the-art and future trends. *Computers & Operations Research 34(10),* 2902-2917. doi.org/10.1016/j.cor.2005.11.007

He,B., Shi,Y., Wan, Q., Zhao, X. (2014). Prediction of Customer Attrition of Commercial Banks based on SVM Model. *Procedia Computer Science,31*. 423-430. doi.org/10.1016/j.procs.2014.05.286

Huang, B. Q., Kechadi, T., Buckley, B.,Keirnan, G.,Keogh, E., Rashid, T. (2010). A new feature set with new window techniques for customer churn prediction in land-line telecommunications. *Expert Systems with Applications, 37(5)*. 3657-3665. doi.org/10.1016/j.eswa.2009.10.025

Idris, A., Rizwan, M., & Khan, A. (2012). Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies. *Computers & Electrical Engineering, 38(6),* 1808–1819. doi.org/10.1016/j.compeleceng.2012.09.001

Jahromi,A., Stakhovych,S., Ewing,M. (2014). Managing B2Bcustomer churn, retention and profitability. *Industrial Marketing Management,43(7)*.1258-1268. doi.org/10.1016/j.indmarman.2014.06.016

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal Of Anesthesiology, 64(5),* 402-409. doi: 10.4097/kjae.2013.64.5.402

Kaya, E., Dong, X., Suhara,Y., Balsicoy, S., Bozkaya, B., Pentland, A. (2018). Behavioral Attributes and Financial Churn Prediction. *EPJ Data Science, 7(1),* 1-18. doi.org/10.1140/epjds/s13688-018-0165-5

Kelleher, J., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. Cambridge, Massachusetts: The MIT Press, 2015.

Kelleher, J., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics. Cambridge (Mass.): The MIT Press.

Khan, M. R., Manoj, J., Singh, A., & Blumenstock, J. (2015). Behavioral Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty. *IEEE International Congress on Big Data.* 1-4, doi.org/10.1109/bigdatacongress.2015.107

Kim, M., Park, M., & Jeong, D. (2004). The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services. *Telecommunications Policy, 28(2),* 145-159. doi: 10.1016/j.telpol.2003.12.003

Kim, S., Shin, K., & Park, K. (2005). An Application of Support Vector Machines for Customer Churn Analysis: Credit Card Case. *Lecture Notes In Computer Science,* 636-647. doi: 10.1007/11539117_91

KORKMAZ, M., GÜNEY, S. and YİĞİTER, Ş. (2012). The importance of logistic regression implementations in the Turkish livestock sector and logistic regression implementations/fields. *Harran University, 16(2),* 25-36.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions On Computer Science And Engineering, 30,* 1-12.

Kumar, D., & Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. *International Journal Of Data Analysis Techniques And Strategies, 1(1),* 4. doi: 10.1504/ijdats.2008.020020

Lee, H.,Lee, Y.,Cho, H., Im, K.,Kim,Y. (2017). Mining churning behaviors and developing retention strategies based on a partial least squares (PLS) model. *Decision Support Systems 52(1),* 207-216.doi.org/10.1016/j.dss.2011.07.005

Maheshwari, S., Jain, R.C.., & Jadon, R.S.. (2017). A Review on Class Imbalance Problem: Analysis and Potential Solutions. *International Journal Of Computer Science Issues, 14(6)*, 43-51. doi: 10.20943/01201706.4351

Malhotra, K. (2007). Marketing research – An applied orientation (5th Edn ed.). New Jersey: Pearson Education.

Manjupriya, R. and Poornima, A. (2018). Customer Churn Prediction in the Mobile Telecommunication Industry Using Decision Tree Classification Algorithm. *Journal of Computational and Theoretical Nanoscience,15(9)*.2789-2793. doi.org/10.1166/jctn.2018.7540

Miao, J., & Niu, L. (2016). A Survey on Feature Selection. *Procedia Computer Science, 91*, 919-926. doi: 10.1016/j.procs.2016.07.111

Mukaka, M. (2012). Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research. Malawi Medical Journal: *The Journal Of Medical Association Of Malawi, 24(3)*, 69-71.

Nashwan, S., & Hassan, H. (2017). Impact of customer relationship management (CRM) on customer satisfaction and loyalty: A systematic review. *Journal Of Advanced Research In Business And Management Studies*, *6(1)*, 86-107. Retrieved from https://www.researchgate.net/publication/318206357_Impact_of_customer_relationship_management_CRM_on_customer_satisfaction_and_loyalty_A_systematic_review

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems With Applications, 38(12)*, 15273-15285. doi: 10.1016/j.eswa.2011.06.028.

Oyeniyi, A. O., Adeyemo, A. B. (2015). Customer Churn Analysis In Banking Sector Using Data Mining Techniques. *African Journal of Computing and ICT, 8(3)*, 165 - 174. 10.1109/IWBIS.2019.8935884

Poel, D., Lariviere, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research*, 196-217. doi.org/10.1016/S0377-2217(03)00069-9

Pretorius, A., Bierman, S., & Steel, S. (2016). A meta-analysis of research in random forests for classification. 2016 *Pattern Recognition Association Of South Africa And Robotics And Mechatronics International Conference (PRASA-Robmech).*, 1-6. doi: 10.1109/robomech.2016.7813171

Pretorius, A., Bierman, S., & Steel, S. J. (2016). A meta-analysis of research in random forests for classification. *Pattern Recognition Association of South Africa and Robotics and*

*Mechatronics International Conference (PRASA-RobMech).*1-10, doi.org/10.1109/robomech.2016.7813171

Saunders, M., Lewis, P., Thornbill, A. (2009). Research Methods for Business Students (5th Edn ed.). England: Pearson Education.

Sayed, H., A., M., & Kholief, S. (2018). Predicting Potential Banking Customer Churn using Apache Spark ML and MLlib Packages: A Comparative Study. *International Journal of Advanced Computer Science and Applications, 9(11).* doi.org/10.14569/ijacsa.2018.091196

Senanayake, D., Muthugama, L., Mendis, L., & Madushanka, T. (2015). Customer Churn Prediction: A Cognitive Approach. *World Academy Of Science, Engineering And Technology International Journal Of Computer And Information Engineering, 9(3),* 767-773. doi:org/10.5281/zenodo.1100190

Shaaban, E., Helmy, Y., Khedr, A., & Nasr, M. (2012). A Proposed Churn Prediction Model. *International Journal Of Engineering Research And Applications (IJERA), 2(4),* 693-697.

Sharma, A., & Kumar Panigrahi, P. (2011). A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services. *International Journal of Computer Applications, 27(11),* 26–31. doi.org/10.5120/3344-4605

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom),* 1310-1315.

Subramanian, V., Hung, M., Hu, M.(1992). An Experimental Evaluation of Neural Network for Classification. *Computers & Operations Research, 20(7).*769-782.doi.org/10.1016/0305-0548(93)90063-O

Tian, Y., Shi, Y., & Liu, X. (2012). Recent Advances On Support Vector Machines Research. *Technological and Economic Development of Economy, 18(1),* 5–33. doi.org/10.3846/20294913.2012.661205

Tsai, C., Lu, Y. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications, 36(10), 12547-* 12553. doi.org/10.1016/j.eswa.2009.05.032

Umayaparvathi, V., & Iyakutti, K. (2012). Applications of Data Mining Techniques in Telecom Churn Prediction. *International Journal Of Computer Applications, 42(20),* 5-9. doi: 10.5120/5814-8122

Vafeiadis, T., Diamantaras, K., Chatzisavvas, K., Sarigiannidis, G. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory, 55,* 1-9, doi: 10.1016/j.simpat.2015.03.003.

Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal Of Operational Research, 157(1),* 196-217. doi: 10.1016/s0377-2217(03)00069-9

Verbeke, W., Dejaeger,K., Martens, D., Hur,J.,Baesens, B.(2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research 218(1).* 211-229.

Wieringa, R., Maiden, N., Mead, N., & Rolland, C. (2005). Requirements engineering paper classification and evaluation criteria: a proposal and a discussion. *Requirements Engineering, 11(1),* 102-107. doi: 10.1007/s00766-005-0021-6

Wirth, R., Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining

Xie, Y., Li, X., Ngai, E., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems With Applications, 36(3),* 5445-5449. doi: 10.1016/j.eswa.2008.06.121

Zhang, S., Zhang, C., Yang, Q. (2003). Data Preparation for Data Mining. *Applied Artificial Intelligence,17,*375–381.DOI:10.1080/08839510390219264

Zorich, A. (2018). Predicting Customer Churn In Banking Industry Using Neural Networks. *Interdisciplinary Description of Complex Systems, 14.* 116-124. https://doi.org/10.7906/indecs.14.2.1

APPENDIX A

Results without Sampling the data

Logistic Regression

*Classification Report:

	precision	recall	f1-score	support
0	0.90	0.95	0.92	13595
1	0.81	0.67	0.73	4528
accuracy			0.88	18123
macro avg	0.85	0.81	0.83	18123
weighted avg	0.87	0.88	0.87	18123

Random Forest

*Classification Report:

	precision	recall	f1-score	support
0	0.99	0.98	0.98	13595
1	0.93	0.98	0.95	4528
accuracy			0.98	18123
macro avg	0.96	0.98	0.97	18123
weighted avg	0.98	0.98	0.98	18123

SVM

*Classification Report:

	precision	recall	f1-score	support
0	0.91	0.95	0.93	13595
1	0.84	0.71	0.77	4528
accuracy			0.89	18123
macro avg	0.87	0.83	0.85	18123
weighted avg	0.89	0.89	0.89	18123

Neural Network

*Classification Report:

	precision	recall	f1-score	support
0	0.92	0.98	0.95	13595
1	0.91	0.74	0.82	4528
accuracy			0.92	18123
macro avg	0.92	0.86	0.88	18123
weighted avg	0.92	0.92	0.92	18123

Results for SMOTED Sampled data

Logistic Regression

*Classification Report:

	precision	recall	f1-score	support
0	0.94	0.84	0.89	13584
1	0.63	0.84	0.72	4539
accuracy			0.84	18123
macro avg	0.79	0.84	0.80	18123
weighted avg	0.86	0.84	0.85	18123

Random Forest

*Classification Report:

	precision	recall	f1-score	support
0	0.99	0.97	0.98	13584
1	0.93	0.98	0.95	4539
accuracy			0.98	18123
macro avg	0.96	0.98	0.97	18123
weighted avg	0.98	0.98	0.98	18123

SVM

*Classification Report:

	precision	recall	f1-score	support
0	0.95	0.86	0.90	13584
1	0.68	0.86	0.76	4539
accuracy			0.86	18123
macro avg	0.81	0.86	0.83	18123
weighted avg	0.88	0.86	0.87	18123

Neural Network

*Classification Report:

	precision	recall	f1-score	support
0	0.97	0.92	0.94	13584
1	0.79	0.92	0.85	4539
accuracy			0.92	18123
macro avg	0.88	0.92	0.90	18123
weighted avg	0.93	0.92	0.92	18123