PREDICTING CUSTOMERS' CHURN USING DATA MINING TECHNIQUES (MACHINE LEARNING APPROACH)

BY

ABDUL AHMED ABDUL

SPS/19/MCA/00001

A THESIS SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE, BAYERO UNIVERSITY, KANO-NIGERIA IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF MASTERS OF COMPUTER APPLICATION (MCA Hons.) IN COMPUTER SCIENCE.

OCTOBER 2021

DECLARATION

I, Abdul Ahmed Abdul, hereby declare that this thesis titled '**PREDICTING CUSTOMERS'**

CHURN USING DATA MINING TECHNIQUES (MACHINE LEARNING

APPROACH)', has been carried out by me under the supervision of Mr. Ismail Abu Zibiri. It

has not been presented for award of any degree in any institution. All sources of information

are specifically acknowledged by means of reference.

……………………………….   …………………………..

        Signature   Date

CERTIFICATION

This thesis titled '**Predicting Customers' Churn Using Data Mining Techniques (Machine Learning Approach)**'by Abdul Ahmed Abdul meets the requirements governing the award of the degree of Masters (MSc) in Computer Applications and is approved for its contribution to knowledge and literary representation.

……………………………   …………………………..

Mr. Ismail Abu Zibiri   Date
Thesis Advisor

……………………………   …………………………..

Dr. Habiba Kakudi   Date
Head of Department

External Supervisor

Name…………………………….   ……………………………….

Signature………………………….   Date

DEDICATION

I wish to dedicate this work to my parents whom has tirelessly support me for much longer than reasonable without asking.

ACKNOWLEDGEMENT

I was told that the important factor of a research career is having the right people as your mentors and, through little doing of my own. I have been extraordinary fortune in the people I have had in this role.

I would like to thank my supervisor, apart from directing me, made a huge sacrifice of times to read through my work thoroughly leading to the successful compilation of the research. I will also like to appreciate the entire Lecturers and Staffs of Department of Computer Science Bayero University, Kano and for their kind support toward the fulfilment of this research thesis.

Also. my appreciation goes to all the graduating students of SPS/19 class of 2021 BUK; both of which have lived up to their reputations, but more importantly are immensely friendly fostering lasting collaboration and friendships. I 'm honoured to have been able to learn in this environment. I thank my family and friends those whose names have not been mentioned here and who in one way or the other contributed towards the successful completion of my Masters study, I appreciate you all and God bless.

ABSTRACT

Customer churn is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenues of the companies, especially in the telecommunication field, companies are seeking to develop means to predict potential customer to churn. Customers are the most important assets in any industry since they are considered as the main profit source. Churners are persons who quit a company's service for some reasons. Companies should be able to predict the behaviour of customer correctly in order to reduce customer churn rate. Customer churn has emerged as one of the major issues in every Industry. Researches indicates that it is more expensive to gain a new customer than to retain an existing one. In order to retain existing customers, service providers need to know the reasons of churn, which can be realized through the knowledge extracted from the data. To prevent the customer churn, a churn prediction model is developed in this project which can assist companies to predict customers who are most likely to churn. It uses machine learning techniques such as Logistic Regression, Decision Trees, and Random Forest to identify the primary determinants of customer churn along with the algorithm fit for such predictions. The dataset contains demographic details of customers, their data usage and the type of service they receive from the company. It comprises of churn data of 3333 customers spread over 11 attributes obtained from Global AI. Further on this investigation, the usage of the above-mentioned algorithms is described for predicting customer churn. Also, it has been found that the significance of data services is constantly increasing for the subscribers. These findings will be useful to both service providers and researchers to comprehend the service quality parameters considering mobile subscribers view point. This can help to re-look the existing benchmarks and increase customer satisfaction. However, the best results were obtained by applying Random Forest algorithm.

TABLE OF CONTENTS

LISTS OF FIGURES

LISTS OF TABLES

CHAPTER ONE

INTRODUCTION

1.1 **Background of study:** Emphasizing the higher costs associated with attracting new customers compared with retaining existing customers, and the fact that long-term customers tend to produce more profits, (Verbeke et al. 2011) assert that customer retention increases with profitability. Many competitive organizations have realized that a key strategy for survival within the industry is to retain existing customers. (Tsai and Chen. 2010) argued that "this leads to the importance of churn management." Customer churn represents a basic problem within the competitive atmosphere of telecommunication industry. The telecommunications sector has become one of the main industries in developed and developing countries. The technical progress and the increasing number of operators raised the level of competition (Gerpott TJ, Rams W, Schindler A, 2001). Companies are working hard to survive in this competitive market depending on multiple strategies. Tree main strategies have been proposed to generate more revenues (Wei CP, Chiu IT, 2002):

- acquire new customers,

- upsell the existing customers, and

- increase the retention period of customers.

However, comparing these strategies taking the value of return on investment (RoI) of each into account has shown that the third strategy is the most profitable strategy (Wei CP, Chiu IT, 2002), proves that retaining an existing customer cost much lower than acquiring a new one (Qureshii SA, et al. 2013), in addition to being considered much easier than the upselling strategy (Ascarza E, 2016). To apply the third strategy, companies have to decrease the potential of customer's churn, known as "the customer movement from one provider to

another" (Bott. 2014). The development and digitalization of the world has led to new ways of doing business and companies all over the globe have been forced to adapt (R. Shkurti and A. Muca, 2014). Subscription based services are one of the outcomes of the explosive digitalization that has taken the world by storm and with this comes both possibilities and challenges that require modern day solutions (C. Blank and T. Hermansson, 2018). Digitalization has not only changed the way business is conducted but the abundance of information available has also led to consumers facing a higher supply of subscription-based services. This can be viewed as a challenge for companies since retaining customers can potentially become more difficult. Digitalization within companies can lead to a decrease in labour costs, an increase in efficiency and a better overview of the company's operations within the organization (R. Shkurti and A. Muca, 2014). All of this is essential for staying competitive, and to gain an edge over other companies.

As information technology is a growing trend, the available amount of data and information has increased significantly during the past years. This rapid growth has enabled storage and processing of great amounts of data while increasing the necessity of automatically finding valuable information and creating knowledge (D. Buo and M. Kjellander, 2014). With meaningful information extracted from the stored data, firms can make appropriate decisions in order to grow the business. With this growth, the use of data mining techniques and machine learning has increased, due to its ability of handling and analysing great amounts of data (K. Mishra and R. Rani, 2017). The digitalization has also brought forward an ongoing trend to improve current data processing activities as a part of customer relationship management (CRM) strategies. The idea of knowledge management and customer relationship management has lately obtained more attention in the subscription-based business model and the concepts focus on distribution of resources to activities that are customer-centric, to be able to increase competitive advantages (H. Gebert, 2003 and M. Sergue, 2020). Customer knowledge is the

knowledge that businesses can obtain through interaction with their customers (F. Khodakarami and Y. Chan, 2014). Customer relationship management systems are systems that support interaction between businesses and customers with the objective of collecting, storing and analysing data to get an overview of their customers (F. Khodakarami and Y. Chan, 2014). Such systems have evolved through the past years and by using technology and different data analysis tools, enterprises can find patterns in customer's behaviour, which would be almost impossible to discover manually (D. Buo and M. Kjellander, 2014). Such patterns could vary from a customer's purchasing behaviour or patterns related to customer churning. In a subscription-based business model, a fundamental part of success is to minimize the rate of customers ending their subscriptions, in other words, to minimize churn (D. Buo and M. Kjellander, 2014).

Customer churning refers to the action of when a customer chooses to abandon their service provider (D. L. Garcia, et. al 2017). The term is relatively new and has gained more relevance with the emergence of online services. Firms across the globe recognize customer churning as a great loss since they have already invested in attracting these customers. This is one of the major reasons that customer retention is beneficial for a firm. Customers can churn for many reasons and it is hard to pinpoint a general reason for churning. The availability of information has given consumers a bargaining power, and nowadays customers can easily find the service provider, which provides the same product with a more satisfying deal (N. Gordini and V. Veglio, 2017). To manage this, firms invest in customer churn prediction, which means that companies try to predict which of their customers will churn, so that they can apply preventative measures. These preventive measures could differ depending on the reason a customer might churn, and could be for example, offering a lower price or including an extra service. As mentioned earlier, analysing customer behaviour serves as the basis for predicting customers who might churn, which is important for many reasons. One reason is that for

companies who rely on subscription-based income, it can make a big difference on whether they can keep a steady income level or if they need to make changes to their services to keep customers. Another reason is that, compared to retaining customers, attracting new ones is costlier and firms can save money by retaining their existing customer base (I. Ullah, 2019).

More so, with new companies continuously emerging in the telecommunications industry, and new customers to those companies sparse, markets have matured to a point where they have saturated, so possible new customer acquisition is scarce. Because of this, companies are beginning to recognise that their most valuable assets are their existing customers. Holding onto existing customers is not an undemanding procedure, with further complications arising from governing bodies such as NCC. These authorities have emerged to act in the customers' best interests, and the promotion of healthy competition amongst service providers to ensure that the customer has a choice, and services are not monopolised. A further challenge that is commonly being recognised regarding the acquisition of new customers is lack of data, making targeted sales campaigns difficult to deploy.

In today's business world companies are recognising that customer value and increased revenue is more likely to come from their existing customer base than from new customer acquisition. The reasoning behind this is that companies know their existing customers, already have a relationship with them, and amplitude of data on them. In common recognition of this problem, industry has seen an emergence of customer relationship management (CRM) products. Software companies have realised that CRM has become a receptive topic within industry and therefore an area of potential wealth and opportunity for the development and promotion of products that promise to not only boost customer retention, but also for increasing selling opportunities (Hadden J., et al 2007).

Also, in the past years, companies have been able to store and process huge amounts of data while realizing that being customer-centric was becoming a main requirement to stand out from the competition. Indeed, due to saturated markets, focusing on Customer Relationship Management (CRM) to retain an existing customer base is not optional anymore, but an absolute necessity for competitive survival. In its research for Bain and Company, Frederick Reichheld stated that the cost of acquiring a new customer could be higher than that of retaining a customer by as much as 70%, and that increasing customer retention rates by a mere 5% could increase profits by 25% to 95%. More generally, data-driven decision making is way for businesses to make sure their next move will benefit both them and their customers. Almost every company, especially in the Tech ecosystem has now put into place a tracking process, to gather data related to their customers' behaviour. The data to track varies along with the specific business model of each company and the problem service they aim to address. By analysing how, when and why customers behave a certain way, it is possible to predict their next steps and have time to work on fixing issues beforehand. This prediction and quantification of the risk of losing customers can be done globally or individually and is mainly used in areas where the product or service is marketed on a subscription basis. The prediction of churn is generally done by studying consumer behaviour or by observing individual behaviour that indicates a risk of attrition. It involves the use of modelling and machine learning techniques that can sometimes use a considerable amount of data.

1.2 Problem formulation: The subscription-based business model is continuously growing due to digitalization and offers companies an innovative way of conducting their business (C. Blank and T. Hermansson, 2018). At the same time, more and more services are being digitalized and data has become much easier to collect, store, and process (M. Sergue,

2020). There is an abundance of different service providers to choose from, which has increased competition and made it more difficult to retain customers, in this modern-day service market (C. Blank and T. Hermansson, 2018). Due to the availability of data and substitutes, subscription-based businesses must adapt by focusing more on Customer Relationship Management, specifically customer churn management (M. Sergue, 2020). According to Blank and Hermansson (2018) the key to success within a subscription-based business is to keep a low churn rate, which is defined as the number of customers leaving their service provider during a given period of time (C. Blank and T. Hermansson, 2018). As a service provider, there is a greater chance of selling to an existing customer rather than a com pletely new one. This can be highlighted by the cost of attracting new customers. According t o Verbeke et al. (2012), attracting new customers can cost somewhere between five to six times more relative to retaining customers, which states the importance of preventing churn. Apart from the cost of attracting and retaining customers, the net return on investments (ROI) for strategies related to retention is in general higher than for acquisition, which may increase revenue of customers who continue transacting (A. T. Jahromi, 2014).

Also, with the proliferation of smartphones, together with higher data usage, expanding the customer base, by capturing new customers, retaining the existing ones and preventing few of them from churning out, are the major concerns for any cellular service provider. According to Cisco (2017), "average monthly smartphone traffic will grow for all devices by 10 times between 2016 and 2022, while the mobile connected tablet traffic alone will exceed 1.1 exabytes roughly the amount of mobile data traffic worldwide in 2011". From the issues imposed through market saturation and cost implications as described above, there has been an identification of a need for a computer-based churn prediction methodology that is capable of accurately identifying a loss of customer in advance, so that proactive retention strategies can

be deployed in a bid to retain the customer. The churn prediction has to be accurate because retention strategies can be costly.

However, managing customer churn is one major challenge facing companies, especially those that offer subscription-based services. Customer churn is basically the loss of customers, and it is caused by a change in taste, lack of proper customer relationship strategy, change of residence and several other reasons. If businesses can effectively predict customer attrition, they can segment those customers that are highly likely to churn and provide better services to them. Hence, a churn prediction model is developed in this project that uses machine learning techniques such as Logistic Regression, Decision Trees and Random Forest algorithms to assist companies in predicting customers who are most likely to churn. In this way, they can achieve a high customer retention rate and maximize their revenue.

1.3   Research Objectives and Goals: As aforementioned, digitalization and access to information are in large proportions, companies are under constant threats of customer churning, in particularly subscription-based service providers that rely on such income Hence, it is important to study churn prediction and to support companies in this problem. The purpose of this study is to contribute knowledge in the field of churn prediction in a subscription-based service context and to develop a customer profiling methodology for predicting churn in advance. In this study we employ different machine learning techniques, in order to investigate the effects on the performance of churn prediction models in a subscription-based service context.

More Over, the goal is to propose and evaluate different approaches and techniques of how machine learning can be used to predict churn; hence creating value for businesses offering subscription-based services. The objectives of this study are identified as follows:

- Identify and select appropriate mathematical (i.e., statistical and data mining) techniques for developing predictive models.

- compare different types of algorithms for the built model, and

- study the effects of balancing the training dataset for the considered model

1.4 Research Questions: Based on the background study, the fact that much data is currently available, this thesis focuses on investigating the following research question:

- How can customer churning be predicted in the domain of business-to-customer using machine learning?

- To what extent can statistical analysis and machine learning algorithms be used to highlight churn drivers for a business-to-customer company with little data available so that churn can be explained?

1.5 Scope of Study: This research targets the services that fall under the umbrella of the telecommunications industry, i.e., telephone and mobile services. Also, the research focuses on faults and complaints data as an alternative to demographic and usage data as to avoid conflicts with monopoly regulations. Due to the time limit and complexity, three different supervised learning algorithms were chosen for comparison and evaluation. In a perfect world there are a lot of different algorithms and techniques that can be used for predicting customer churning and using a number of these algorithms means that they are not all utilized, which is why this is considered a delimitation. Due to limitations in raw data, attrition rate and misclassification will not be taken into account. Predicting customer churning can be used to lower costs and improve a firm's revenue streams, however this study will not cover such economical aspects.

CHAPTER TWO

LITERATURE REVIEW

2.1 **Introduction**: Customer churn as newly developed concept has been widely employed in telecommunication, e-retailing and banking industries in recent years. The definition for churn may vary for each organization with regards to the duration that a customer is apart from a company. Many existing studies highlight that in order to boost customer loyalty and boost retention, having a customized and differentiated relationship with its customers is fundamental. To answer this problem, several ways of performing analysis capturing customer behaviours exist.

Among others, Brandusoiu I, et al. (2016) presented an advanced methodology of data mining to predict churn for prepaid customers using dataset for call details of 3333 customers with 21 features, and a dependent churn parameter with two values: Yes/No. Some features include information about the number of incoming and outgoing messages and voicemail for each customer. He applied principal component analysis algorithm "PCA" to reduce data dimensions. Tree machine learning algorithms were used: Neural Networks, Support Vector Machine, and Bayes Networks to predict churn factor. He used Area Under Curve "AUC" to measure the performance of the algorithms. The AUC values were 99.10%, 99.55% and 99.70% for Bayes Networks, Neural networks and support vector machine, respectively. The dataset used in this study is small and no missing values existed.

He Y, et al. (2009) proposed a model for prediction based on the Neural Network algorithm in order to solve the problem of customer churn in a large Chinese telecom company which contains about 5.23 million customers. The prediction accuracy standard was the overall accuracy rate, and reached 91.1%.

Idris A. (2012) proposed an approach based on genetic programming with AdaBoost to model the churn problem in telecommunications. The model was tested on two standard data sets. One by Orange Telecom and the other by cell2cell, with 89% accuracy for the cell2cell dataset and 63% for the other one.

Huang F, et al. (2015) studied the problem of customer churn in the big data platform. The goal of the researchers was to prove that big data greatly enhance the process of predicting the churn depending on the volume, variety, and velocity of the data. Dealing with data from the Operation Support department and Business Support department at China's largest telecommunications company needed a big data platform to engineer the fractures. Random Forest algorithm was used and evaluated using AUC.

Makhtar M, et al. (2017) proposed a model for churn prediction using rough set theory in telecom. As mentioned in his paper Rough Set classification algorithm outperformed the other algorithms like Linear Regression, Decision Tree, and Voted Perception Neural Network. Various researches studied the problem of unbalanced data sets where the churned customer classes are smaller than the active customer classes, as it is a major issue in churn prediction problem. Amin A, et al. (2016) compared six different sampling techniques for oversampling regarding telecom churn prediction problem. The results showed that the algorithms (MTDF and rules-generation based on genetic algorithms) outperformed the other compared oversampling algorithms.

Burez and Van den Poel (2009) studied the problem of unbalance datasets in churn prediction models and compared performance of Random Sampling, Advanced Under Sampling, Gradient Boosting Model, and Weighted Random Forests. They used (AUC, Lift) metrics to evaluate the model. The result showed that under sampling technique outperformed the other tested techniques.

2.2   Customer Relationship Management (CRM): Customer relationship management (CRM) can be viewed as a strategy of a company with the aim of reducing costs, improving p rofitability, and improving customer relations by offering the right product or service to the right customer (S. R. Gulliver, 2013). CRM can be done in different ways, meaning that the method or process used by companies for CRM differs (S. R. Gulliver, 2013; A. Payne and P. Frow, 2005). However, the main principle of CRM stays the same, which is to attract customers, learn about them, find the most suitable way of serving them, and then use this knowledge to retain them (S. R. Gulliver, 2013). This is done through something called CRM systems which is the technological part of it. The purpose of these systems is to enable communication with the customer but to also store and analyse customer data in order to get an overview of the firm's customer base (A. Payne and P. Frow, 2005).

Also, a number of products exist for customer relationship management (CRM) which aims at analysing a company's customer base. CRM is not a new concept, beginning in the mid 1990's with IT based systems being developed to track multiple customer activities (Minami and Dawson, 2008). Today there are many commercial products available for the purpose of CRM with the amount of money being invested in this field exploding over recent years (Ang and Buttle, 2002). With so many products emerging organisations should take great care when deciding to purchase one of these solutions 'off-the-shelf'. Chen and Popovich (2003) state that "CRM vendors might entice organizations with promises of all powerful applications.

2.3   Customer Churn Management: It is the concept of identifying   which   customers might churn and has become a core strategy to survive within an industry (Hadden. J et al. 20 07). Customer churn management   can   be   used   to   identify   potential   churners   and target those customers with proactive marketing campaigns with retention incentives (Hung,

et al, 2006).

According to Hadden. J et al. (2007), customers who churn can be divided into voluntary and nonvoluntary churn. Non-voluntary churners are the ones that had their service removed by the company, which are the easiest churners to identify. Voluntary churn is unlike, non-voluntary churn, harder to identify and occurs when the customer makes a decision to abandon the service provider. Such churn could occur because of changes in circumstances, such as a customer's financial situation, in other words incidental churn. Other reasons for voluntary churn could be the customer's intentions of changing service provider to a competitor, which is called deliberate churn.

Hung et al. (2006) state that churn management is considered a part of customer relationship management and from a business intelligence perspective, churn management includes two tasks. The first is to predict customers who might churn, and the second task is to identify possible retention strategies from an organization's point of view.

2.4   Customer Continuity Management: The possible symptoms that might alert of a possible defection to the competition must be preventively anticipated, and this requires evaluation of their evolution over time. Therefore, the adoption of a suitable *Customer Continuity Management* (CCM) model (Garcia et al. 2007) should make it easier for companies to critically review all aspects that might affect the construction of true loyalty bonds with customers, including policies for everyday management, both for the customer life-cycles and for the predicted and declared cases of customer loss as shown in the figure below.
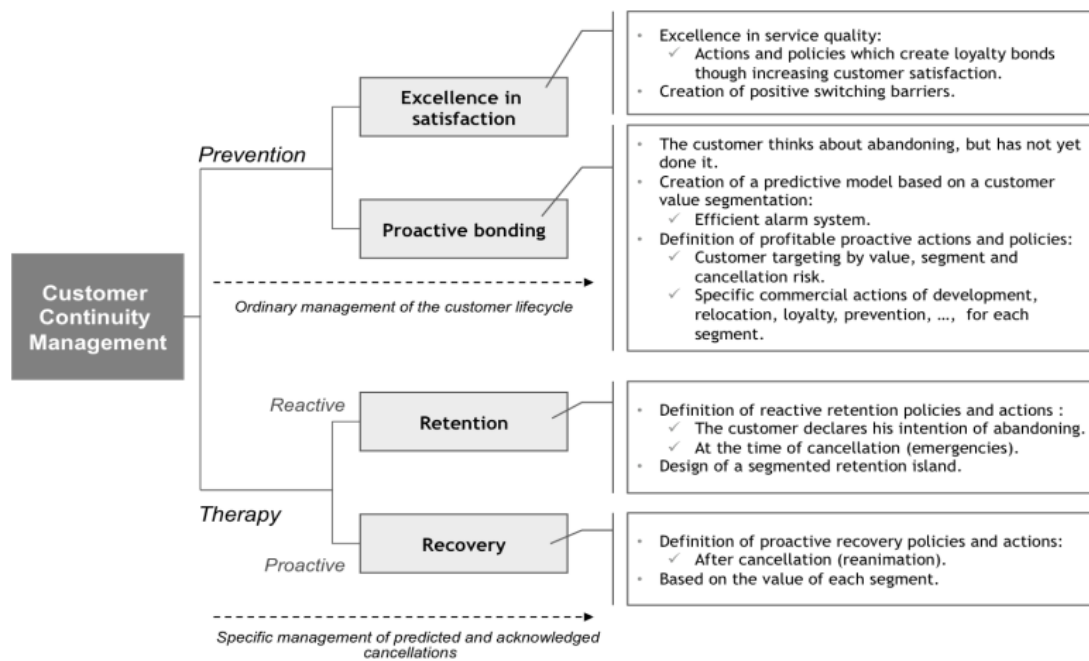
Excellence in service quality:
- ✓ Actions and policies which create loyalty bonds though increasing customer satisfaction.
- Creation of positive switching barriers.

- The customer thinks about abandoning, but has not yet done it.
- Creation of a predictive model based on a customer value segmentation:
  - ✓ Efficient alarm system.
- Definition of profitable proactive actions and policies:
  - ✓ Customer targeting by value, segment and cancellation risk.
  - ✓ Specific commercial actions of development, relocation, loyalty, prevention, ..., for each segment.

- Definition of reactive retention policies and actions :
  - ✓ The customer declares his intention of abandoning.
  - ✓ At the time of cancellation (emergencies).
- Design of a segmented retention island.

- Definition of proactive recovery policies and actions:
  - ✓ After cancellation (reanimation).
- Based on the value of each segment.

Figure 1: Customer Continuity Management Model

The level of customer bonding and, as a result, their life expectancy is intimately bound to the level of customer satisfaction relative to the quality of the service provided by a company. The higher the level of quality of service that customers perceive, the stronger the loyalty bonds (Cronin et al. 2000; Jones et al. 2000; Jones et al. 2002; Patterson et al. 2003; Kim et al. 2004). Thus, consumers who experience high levels of satisfaction about the service usually continue with their current provider.

However, despite customer satisfaction having a positive influence on the level of bonding, it does not always suffice. There are numerous situations in which better service quality does not have a significant impact on consumer loyalty: for instance, customers that change mobile phone operators in spite of the fact that their current provider offers greater coverage; customers that fill up in slow petrol stations, with bad accesses and no additional services for the driver; customers who prefer to travel with certain airlines despite the continuous delays to their flights, etc. In consequence, there must be other factors, beyond satisfaction with service, influencing customer loyalty.

2.5   Data Mining Approach: Data mining refers to the discovery of knowledge from a huge amount of data (Nie et al. 2011). Tsai and Lu (2009) described data mining as discovering interesting patterns within the data and predicting or classifying the behaviour exhibited by the model. Seng and Chen (2010) suggested that the basic challenge is how to convert seemingly meaningless data into useful information and competitive intelligence. Tsai and Lu (2009) stipulated that "in literature, statistical and data mining techniques have been used to create the prediction models." Classification tools are often used to model and predict customer churn. Some of the techniques commonly used to achieve this are neural networks, decision trees (DT), random forests, support vector machines (SVM) and logistic regression (Miguéis et al. 2012).

Guo-en and Wei-dong (2008) focused on building a customer churn prediction model using SVM in the telecommunication industry. They compared this method with other techniques such as DT, artificial neural networks, naïve Bayesian (NB) and logistic regression. The results proved SVM to be a simple classification method of high capability yet good precision. Anil Kumar and Ravi (2008) used data mining to predict credit card customer churn. They used multilayer perceptron (MLP), logistic regression, DT, random forest, radial basis function, and SVM techniques. Nie et al. (2011) built a customer churn prediction model by using logistic regression and DT-based techniques within the context of the banking industry. In their study, Lin et al. (2011) used rough set theory and rule-based decision-making techniques to extract rules related to customer churn in credit card accounts using a flow network graph (a path-dependent approach to deriving decision rules and variables). They further showed how rules and different kinds of churn are related.

Sharma and Panigrahi (2011) applied neural networks to predict customer churn from cellular network services. The results indicated that neural networks could predict customer churn with an accuracy of higher than 92 %. Saradhi and Palshikar (2011) compared machine learning

techniques used to build an employee churn prediction model. Yu et al. (2011) applied neural network, SVM, DT, and extended SVM (ESVM) techniques to forecast customer churn. Of the methods studied, ESVM performed best. Huang et al. (2012) presented new-features-based logistic regression (LR), linear classifier (LC), NB, DT, MLP neural networks, and SVM. In their experiments, each technique produced a different output. Data mining by evolutionary learning (DMEL) could show the reason or probability of a churning phenomenon; DT, however, could only show the reason. LR, NB, and MLP could provide probabilities of different customer behaviours. LC and SVM could distinguish between a churner and a non-churner. Farquad et al. (2014) used SVM to predict customer churn from bank credit cards. They introduced a hybrid approach to extract rules from SVM for customer relationship management purposes. The approach is composed of three phases where:

- SVM-recursive feature elimination is applied to reduce the feature set;
- the obtained dataset is used to build the SVM model; and
- using NB, tree rules are generated.

Keramati et al. (2014) not only presented different approaches to data mining and classification methods such as DT, neural networks, SVM, and k-nearest neighbours, but also had the performances of these approaches compared.

2.6   Time of Churn Prediction: Junxiang (2002) and Masarifoglu and Buyuklu (2019), study customer churn prediction using survival analysis, which is a group of statistical methods used for studying events based on certain circumstances, such as customer churn prediction. Using survival analysis, Masarifoglu and Buyuklu (2019) emphasize that the question they answer is the time until a certain event occurs. Their model estimation is based on the survival function and the hazard function. They employed the Kaplan-Meier method and the Cox model in R. Using this approach, the preliminary risks of customer churning are explained and

15

quantified.

Survival curves and hazard ratios are also presented, which they argue, allows service providers to take preventative measures against churners. Masarifoglu and Buyuklu (2019) use historical telecom data acquired from a mobile operator for their study, which contains information regarding 10365 randomly selected customers. The data consists of categorical variables represented as dummy variables. These variables represent information regarding subscription type, length of subscription but also customer specific information such as gender. The modelling process used by Junxiang (2002) is done using four steps: explanatory data analysis, variable reduction, model estimation, and model validation, where the first two steps can be viewed as preparing the data for the survival analysis. Just like Masarifoglu and Buyuklu (2019), the model estimation is based on the survival function and the hazard function. They agreed that the purpose of this estimation is to identify potential churn characteristics and estimate customer churn by calculating probabilities for survival. The last step in the modelling process, that is, model validation, is done by scoring predicted survival probabilities at a specified time for each customer. These survival probabilities are ranked in ascending order and they state that the customers with the lowest predicted survival probabilities are the ones that will most likely churn. Junixiang (2002) ranks the survival probabilities in different deciles and compares the predicted number of churners at a specific time in each decile. The raw data used in the study consist of demographic data, customer contact data, and customer internal data, which is further divided into warehouse- and telecommunications data. The study started with a dataset containing 212 variables, which was later reduced to 115 variables. These variables are categorized into both numerical and categorical values and only 29 variables are explanatory variables, which Junixiang (2002) uses in the survival analysis.

2.7   Churn Prediction in Business to customer Domain: Vafeiadis et al. (2015) compare different machine learning classifiers in order to predict customer churning. The data used in

this study is provided by a telecom provider and consists of 5000 samples and multiple variables. The data mostly contains usage information such as, call duration and number of texts sent. In addition, specific subscription information is also used, such as subscription period. The different classifiers compared are Artificial Neural Network (ANN), Support Vector Machine (SVN), Decision Tree (DT), Naïve Bayes, and Logistic Regression. Furthermore, they use boosting in order to compare the single learners with their boosted versions; however, Naïve Bayes and Logistic Regression are not boosted due to imitations in their parameters. In order to evaluate the performance of the different classifiers, they use precision, recall, accuracy and F-measure, which are calculated from the confusion matrix. Based on these measures they conclude that boosting significantly improves the performance in terms of both accuracy and F-measure for all of the considered classifiers. On the other hand, without the use of boosting, they suggest that Support Vector Machine is a good tool for customer churn prediction. The accuracy of all the studied classifiers ranges from 93 to 99 percent and the F-measure between 73 and 77 percent.

According to Brandusoiu et al. (2016), prediction tasks depend strongly on data mining techniques, since machine learning algorithms improve the performance of the prediction process. They compare different machine learning algorithms with the aim of predicting churn. The problem addressed is a classification problem with churn/non-churn as categorical variables. They compare Support Vector Machines with Bayesian Networks and Neural Networks. The result is based on a confusion matrix and a gain measure, where the overall accuracy for all the considered algorithms ranges between 99 and 100 percent. The dataset used in the study consists of usage data from a telecom company, such as how many calls or texts the subscriber has made. The data contains information regarding 3333 subscribers with 20 variables, 15 of them being continuous and five being discrete. The dependent variable is categorical, churn or non-churn. Based on their result they conclude that all three classifiers

performed well on predicting churn. They further suggest that the predictive performance can be improved by applying ensemble learning structures.

Lalwani et al. (2021), conduct a study related to churn prediction using a machine learning approach. They use Logistic Regression, Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, and various boosting algorithms such as XGBoost and AdaBoost. To conduct research of this magnitude, Lalwani et al. (2021), use comprehensive data from a telecom provider that consists of 7000 instances. Furthermore, the dataset includes 21 features, categorized into numerical and categorical values. The dataset consists of usage information such as number of calls, but also specific customer data, such as gender. In order to address their research question, they follow six steps. The first three steps are related to data processing, feature analysis, and feature selection. In the two following steps, they develop their model based on the different algorithms. In the last step of the process, they evaluate the performance based on the metrics: precision, recall, accuracy, and F-measure, calculated from a confusion matrix. The accuracy ranges from 74 to 82 percent, recall from 74 to 82 percent, precision from 57 to 81 percent, and F-measure from 63 to 81 percent. Lalwani et al. (2021) concludes that the boosted classifiers, XGBoost and AdaBoost, performed best and achieved the best accuracy. In addition, they state that machine learning is the most efficient way of dealing with predictions, which is why they consider that the future of churn prediction will revolve around machine learning.

2.8   Review on factors affecting churn: it is more important   to   retain   profitable subscriber (Howard, 2010). Zineldin (2000), customer retention may be defined as the percentage of customers who are with the company at the beginning of the period as well as at the end of period. Fill (2005), customer retention is the stage where there is an understanding between the customer and the company. The other side of retention is when customer is locked

with the company due to some contract for e.g., tariff or number portability. The switching of cellular subscribers or churning is one of the main concerns for the industry. Churn analysis not only helps the service provider to know which customers are at risk of moving to other service provider but it also helps in determining the customers which are profitable for the company. Loyal and satisfied customers are assets to any cellular service provider resulting in constant billing to their service providers and their billing increases with their tenure (Reichheld and Sasser, 1990). It is expected that the costs of attaining a new customer is five to seven times more than retaining an existing customer (Mittal and Lassar, 1998). These customers are not so much concerned about the tariff and are open to pay even more for superior services (Gould, 1995; Reichheld and Sasser, 1990). Thereby it is imperative for the service providers to employ churn prediction techniques and based on them, apply appropriate marketing strategies to retain the existing subscriber. Service providers are constantly working on optimising the performance of churn prediction model. Previous studies have employed different data mining techniques to forecast subscribers who are at the propensity of churn (Tsai and Lu, 2010; Hung et al., 2006; Balasubramanian and Selvarani, 2014; Almana et al., 2014; Dahiya and Talwar, 2015; Glatz et al., 2014; Vafeiadis et al., 2015; Khizindar et al.,2015; Churi et al., 2015). Owczarczuk (2010) validated efficacy of prevalent data mining models to estimate churn and establish that linear models, specifically logistic regression, when modelling based on two groups that is churn of the prepaid clients/post-paid subscribers. The most important aspect of a customer retention program is whether it is proactive or reactive. In proactive method, the company forecasts customer behaviour using data analytics and design strategies based on statistical test and calculations. If the company collects more information about the customers and if their data analytics and retention strategies are in place, then they will be better prepared for driving loyalty (Agyeman,

2013). The important drivers for customer retention are: customer satisfaction, trust and customer engagement with the company (Ranaweera and Prabhu, 2003; Gounaris, 2005, Richards, 1996). They explain that if the customers are more engaged with the company, there are chances that they will remain loyal for the company. An engaged customer will first try to resolve the problem with the service provider instead of moving to other service provider. Another strong enabler to strengthen the bond between customer and the company is an effective communication and timely sharing of information between the customer and the client (Sharma and Patterson, 2000).

The studies investigating the relationship between customer satisfaction and customer loyalty is ambiguous, in the sense that few companies believed that having satisfied customers lead to customer loyalty. Customer satisfaction is usually companies' top priority. However, other researchers suggest that satisfaction alone does not ensure loyalty. Studies have explored the impact of relative satisfaction on customer loyalty. Dick and Basu (1994) and Keaveney (1995) claim that service providers must evaluate the strategy adopted by the other service providers to device plan for increasing loyalty of subscribers. The work done by Lee et al. (2001) demonstrates "the extent to which switching intention is influenced by the level of perceived switching costs and whether switching costs moderate the satisfaction-loyalty linkage".

CHAPTER THREE

METHODOLOGY AND DATA COLLECTION

3.1 **Introduction**: The basic layer for predicting future churn is from past data. We look at data from customers already have churned (response) and their characteristics (predictors) before the churn happened. By fitting a statistical model that relates the predictors to the response, the response for existing customers is predicted. The scope of work to forecast customer attrition may look like the following:

- Data collection

- Data preprocessing and preparation

- Modelling and testing

- Model deployment and monitoring

Also, it is important to understand what insights one needs to get from the analysis. We must decide what question to ask and consequently what type of machine learning problem to solve: classification or regression.

Classification: the goal of classification is to determine to which class or category a data point (that is, customer) belongs to. For classification problems, historical data is used with predefined target variables, that is labels (churner/non churner) that need to be predicted to train an algorithm. With classification, business can answer the following questions:

- Will this customer churn or not?

- Will a customer renew their subscription?

- Will a user downgrade a pricing plan?

- Are there any signs of unusual customer behaviour?

Regression: Customer churn prediction can also be formulated as regression task. Regression analysis is a statistical technique to estimate the relationship between a target variable and other data values that influence the target variable, expressed in continuous values. The result of regression is always some number, while classification always suggest a category. Furthermore, regression analysis allows for estimating how many different variables in data influence a target variable. With regression, businesses can forecast in what period of time, a specific customer is likely to churn or receive some probability estimate of churn per customer.

3.2   Research Design: As aforementioned, we decided to use the machine learning process proposed by Marsland (2014) as shown below. The proposed machine learning process includes data collection and preparation, feature selection, algorithm choice, parameter and model selection, training and evaluation. We do not consider the proposed machine learning process as a linear process; there is a lot of back and forth between the steps in other to create the most optimal model, in our case a churn prediction model.
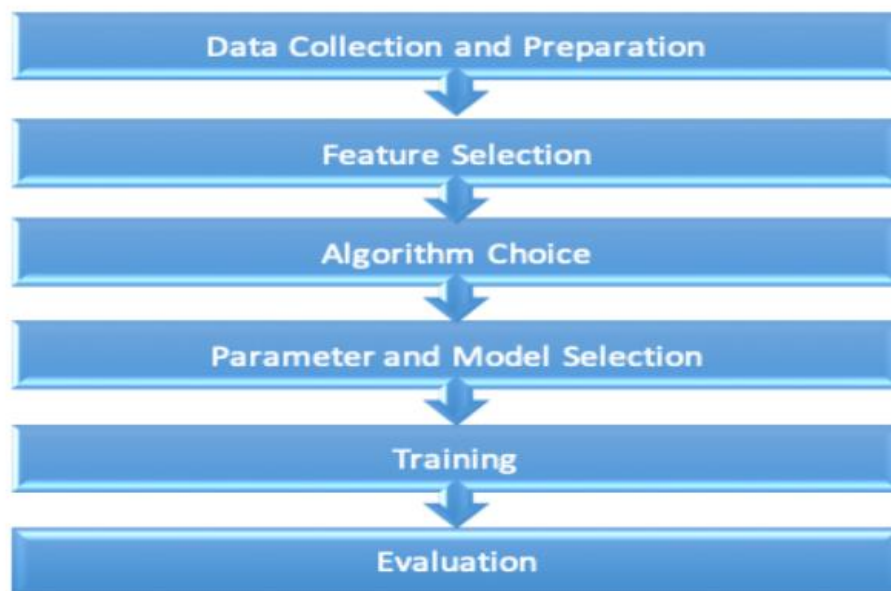

Figure 2 Machine learning process proposed by Marsland (2014)

3.2.1. Data Collection and preparation: Once kinds of insights to look for are identified, the data sources necessary for further predictive modelling can be decided. The dataset used for this project contains demographic details of customers, their total charges and the type of service they receive from the company. It comprises of churn data of over 3333 customers spread over 11 attributes obtained from the Global AI. (As shown in Figure 3). It can be used to analyse all relevant customer data and develop focused customer retention programs.

```
# Read csv
data = pd.read_csv("churn.csv")
data.head()
```

	Churn	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins	DayCalls	MonthlyCharge	OverageFee	RoamMins
0	0	128	1	1	2.7	1	265.1	110	89.0	9.87	10.0
1	0	107	1	1	3.7	1	161.6	123	82.0	9.78	13.7
2	0	137	1	0	0.0	0	243.4	114	52.0	6.06	12.2
3	0	84	0	0	0.0	2	299.4	71	57.0	3.10	6.6
4	0	75	0	0	0.0	3	166.7	113	41.0	7.42	10.1

Figure 3 A Snapshot of the dataset been used

In the given figure above, each row represents a customer, and each column contains customer's attributes described on the column Metadata.

3.2.2. Feature Selection: Consists of determining the most impactful features for a problem. Feature selection is used in order to identify the most relevant features and is often used due to its performance enhancing properties (Tang et al. 2014). This requires in depth knowledge of the data being used to determine which features should be included (Marsland, 2014). According to Tang et al., (2014) too many features can result in overfitting, meaning that the model learns from what can be considered noise to the point that the result is negatively affected. This means that learning from noise complicates generalization and the ability of correctly classify unseen data instances. We decided to identify non-important features using a less complicated and straightforward approach, where the importance of the features is scored

based on their prediction ability. We ranked the features based on their importance score and removed all non-important features with a score below 0.05. There exist more advanced methods for feature selection, such as recursive feature elimination (RFE) (Chen and Jeong, 2007). We decided not to use such methods since they are more time consuming and exceed the purpose of our study. The results we obtained were sufficiently good using feature importance. Since Naïve Bayes assumes that each feature is independent, feature selection using feature importance was not performed.

3.2.3. Algorithm Choice: Based on the scope of the problem, the optimal algorithm for the considered problem can vary. There are numerous algorithms to choose from when faced with a problem, for instance prediction tasks. However, not every algorithm is suitable and fits the criteria for the task at hand, which is why several algorithms are rejected. Since the problem we are trying to solve is a classification problem, all algorithms used for regression problems are rejected. Common classifiers, such as Naïve Bayes, Random Forest, XGBoost, Support Vector Machine, Logistic Regression, and Hidden Markov's Model are used in prediction problems, however previous studies show that ensemble learners are preferred for classification problems and they are rapidly becoming the standard choice among algorithms due to their performance enhancing characteristics (Kumar and Jain, 2020; Van Wezel and Potharst, 2007). Hence, our choice of algorithms is based on previous studies. We chose to use Decision Tree, which applies classification and regression tree algorithm (CART); Random Forest, which applies bagging and Logistic Regression which applies stacking, which performs well for several real-world problems including churn prediction.

3.2.4. Parameter and Model Selection: Many machine learning algorithms need to be fed with parameters such as the number of estimators. Parameter turning is the process to optimize the parameters of a machine learning algorithm and choosing the ones that obtain the best result for the desired problem. In our study, we investigate how machine learning can be used to

predict customer churning within the telecommunication industry; hence we considered parameter tuning to be beyond the scope of this study. In addition, we considered it to be too big of a task to accomplish. We decided to use default parameters and values for our algorithms and models.

3.2.5. Training: The process of training refers to using labelled data to build a model, which preferably performs well on unseen instances. The dataset should be split in a training set used for training and fitting of a model, and a test set used for evaluation. As mentioned above, supervised learning refers to learning by examples, meaning that each example contains inputs (features) and a corresponding output. To train the model, correlations are found between inputs and outputs. We decided to use supervised learning since our dataset is labelled and includes inputs with the corresponding output for each customer. We decided to randomly split the dataset into a training set and a test set with the proportions of 70:30 percent. Unfortunately, there is no distinct answer to how the split between training- and test data should be done and the choice is made from the researchers. Previous studies related to machine learning have used different splits such as 50/50, 60/40, 70/30, 80/20, and 90/10, but there is no research that suggests that one proportion preferred over another (S. Raschka, 2018). However, researchers argue that the bigger the dataset, the less test data is needed. For example, if there is data that contains 200 000 data instances, 10% can be more than enough to see if the model performs adequately but if the sample size contained 3000 data instances, 10% might be too little to validate the performance of the model and you might need to choose a larger test sample. On the other hand, there is no distinct answer to what a big or small dataset is. After consideration and a discussion with a domain expert we came to the conclusion that our dataset, including approximately 3333 data instances, is relatively big and we chose to split our dataset with the proportions of 70:30 percent.

3.2.6. Evaluation: Evaluation is an important part of the machine learning process, where the performance of the classifier should be evaluated. In order to select the optimal machine learning algorithm different indicators are used for different types of problems (Hongquig et al. 2018). In previous literature regarding churn prediction, the majority of researchers evaluate their model's using accuracy, precision, recall, and F-measure, which are all calculated from the confusion matrix. We chose to evaluate the performance and efficiency of our algorithms using the above-mentioned metrics. In the case of imbalanced data, accuracy is not the most optimal metric used for evaluation since it does not always fully reflect the performance of the algorithm (Tyagi and Mittal, 2020). To handle this problem, we evaluate the classifiers based on precision, recall, and F1-score on the target class, that is Churn. However, we will focus on F1-score as it illustrates a balance between precision and recall. In our application, precision, that is, the rate of correctly classified churn instances, is less important relative to recall, which measures the model's ability to predict actual churners. Evaluation is done on unseen instances from the test set, in other words data, which the algorithm has not been trained on. This is why the data should be split into a training and a test set to be able to determine how well the model performs on examples not used during training.

3.3   UML Diagram:

3.3.1   Data Flow Diagram: The data flow diagram shown below shows the flow of data right from acquiring the dataset to model building, extracting features importance and comparison of model performances.
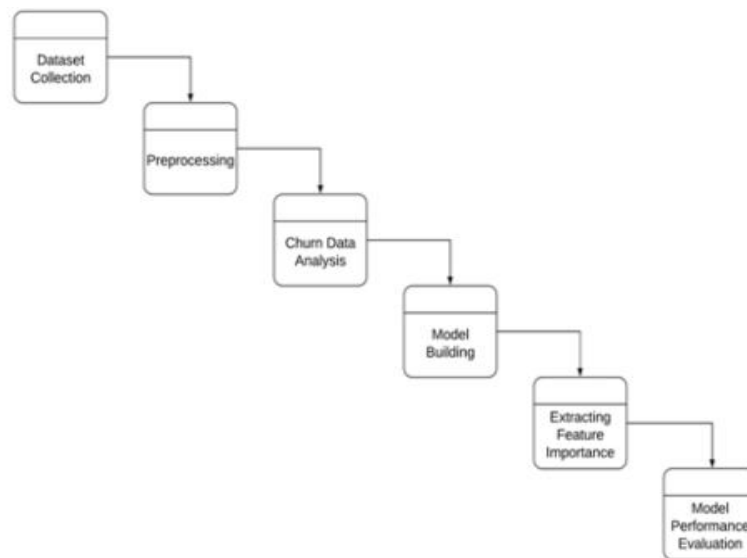
Figure 4 Data flow diagram for churn prediction

3.3.2 Sequence Diagram: The sequence diagram shown below shows the sequence of executing of various processes with acquiring dataset, preprocessing, feature extraction, predicting results and accuracy determination.
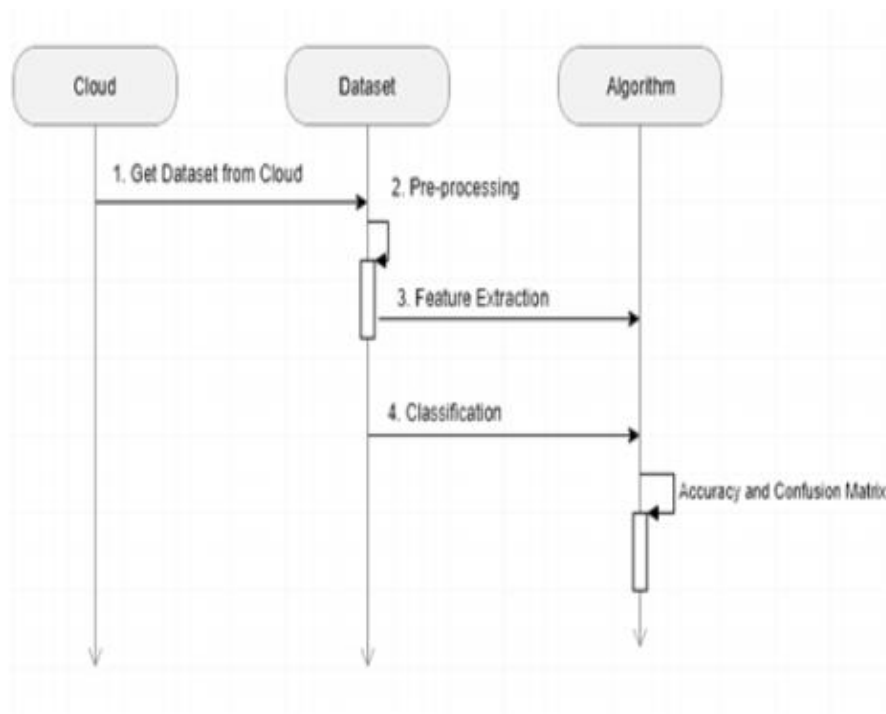

Figure 5 Sequence Diagram for customer Churn Prediction

3.4   Machine Learning Algorithms: Machine learning is a part of artificial intelligence and thereby a part of data science, which focuses on developing models in the form of algorithms that learn from data and experience (Jiang et al. 2020; Mehta et al. 2019). By training algorithms on data, they can improve their decision-making and the accuracy of their predictions over time (Jiang et al 2020). Machine learning is a frequently used approach considering automation of a variety of tasks and is categorized in different types of learning based on how the algorithms learn to become better in form of accuracy. These types of learning are usually categorized as the following: Supervised learning, unsupervised learning, semi supervised learning and reinforcement learning and based on the scope of the problem and the type of data used, one algorithm could be more suitable than the others (Simeone 2018; Marsland 2014).

According to Simeone (2018) and Mehta et al. (2019), supervised learning is often described as learning by examples, meaning that the algorithm uses labelled data for training. Unlike supervised learning, unsupervised learning is a technique where the algorithms do not learn from a supervisor, meaning that the dataset is unlabelled. The algorithms need to find their own patterns from the input. Semi-supervised learning combines the above-mentioned techniques but the algorithms use more of the unlabelled dataset than the labelled for training. Reinforcement learning is a technique, where the algorithm learns by employing a reward and punishment system to reach an end goal.

3.5   Supervised Learning: Supervised learning is considered a sub-category of machine learning, which is frequently used within prediction problems (Jiang et al. 2020). Supervised learning can be described as learning by examples, which means that the algorithms are trained by a labelled dataset (Mohammed et al 2016). A labelled dataset is a dataset, which consists of both inputs and outputs, and the main idea is to train the algorithm by mapping these inputs and outputs from the dataset (Mohammed et al 2016). This type of learning works by finding

patterns in the dataset in order to identify outputs of unseen instances (Oral et. al, 2012). Supervised learning can be further divided into regression and classification (V. Verdhan, 2020). A regression problem is where the output is a continuous value, such as the days until a potential customer might churn. Classification is on the other hand used when the output is categorical, for instance churn or not churn (V. Verdhan, 2020). The following supervised machine learning algorithms have been used for predicting customer churn:

3.5.1   Logistic Regression: is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where two values are labelled "0" and "1".

Also, Logistic regression is a fundamental classification technique. It belongs to the group of linear classifiers and is somewhat similar to polynomial and linear regression. Logistic regression is fast and relatively uncomplicated, and it's convenient for you to interpret the results. Although it's essentially a method for binary classification, it can also be applied to multiclass problems.

Logistic regression is named for the function used at the core of the method, logistic function. The logistic function, also called the sigmoid function is an S-shaped curve (as shown in Figure below) that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.
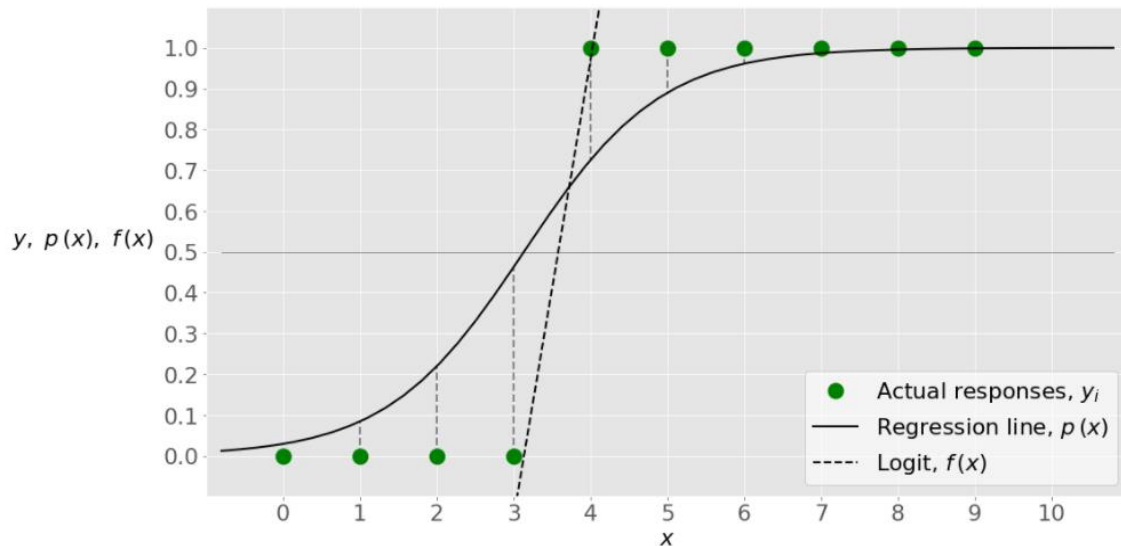
Figure 6 Logistic function

3.5.2   Decision Tree: Decision tree learning is one of the predictive modelling approaches that uses a decision tree (as a predictive model) to go from observations about an item i.e., attribute (represented in the branches) to conclusions about the item's target value i.e. churn or not (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. This algorithm splits a data sample into two or more homogeneous sets based on the most significant differentiator in input variables to make a prediction. With each split, a part of a tree is being generated. As a result, a tree with decision nodes and leaf nodes (which are decisions or classifications) is developed. A tree starts from a root node – the best predictor. Below is a decision tree structure.
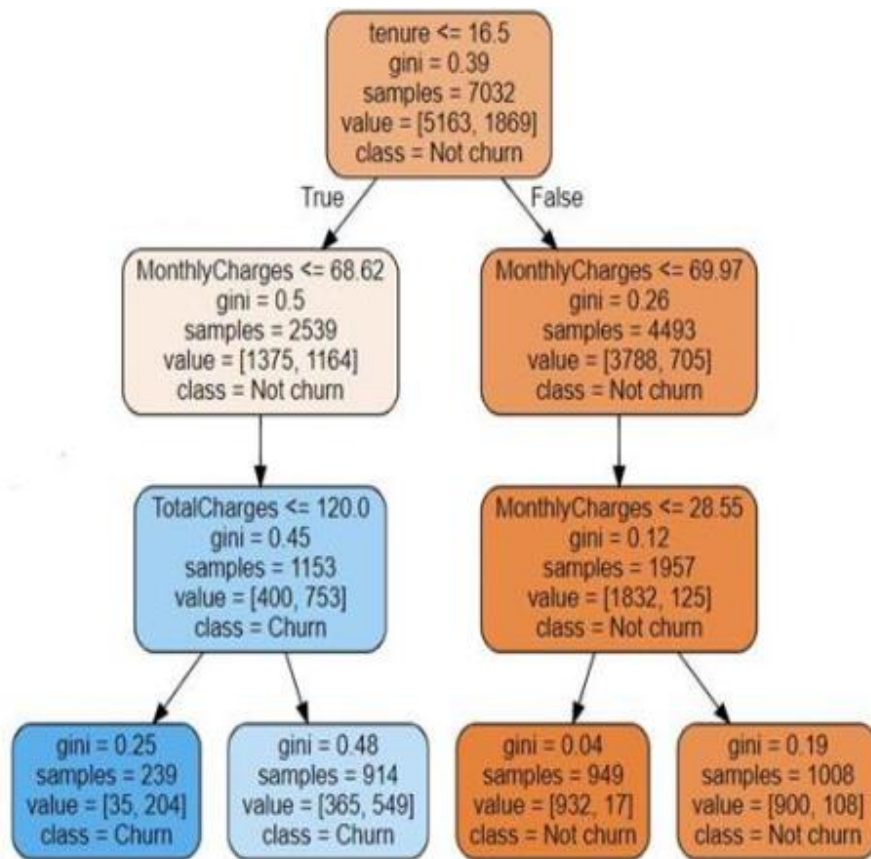
Figure 7 Basic structure of a decision tree

Prediction results of decision trees can be easily interpreted and visualized. Even people without an analytical or data science background can understand how a certain output appeared. Compared to other algorithms, decision trees require less data preparation, which is also an advantage. However, they may be unstable if any small changes were made in data. In other words, variations in data may lead to radically different trees being generated.

3.5.3 Random Forest: Random Forest classifier is an ensemble learner, which is considered to be a supervised learning algorithm. The Random Forest classifier utilizes the bagging approach consisting of multiple classifiers, where the base learners are decision trees Each decision tree takes a random sample of the original training dataset with replacement, which means that some data is used more than once in training. However, all features are not utilized

for every decision tree. In the Random Forest classifier, the division of the nodes in the trees are based on the best split of a random subset of features rather than using the best split considering all features. This reduces the correlation between the trees and decreases the generalization error. The diversity of each tree is increased in Random Forest due to the fact that different subsets of the training data is used for each tree, which leads to the classifier being more stable and robust towards noise and overtraining. Each individual decision tree in the Random Forest contributes to the final output by majority voting. The class with the most votes is chosen as the final output, see the figure below.
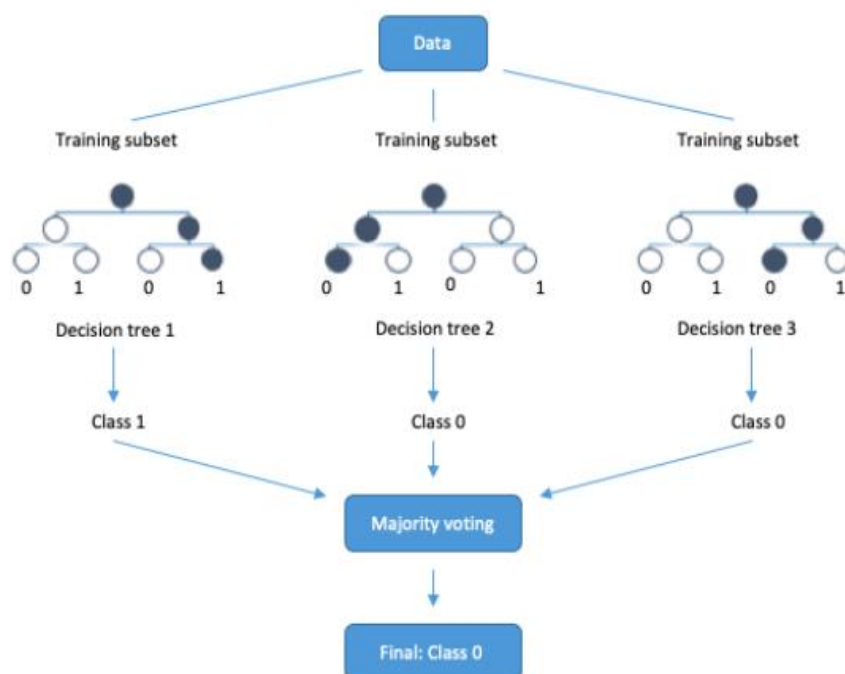

Figure 8 Random Forest classifier

3.6   Evaluation Metrics: Machine learning models are not always perfect for the given data and are in need of evaluation to see how well the model performs. The most common ways to evaluate binary classifiers are using certain metrics, which are accuracy, recall, precision and F1-score (Sokolova and Lapalme, 2009). These metrics can be calculated through a confusion matrix, shown in table below.

	Actual	
Predicted	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

Table 1 Confusion Matrix for binary classification

A confusion matrix is a machine learning concept, which includes information about the actual and predicted classifications, used to describe the performance of the classifier (Deng et al. 2016). True Positives (TP) and True Negatives (TN) represent the correctly classified test instances while False Negatives (FN) and False Positives (FP) represent the incorrectly classified test instances (Witten, 2016).

Accuracy is a measure that shows the overall effectiveness of the classifier (Sokolova and Lapalme, 2009). It is a metric showing the rate of total correctly classified instances. According to Deng et al. (2016), accuracy is defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \qquad\qquad (Eq.1)$$

Precision is a measure that shows the proportion of correctly predicted positive instances. The metric shows how often the model is correct when predicting the target class, in our application, churners. According to Deng et al. (2016) precision shows the accuracy of predicting a specific class and is calculated as follows:

$$Precision = \frac{TP}{TP+FP} \qquad\qquad (Eq.2)$$

Recall is a measure, which shows the effectiveness of the classifier to determine examples labelled as positive (Sokolova and Lapalme, 2009). It shows the ability of the binary classifier to identify instances of a specific class (Deng et al. 2016). Recall is calculated as follows:

$$Recall = \frac{TP}{TP+FN} \qquad\qquad (Eq.3)$$

The F-score is often used for evaluating the performance of a classifier. The F-score is a measure that takes both precision and recall into consideration and is typically defined as the harmonic mean of precision and recall. A better combined recall and precision is achieved as F-score is closer to 1. (Vafeiadis et al. 2015)

$$F - Score = \frac{2*Precision*Recall}{Precision+Recall}$$ (Eq.4)

Chapter Four

Data Analysis

4.1   Data Visualization: Before training a machine learning model, it's always a good idea to explore the distributions of various columns and see how they are related to the target column. The figures below show most relevant features. we notice that except for the number of integrations, all the other features have an almost uniform distribution, but with a churn rate that varies a lot along the distribution. These features can thus be considered as good explanatory features. Further modelling will show that they are as well of great importance in churn prediction.
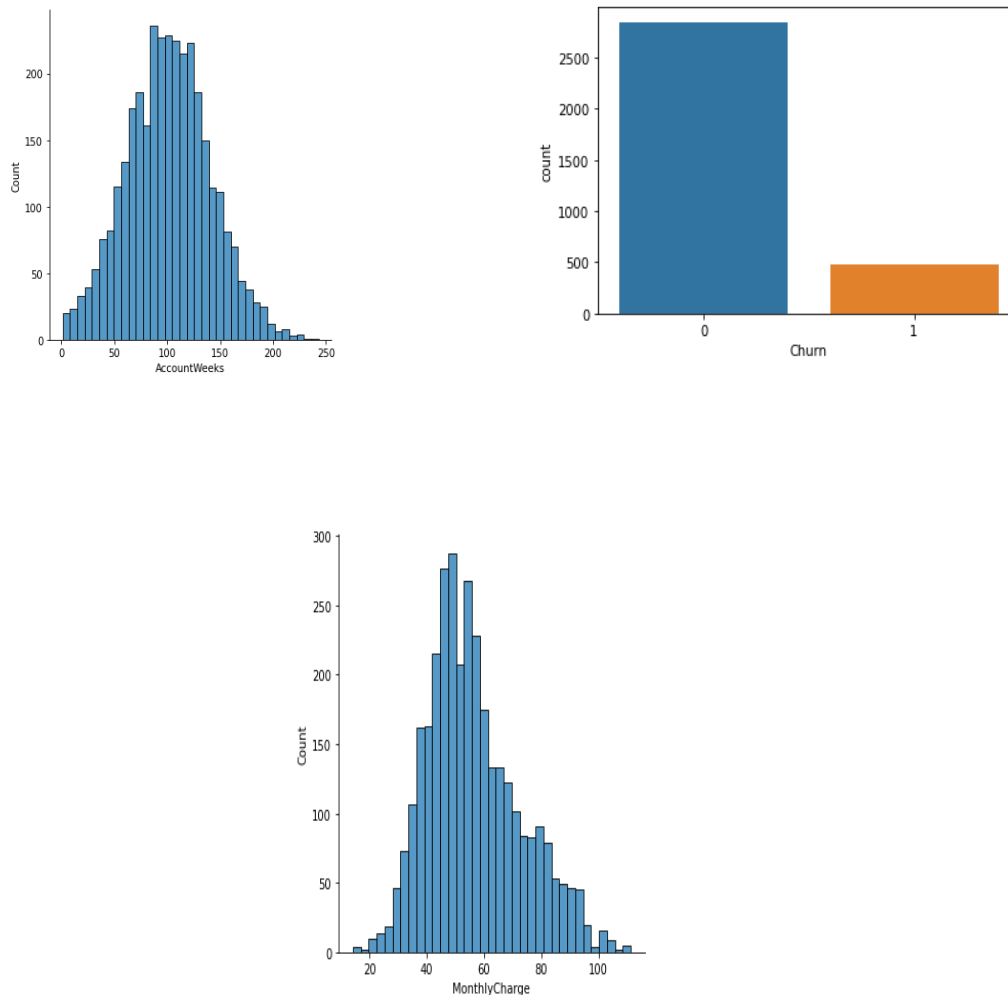

Figure 9 Churn breakdown with some key factors for churning

4.2   Confusion Matrix Analysis: The following model fit on a training resulting from a simple stratified train-test spilt. No resampling technique is implemented.

Figure 10 Confusion Matrix - Decision Tree

```
              precision    recall  f1-score   support

           0       0.94      0.92      0.93      2001
           1       0.56      0.63      0.59       333

    accuracy                           0.88      2334
   macro avg       0.75      0.77      0.76      2334
weighted avg       0.88      0.88      0.88      2334
```
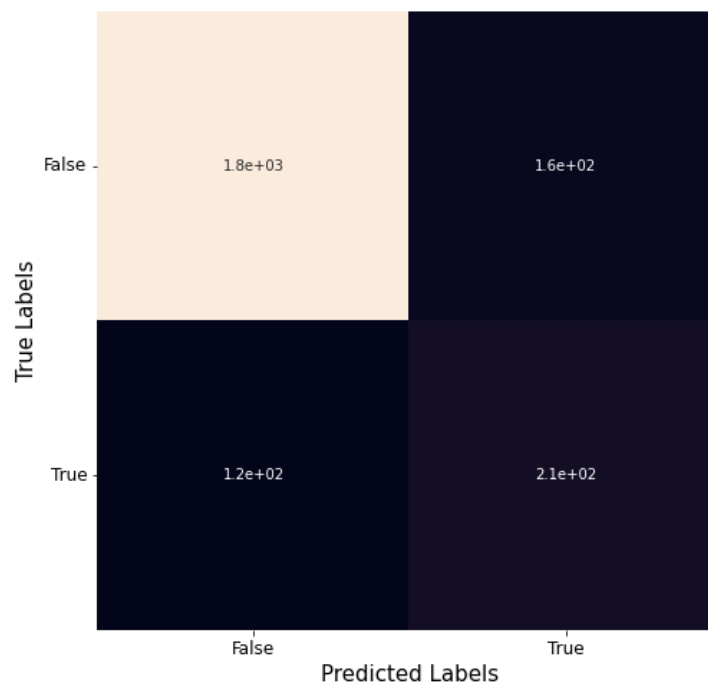
Figure 11 Precision, Recall, F1-Score -Decision Tree

The model figure above has an accuracy of approximately 88% but the confusion matrix has a specificity of 63.6% of the model indicating the actual churners that are correctly identified.

Figure 12 Confusion Matrix - Random Forest

```
              precision    recall  f1-score   support

           0       0.93      0.98      0.96      2001
           1       0.81      0.58      0.68       333

    accuracy                           0.92      2334
   macro avg       0.87      0.78      0.82      2334
weighted avg       0.92      0.92      0.92      2334
```
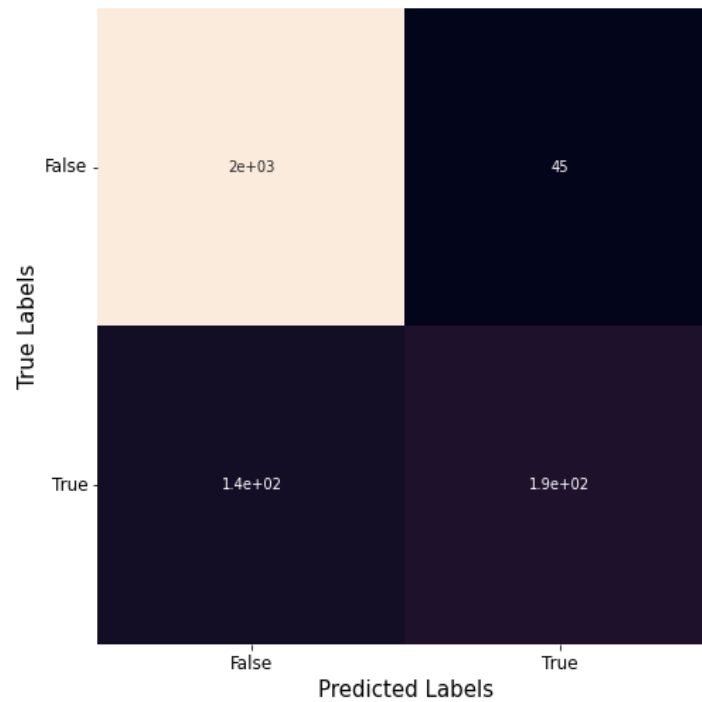
Figure 13 Precision, Recall, F1-Score -Random Forest

Also, the model figure above has an accuracy of approximately 92% but the confusion matrix has a specificity of 57.6% of the model showing the actual churners that are correctly identified.

Figure 14 Confusion matrix - Logistic Regression

```
              precision    recall  f1-score   support

           0       0.88      0.98      0.93      2001
           1       0.58      0.17      0.27       333

    accuracy                           0.86      2334
   macro avg       0.73      0.58      0.60      2334
weighted avg       0.83      0.86      0.83      2334
```
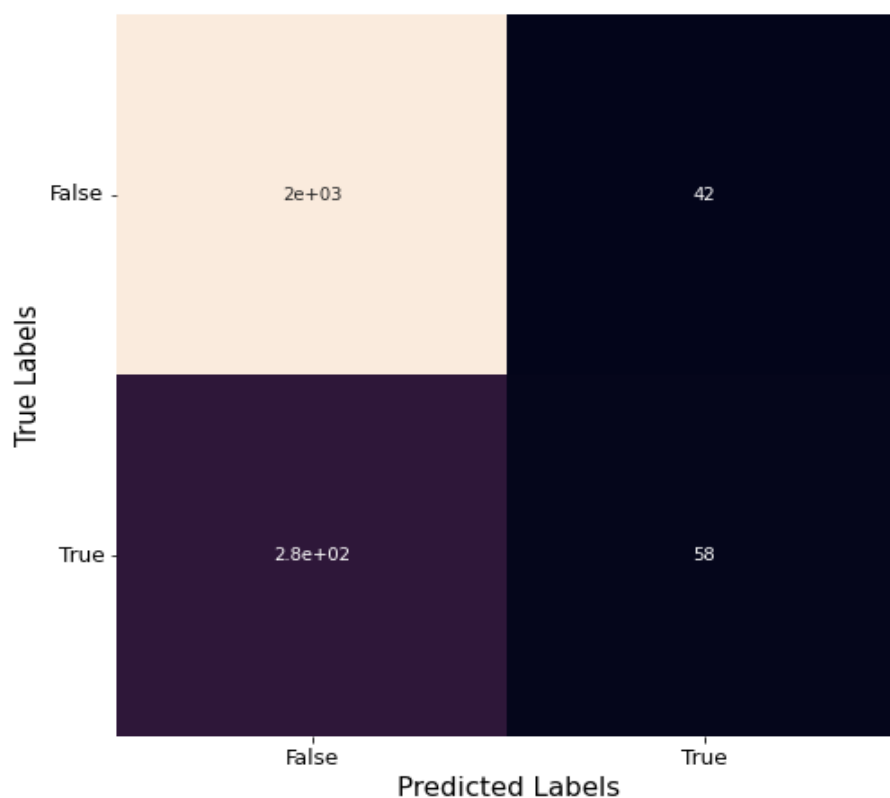
Figure 15 Precision, Recall, F1-Score - Logistic Regression

Furthermore, the model figure above has an accuracy of approximately 86% but the confusion matrix has a specificity of 17.16% of the model showing the true churners that have been detected.

4.3   Comparison of Models: A thorough comparison of algorithms based on the metrics mentioned above gives a comprehensive insight into the performance and efficiency of each of them. From the table below, we observe that by comparing the three models, random forest seems to yield the highest level of accuracy. it also has a better f1-score for predicting customer churn, we concluded that Random Forest is the best model of the three models used.

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.8770351328191945	0.5618279569892473	0.6276276276276276	0.5929078014184397
Random Forest	0.921165381319623	0.8117154811715481	0.5825825825825826	0.6783216783216783
Logistic Regression	0.8641816623821765	0.58	0.17417417417416	0.2678983833718245

Table 2 comparison of model using evaluation metrics

More so, accuracy could be a misleading metric in the case of an imbalanced dataset, since it does not tell much on how the classifiers perform on specific classes. However, in our research it is more important to predict actual churners with a high accuracy and minimizing False Negatives rather than minimizing the number of False Positives, that is, non-churners classified as churners. In Table 2 above, we present the performance of the classifiers for the target class, that is churners. In our study, the focal point is to predict churners with a good accuracy. By analysing the results, Logistic Regression does not perform well on predicting churners, in particular with the default imbalanced dataset, the results were far from adequate. The Random Forest classifier and the Decision Tree classifier perform remarkably

better on churn prediction than the Logistic Regression classifier does. The Decision Tree classifier and the Random Forest Classifier perform similarly.

4.4 **Performance of Models**: Model evaluation is a big part of Design Science and with a great variety of metrics and measures, the choice of which metric(s) to use is not always obvious, and it depends on the considered application. Our results show that for the considered classifiers, Logistic Regression, Decision Tree, and Random Forest, the overall accuracy ranges between 86.4% and 92.1%, indicating an overall good performance. On the other hand, one can question what a good and suitable score of accuracy actually is. Accuracy measures the classifiers' ability of correctly classifying all the test instances, in our case including both non-churners and churners. But what is actually a good score? This is a common question when evaluating machine learning algorithms, and unfortunately there is no distinct answer to this question. We argue that a good score of accuracy depends on the considered application and what prediction problem is investigated. 90% accuracy in one application may not be as good as 90% in another application. For example, for email spam detection, 90% accuracy is not bad, meanwhile that only 10% of the emails are wrongly classified. In medical predictions, regarding potential patients in a certain risk group, 10% wrong classifications are far from acceptable. It should be emphasized that accuracy alone is not a sufficient metric used for evaluation (Jackson et al. 2017), which is also shown by the above examples. In particular, when the considered dataset is imbalanced to a high degree, accuracy can be a misleading metric. In our case, the used dataset consists of 98% non-churners and 2% churners, which in fact could give us a 98% accuracy by predicting all test instances as non-churners. If this would be the case, the result is far from acceptable and a 98% accuracy would not be good. Based on this reasoning, accuracy gives in our case a skewed view of the actual performance of the classifiers. Based on the above reasoning, we can therefore not assume that the performance of the classifier is good by only looking at accuracy. Alternative evaluation metrics, which are commonly used in the liter

ature are precision, recall, and F1-score, all indicating the performance of the classifier on a s pecific target class. In our case the target class is churn, and our results based on the above-m entioned alternative metrics show that, except for Decision Tree and Logistic Regression, Ran dom Forest has the best obtained precision of 81.1%, recall of 58.3%, and F1-score of 67.8% i s actually good at predicting churners in our application. This further proves our point stated a bove that accuracy can be a misleading metric.

CHAPTER FIVE

CONCLUSION AND FUTURE WORK

5.1   Conclusion: Churn prediction is one of the most effective strategies used in telecom sector to retain existing customers. It leads directly to improved cost allocation in customer relationship management activities, retaining revenue and profits in future. It also has several positive indirect impacts such as increasing customer's loyalty, lowering customer's sensitivity to competitors marketing activities, and helps to build positive image through satisfied customers. The results predicted by the Random Forest algorithm were the most efficient with an accuracy of 92.1%. Therefore, companies that want to prevent customer churn should utilize this algorithm thereby giving them more flexibility and providing additional services such as device protection and multiple phone lines proves to be of little value to customer attrition.

More so, focusing on enhancing the experience of loyal customers who have stayed with the company for long will prove worthwhile, ensuring their retention. The ability to identify customers that aren't happy with provided solutions allows businesses to learn about product or pricing plan weak points, operation issues, as well as customer preferences and expectations to proactively reduce reasons for churn.

Lastly, we consider our findings valuable for researchers in our field of study, but also for organizations developing churn management systems as a part of their customer relationship management. In particular, our findings fill the gap in the telecommunication sector, and give researchers and firms more knowledge regarding churn prediction in the subscription-based service context, which is still underdeveloped.

5.2   Future Work: An important area for future research is to use a customer profiling methodology for developing a real-time monitoring system for churn prediction. Research dedicated to the development of an exhaustive customer loyalty value would have significant

benefits to industry. It is anticipated that the profiling methodology could provide an insight into customer behaviour, spending patterns, cross-selling and up-selling opportunities. Seasonal trends could be apparent if the same data was studied over a period of several years. However, a comparative analysis of prediction model building time with respect to different classifiers could be done in order to assist telecom analysts to pick a classifier which not only gives accurate results in terms of True positive (TP) rate, Area under curve (AUC) and lift curve but also scales well with high dimension and large volume of call records data. As concrete findings are related to the telecom dataset, other domains' datasets might be subject for further exploration and testing. Also, different and a greater number of performance metrics with respect to business context and interpretability might be explored in future.

Another aspect, which would be interesting to further investigation, is what other methods can be used for feature selection and sampling and how they would impact the result. In addition, what other algorithms that could be used and how they could be modified in other to achieve the best result.

REFERENCES

A. Kumar and M. Jain, (2020) Ensemble Learning for AI Developers: Learn Bagging, Stacking, and Boosting Methods with Use Cases, Apress.

A. Payne and P. Frow, (2005) "A Strategic Framework for Customer Relationship Management," Journal of Marketing, vol. 69, no. 4, pp. 167-176.

A. T. Jahromi, S. Stakhovych and M. Ewing, (2014) "Managing B2B customer churn, retention and profitability," Industrial Marketing Management, vol. 43, no. 7, pp. 1258-1268.

Agyeman, C.M. (2013) A Study on Customer Satisfaction and Retention in the Telecommunication Industry; an Empirical Study of the New Juaben Municipality, Lulu.com Publications.

Almana, et al. (2014) International Journal of Engineering Research and Applications ISSN: 2248–9622, Vol. 4, No. 5 (Version 6), pp.165–171

Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, Hawalah A, and Hussain A. (2016) Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study. IEEE Access. 4:7940–57

Ang, L. and Buttle, F. A. (2002) ROI on CRM: a customer journey approach. 18th IMP-conference. Australia, IMP Group in Asia.

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

Ascarza E, Iyengar R, and Schleicher (2016) M. The perils of proactive churn prevention using plan recommendations: evidence from a feld experiment. J Market Res; 53(1):46–60

B. J. Jacksson, P. Korfiatis, Z. Akkus and T. L. Kline, (2017, February)"Machine Learning for Medical Imaging," Radio-Graphics, vol. 37, no. 2, pp. 505-515.

Balasubramanian, M. and Selvarani, M (2014), 'Churn prediction in mobile telecom system using data mining techniques', International Journal of Scientific and Research Publications, Vol. 4, No. 4, pp.1–5

Basu, K. and Dick, A.S. (1994) 'Customer loyalty: toward an integrated conceptual framework', Journal of the Academy of Marketing Science, Vol. 22, No. 2, pp.99–113

Bott. (2014) Predicting customer churn in telecom industry using multilayer perceptron neural networks: modelling and analysis. Igarss;11(1):1–5

Brandusoiu I, Toderean G, and Ha B. (2016) Methods for churn prediction in the prepaid mobile telecommunications industry. In: International conference on communications. p. 97–100

Burez D and Van den Poel. (2009) Handling class imbalance in customer churn prediction. Expert Syst Appl. 36(3):4626–36

C. Blank and T. Hermansson, (2018)"A Machine Learning approach to churn prediction in a subscription-based service," KTH, Stockholm.

Chen, I. J. and Popovich, K. (2003) Understanding Customer Relationship Management (CRM). Business Process Management Journal, 9, 672-688

Churi, A., Divekar, M., Dashpute, S. and Kamble, P. (2015) 'Analysis of customer churn in mobile industry using data mining', International Journal of Emerging Technology and Advanced Engineering, Vol. 5, No. 3, pp.225–230

Cisco Visual Networking Index (2017) Forecast and Trends, 2017–2022. Available online at: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-indexvni/white-paper-c11-741490.html

Cronin JJ, Brady MK, and Hule GT (2000) Assessing the effects of quality, value, and customer satisfaction on customer behavioural intentions in service environments. Journal of Retailing 76(2):193–218

D. Buo and M. Kjellander, (2014)"Predicting Customer Churn at a Swedish CRM-system Company," Linkoping's Universitet, Linkoping.

D. L. Garcia, A. Nebot and A. Vellido, (2017)"Intelligent data analysis approaches to churn as a business problem: a survey," Knowledge and Information Systems, vol. 51, no. 3, pp. 1-56.

Dahiya, K. and Talwar, K. (2015) 'Customer churn prediction in telecommunication industries using data mining techniques – a review', International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, No. 4, pp.417–433.

Deng X, Liu Q, Deng Y, and Mahadevan S (2016) An improved method to construct basic probability assignment based on the confusion matrix for classification problem. Inf Sci 340–341:250–161

F. Khodakarami and Y. Chan, (2014) "Exploring the role of customer relationship

management (CRM) systems in customer knowledge creation," Information & Management, vol. 51, pp. 27-42.

Farquad M, Ravi V, and Raju S (2014) Churn prediction using comprehensible support vector machine: An analytical CRM application. Applied Soft Computing 19:31–40

Fill, C., (2005) Marketing Communications: Contexts, Strategies, and Applications, 4th ed., Pearson Education Limited, New York.

G. Hongqing, S. Peiyong, G. Wenzhong and G. Kun, (2018) "Component-based Assembling Tool and Runtime Engine for the Machine Learning Process," in International Conference on Cloud Computing, Big Data and Blockchain.

Garcıa DL, Vellido A, Nebot A (2007) Customer continuity management as a foundation for `churn data mining. Technical report LSI-07-2-R, Universitat Polit`ecnica de Catalunya (UPC), Barcelona, Spain.

Gerpott TJ, Rams W, Schindler A. (2001) Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market. Telecommunication Policy. 25:249–69

Glatz, E., Mavromatidis, S., and Ager, B. et al. (2014) 'Visualizing big network traffic data using frequent pattern mining and hypergraphs', Computing, Vol. 96, No. 1, pp.27–38

Gould, G. (1995) 'Why it is customer loyalty that counts (and how to measure it)', Managing Service Quality: An International Journal, Vol. 5, No. 1.

Gounaris, S.P. (2005) 'Trust and commitment influences on customer retention: insights from business-to-business services', Journal of Business Research, Vol. 58, No. 2, pp.126–140

Guo-en X, Wei-dong J (2008) Model of customer churn prediction on support vector machine. Syst Eng Theory Pract 28(1):71–77

H. Gebert, M. Geib, L. Kolbe and W. Brenner, (2003) "Knowledge-enabled customer relationship management: integrating customer relationship management and knowledge management concepts," Journal of Knowledge Management, vol. 7, no. 5, pp. 107-123.

Hadden J, Tiwari A, Roy R, Ruta D (2007) Computer-assisted customer churn

management: State-of-the-art and future trends. Computers and Operations Research 34(10):2902–2917

He Y, He Z, Zhang D. (2009) A study on prediction of customer churn in fixed communication network based on data mining. In: Sixth international conference on fuzzy systems and knowledge discovery, vol. 1. p. 92–4

Howard (2010) Customer Retention: The Art of Keeping Good Customers. Available online at: http://howard-marketing.com/speaking-training/customer-retention-the-art-of-keeping-goodcustomers-2/

Huang F, Zhu M, Yuan K, Deng EO. (2015) Telco churn prediction with big data. In: ACM SIGMOD international conference on management of data. p .607–18

Hung S, Yen DC, Wang HY (2006) Applying data mining to telecom churn management. Expert Systems with Applications 31(3):512–524

I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, (2019) "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector," IEEE Access, pp. 60134-60149.

I. H. Witten, E. Frank, M. A. Hall and C. J. Pal, (2016) Data Mining: Practical Machine Learning Tools and Techniques, San Francisco: Elsevier Science & Technology.

Idris A, Khan A, Lee YS. (2012) Genetic programming and adaboosting based churn prediction for telecom. In: IEEE international conference on systems, man, and cybernetics (SMC). p. 1328–32

J. Tang, S. Alelyani and H. Li, (2014) "Feature Selection for Classification: A Review," p. 37.

Jones MA, Mothersbaugh DL, Beatty SE (2000) Switching barriers and repurchase intentions in services. Journal of Retailing 70(2):259–270

Jones MA, Mothersbaugh DL, Beatty SE (2002) Why customers stay: measuring the underlying dimensions of services switching costs and managing their differential strategic outcomes. Journal of Business Research 55(6):441–450

K. Mishra and R. Rani, (2017)"An inclusive survey on machine learning for CRM: a

paradigm shift," in 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India.

Keaveney, S.M. (1995) 'Customer switching behaviour in service industries: an exploratory study', The Journal of Marketing, pp.71–82

Keramatia A, Jafari-Marandi R, Aliannejadi M, Ahmadian I, Mozaffari M, Abbasi U (2014) Improved churn prediction in telecommunication industry using datamining techniques. Applied Soft Computing 24:994–1012.

Khizindar, T.M., AI-Azzam, A.F.M. and Khanfar, I.A. (2015), 'An empirical study of factors affecting customer loyalty of telecommunication industry in the kingdom of Saudi Arabia', British Journal of Marketing Studies, Vol. 3, No. 5, pp.98–115

Kim MK, Park MC, Jeong DH (2004) The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services. Telecommunications Policy 28(2):145–159

L. Junxiang, (2002) "Predicting Customer Churn in the Telecommunications Industry — An Application of Survival Analysis Modelling Using SAS," Sprint Communications Company.

Lee, J. and Feick, L. (2001) 'The impact of switching costs on the customer satisfaction loyalty link: mobile phone service in France', Journal of Services Marketing, Vol. 15, No. 1, pp.35–48

Lin CS, Tzeng GH, and Chin YC (2011) Combined rough set theory and flow network graph to predict customer churn in credit card accounts. Expert Syst Appl 38:8–15

M. Masarifoglu and A. H. Buyuklu, (2019, November)"Applying Survival Analysis to Telecom Churn Data," American Journal of Theoretical and Applied Statistics, vol. 8, no. 6, pp. 261-275.

M. Mohammed, M. B. Khan and E. B. M. Bashier, (2016) Machine Learning: Algorithms and Applications, CRC Press.

M. Oral, E. L. Oral and A. Aydin, (2012)"Supervised vs. unsupervised learning for construction crew productivity prediction," Automation in Construction, vol. 22, pp. 271-276.

M. Sergue, (2020)"Customer Churn Analysis and Prediction using Machine Learning for a B2B SaaS company," KTH, Stockholm.

M. Sokolova and G. Lapalme, (2009, July) "A systematic analysis of performance measures for classification tasks," Information Processing & Management, vol. 45, no. 4, pp. 427-437.

M. Van Wezel and R. Potharst, (2007) "Improved customer choice predictions using ensemble methods," European Journal of Operational Research, vol. 181, no. 1, pp. 436-452.

Makhtar M, Nafs S, Mohamed M, Awang M, Rahman M, Deris M. (2017) Churn classification model for local telecommunication company based on rough set theory. J Fundam Appl Sci.9(6):854–68.

Miguéis VL, Van den Poel D, Camanho AS, and Falcao e Cunha J (2012) Modelling partial customer churn: On the value of first product-category purchase sequences. Expert Syst Appl 39:11250–11256

Minami, C. and Dawson, J. (2008) The CRM process in retail and service sector firms in Japan: Loyalty development and financial return. Journal of Retailing and Consumer Services, xxx, In Press, Corrected Proof

Mittal, B. and Lassar W. M. (1998) 'Factor affecting, selection of cell services: a cross-segmental study', Synergy, Vol. 4, No. 1, pp.74–85

N. Gordini and V. Veglio, (2017) "Customer's 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, vol. 62, pp. 100-107.

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):15,273–15,285

O. Simeone, (2018) "A Very Brief Introduction to Machine Learning with Applications to Communication Systems," IEEE Transactions on Cognitive Communications and Networking, vol. 4, no. 4, pp. 648-664.

Owczarczuk, M. (2010) 'Churn models for prepaid customers in the cellular

telecommunication industry using large data marts', Expert Systems with Applications, Vol. 37, pp.4710–4712

P. Lalwani, M. M. Kumar, J. Singh Chadha and P. Sethi, (2021) "Customer churn prediction system: a machine learning approach," Computing, pp. 1-24.

P. Mehta, M. Bukov, C. H. Wang, A. G. Day, C. Richardson, C. K. Fisher and D. J. Schwab, (2019) "A high-bias, low-variance introduction to Machine Learning for physicists," Physics Reports, vol. 810, pp. 1-124.

Patterson MP, Smith T (2003) A cross-cultural study of switching barriers and propensity to stay with service providers. Journal of Retailing 79(2):107–120

Qureshii SA, Rehman AS, Qamar AM, Kamal A, Rehman A. (2013) Telecommunication subscribers' churn prediction model using machine learning. In: Eighth international conference on digital information management. p.131–6

R. Shkurti and A. Muca (2014) "An Analysis of Cloud Computing and Its Role in Accounting Industry in Albania," Romanian-American University, vol. 8, no. 2, pp. 219-229.

Ranaweera, C. and Prabhu, J. (2003) 'The influence of satisfaction, trust and switching barriers on customer retention in a continuous purchasing setting', International Journal of Service Industry Management, Vol. 14, No. 4, pp.374–395

Reichheld, F.F. and Sasser, J.W. (1990) 'Zero defections: quality comes to services', Harvard Business Review, Vol. 68, No. 5, pp.105–111

S. Marsland, (2014) Machine Learning: An Algorithmic Perspective, 2nd edition ed., Chapman and Hall/CRC.

S. R. Gulliver, U. B. Joshi and V. Michell, (2013, October)"Adapted customer relationship management implementation framework: Facilitating value creation in nursing homes," Total Quality Management and Business Excellence, vol. 24, pp. 9-10.

S. Raschka, (2018) "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," University of Wisconsin–Madison.

S. Tyagi and S. Mittal, (2020) "Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning," in Proceedings of ICRIC 2019.

Saradhi W and Palshikar GK (2011) Employee churn prediction. Expert Syst Appl 38:1999–

2006

Seng JL and Chen TC (2010) An analytic approach to select data mining for business decision. Expert Syst Appl 37:8042–805

Sharma A and Panigrahi PK (2011) A neural network-based approach for predicting customer churn in cellular network services. Int J Comput Appl 27(11):26–31

Sharma, N. and Patterson, P.G. (2000) 'Switching costs, alternative attractiveness and experience as moderators of relationship commitment in professional, consumer services', International Journal of Service Industry Management, Vol. 11, No. 5, pp.470–490

T. Jiang, J. L. Gradus and A. J. Rosellini, (2020) "Supervised Machine Learning: A Brief Primer," Behaviour Therapy, vol. 51, no. 5, pp. 675-687.

Tsai CF, Lu YH (2009) Customer churn prediction by hybrid neural networks. Expert Systems with Applications 36(10):12,547–12,553.

Tsai, CF. and Lu, YH. (2010) 'Data mining techniques in customer Churn prediction', Recent Patents on Computer Science, Vol. 3, pp.28–32

Tsai C-F, Chen M-Y (2010) Variable selection by association rules for customer churn prediction of multimedia on demand. Expert Syst Appl 37:2006–2015.

V. Verdhan, (2020) Supervised Learning with Python: Concepts and Practical Implementation Using Python, publisher logo Apress.

Vafeiadis T, Diamantaras K, G S, Chatzisavvas K (2015) A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory 55:1–9.

Verbeke W, Dejaeger K, Martens D, Hur J, and Baesens B (2011) New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research 218:211–219

Wei CP, Chiu LT (2002) Turning telecommunications call details to churn prediction: a data mining approach. Expert Systems with Applications 23(2):103–12

X.W. Chen and J.C. Jeong, (2007) "Enhanced recursive feature elimination," in Sixth International Conference on Machine Learning and Applications (ICMLA 2007).

Yu X, Guo S, Guo J, Huang X (2011) An extended support vector machine forecasting framework for customer churn in e-commerce. Expert Syst Appl 38:1425–1430

Zineldin, M. (2000) TRM: Total Relationship Management, Studentlitteratur