Using different machine learning algorithms & sampling techniques to compare the prediction results of telecom industry churn.

A study submitted in partial fulfilment
of the requirements for the degree of
MSC Data Science

at

THE UNIVERSITY OF SHEFFIELD

by

Yash Prayag Lele
Reg no: 200226686

Word-length: **10184** August 2021

Acknowledgements

Firstly, I am grateful to have Dr. Monica Paramita as my supervisor on this dissertation. She was really helpful during the entire study process. She gave me excellent advise and guidance, and patiently answered all of my concerns. Regular supervisory meeting helped me feel confident about the thesis. She always provided valuable advice and was very patient with my queries throughout the research phase. I would like express my sincere gratitude to my supervisor. And it was an honour to have her supervise me on my thesis.

Secondly, it was my family and friends that helped me throughout the research phase with emotional support. And constantly encouraged me to have a positive attitude. It was a tough time this year with global pandemic and these people were the ones providing strong support throughout.

Abstract

Background: Because the telecoms industry is always battling its competitors for customer retention, it has become a hotspot of machine learning and data mining research. Customers' churn patterns must be regularly monitored and properly tracked, which necessitates the deployment of a systematic churn prediction model. Because customer turnover has such a direct influence on income, companies are striving to develop techniques to predict it. This is especially true in the telecom business.

Aims: The study aims to compare different machine learning algorithms in developing a predictive model for telecom industry. Also, to observe how differently sampled datasets affect the overall prediction accuracy achieved by these algorithms. Finally, to find which variables affect the churning of customers from the available dataset.

Methods: Knime Data Analytics Platform is used to implement different machine learning algorithms. Decision tree, Random forest, Naïve bayes, Logistic Regression, KNN, XGBoost tree and MLP are the algorithms used for this study. The dataset used is further randomly under sampled and Oversampled using SMOTE. The dataset achieving best results are further boosted using AdaBoost.SAMME method with upto 100 iterations.

Results: The dataset which was oversampled using the SMOTE technique produced the best possible prediction results with Random Forest algorithm. And after further boosting the SMOTE data the same algorithm out performed other machine learning algorithms used in this study.

Conclusion: Using SMOTE technique for balancing the data resulted in much higher accuracy results in predicting the churning customers. Without implementing boosting methods Random Forest with SMOTE oversampled data correctly predicted 90.7% (recall) of actual churning customers. Whereas after booting it improved to 93.1% (recall). On the other hand, variables like “Tenure”, “Contract”, “Internet Service”, and “Payment Method” were major factors contributing to the churning of a customer in telecom industry.

Table of Contents

Table of Figures and Tables

SECTION I.

1. INTRODUCTION

The telecommunications sector has undergone various reforms in recent years, including market liberalisation, increased competition, new networks, and new technology (Joolfoo, M. B., et.al., 2020). With some telecom firms introducing 4G and 5G networks in recent times have attracted many consumers to their firms making them churn from their previous network firm. In today's culture, the consumer or a customer is rapidly becoming the centre of any company's concern (Jain, H., et.al., 2021). Hence, in order to retain these valued customers, telecom firms try to provide satisfactory service and maintain a better relationship with the consumers.

In recent years, several businesses and industries have planned to change their focus from product to customer, potentially as a result of increasing consumer knowledge and leverage over other buyers (Oh, Y., et al., 2018). Customer churn is a process of a client moving from one firm to a rival firm. Customer churn prediction is a forecast modelling that has been extensively researched across a variety of industries, including financial institutions, social networking services, airlines, video gaming, finance, and also in the telecommunication industry (Ahn, J., et. al., 2020). In the Telecommunications industry, consumer churn refers to the loss of valued consumers or subscriber to rivals companies (Xu, T., et.al., 2021). As stated by SAS institute (2000), the telecommunications industry experiences a churn rate of 25 percent to 30 percent each year. And this churn rate will continue to rise in parallel with business growth in this sector. Because retaining valuable customers costs five to seven times less than gaining new ones, customer churn prediction is becoming a popular practise among the telecommunication firms. If customer turnover can be reduced to 5-10%, the company's growth rate can be increased up to 30-85% (Xia, G. E., et.al., 2016). These facts converged on the importance of customer churn prediction in the corporate decision-making and forecasting processes of telecom firms, which is also the primary goal of Customer Relationship Management (CRM) (Arivazhagan, B., et. al., 2020). Data is usually stored in CRM database systems so that it can be

turned into useful information in order to address the growing problem of customer churn and recognise churn behaviour before consumers are lost, thus increasing customer loyalty.

The key motivation behind developing a predicting algorithm is the pressing need for businesses to retain current clients, as acquiring new customers is much more expensive. Predicting customer churn in specifically telecom industry is the next big thing as many new firms are coming up with their own network and exciting offers to lure customers to their firms. As a result, in order to address this problem, we need to identify churners before they churn, so creating a model that forecasts potential churners appears to be critical. This model must be able to identify consumers who are likely to churn in the near future.

Scholars have proposed a variety of ways for measuring and addressing customer churn in the literature, including both quantitative and qualitative approaches. Quantitative methods include implementing machine learning models to predict the churning customers. One such method is stacking based data mining. In which different algorithms are implemented together on a dataset in order to enhance the accuracy results (Shabankareh, M. J., et. al., 2021). Apart from using Machine Learning algorithms which are traditional statistical models to predict churning customers, social network analysis which is a qualitative approach that is also emerged as a successful method as it takes into consideration different customer behaviours over the period of time and analyse their loyalty towards the firm (Devriendt, F., et. al., 2021). This plays an important role in customer relationship management to give firms the knowledge of their customers. Newly proposed systems like Uplift Modelling by Devriendt et. al. (2021), is a technique of targeting the optimal proportion of the identified probable churning customers those can actually be persuaded into staying with the same firm. This helps in saving the resources which are put into campaigning and advertising to retain customers who will churn regardless of the situation.

The following is a breakdown of the structure of the paper: The first section of the paper explains the notion of how churning affects a firm's business and Churn Prediction Model, followed by a survey of the literature on Churn Prediction Models and data mining techniques in the second section. The working methodology is described in the third section.

The fourth section covers the experimental data analysis along with results and discussion, as well as the conclusion and future research.

Aims:

The aim of this study is to:

There are three main objectives researched for the purpose of this study,

1) Apply Machine Learning algorithms: Decision tree (DT), Random Forest (RF), Logistic Regression, KNN, XGBoost tree, and Artificial Neural Network: Multilayer Perceptron (MLP) method to predict the churning customers from the dataset.

2) Comparing and evaluating the machine learning algorithms used for the purpose of this paper.

3) Apply different sampling technique to balance the dataset and then implement the machine learning algorithms to compare their results.

4) Evaluating the role of different customer behaviours on the churn rate.

Research Questions:

1) How can different Machine Learning algorithms be used to predict the churning customers in order to reduce the lose that will be incurred by the telecom firms.

2) Which algorithm produces the best possible prediction for customer churn?

3) Determine which sampling technique produces best possible prediction results in predicting churning customers.

4) How Customer Relationship Management (CRM) helps in retaining customer or avoiding customer churn?

SECTION II.

2. Literature Review

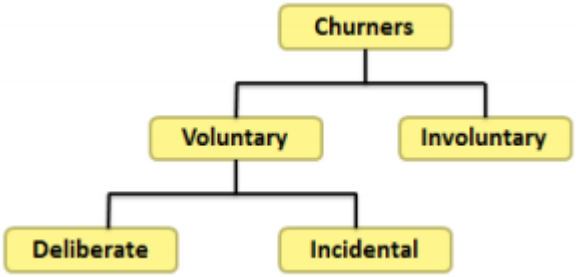
This section provides a review of the literature available on Churn, Customer Churn prediction methods, various approaches adopted to solve the problem and evaluation metrics used for evaluating the models.

2.1 Customer Churn

Customer churn is the process of customers switching from one service provider to another (Zhang, et. al., 2012). Customer churn prediction is essential for all businesses since it helps you acquire a better understanding of your customers and estimate future revenue. It may also help an organization discover and improve areas where customer service is lacking. In many regions, particularly industrialized ones, the market has reached saturation, requiring each new client to be lured away from competitors (Richter, Y., et. al., 2010). There has been a lot of work put into this, and there is still a lot of consumer data from the sector to look at. In today's world, the telecommunications sector is suffering from a significant income shortfall as a result of fierce market rivalry (Umayaparvathi, et. al., 2012). Data mining techniques have been widely utilised to anticipate customer churn by identifying the elements that are most likely leading to losing customers, allowing telecoms service providers to take prompt action to prevent churning. The most extensively used, Supervised data mining approaches for customer churn prediction are among the various data mining techniques available. When developing models that can learn from labelled training data, supervised data mining approaches are acceptable. Various supervised ML algorithms such as linear regression, neural networks, decision trees, k-nearest neighbours, genetic algorithms, Naïve Bayes, support vector machines (SVM) and many more can be implemented for developing such prediction models. When a possible churning customer is discovered, the customer relationship management (CRM) department generally contacts the customer and, if the client is determined to be a churn risk, takes the necessary

steps to retain them. According to Naz, N. A., et. al.,(2018), Churn can be categorised into two categories, namely; Voluntary and Involuntary. As stated by (Shaaban et. al., 2012), Voluntary churners are the ones which choose to switch due to unsatisfactory service or better offers from other service providers. And involuntary churners are the ones which have been discontinued from the service by the company itself due to inactivity or payment dues. Voluntary churners are difficult to be identified due to their unpredictable behaviour as compared to involuntary churners.

Figure 1: Types of Churners



2.1.1 Causes of Churn

Client turnover has become a fairly frequent problem in the telecom industry, therefore there must be a plethora of known and undiscovered reasons for customer churn. As stated by Jain, H., et. al. (2021), Some of the factors contributing towards customer churn are as follows;

- **Customer Service and Network Quality:** This is the most observed problem faced by the customers which causes them to move over to other service providers. Failing to connect to the customer support or not getting satisfactory solution to the problems leads to dissatisfactory experience and hence these customers prefer to leave the service provider. On the other hand, the network quality experience plays a vital role in the churning of a client.

- **High Rates :** Generally short-term contracts are much expensive to the long-term ones. Hence customers think of it as an expensive service overall and hence tend to churn. In contrast, if other telecom firms are providing the services at a lower price then customers are likely to churn.
- **New Service providers:** In order to promote their service, newer firms generally provide incentives to their clients, which attracts customers from other service providers.

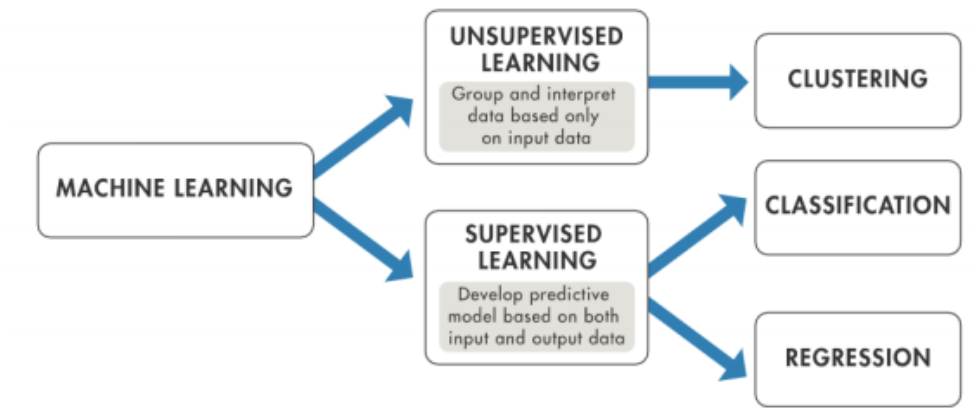
2.2 Predictive Analytics

Many large companies have been collecting consumer data for a long time, resulting in massive amounts of data in their databases. Predictive analytics is a technique used by these big firms in which they gather information and insights from large data sets that they have created over time in order to generate assumptions and forecasts about future occurrences (Larose et. al., 2015). Firms began leveraging this data to enhance estimations and efficiency, as well as forecasting choices, in order to put it to good use (Abbott, 2014). Predictive analytics was the suitable approach to manage enormous amounts of data and generate meaningful insights from it. Predictive analytics was built on a foundation of recognizing patters in the datasets used, quantitative analyses, machine learning, artificial intelligence, and data processing (Abbott, 2014).

2.3 Machine Learning

Machine Learning is a type of data analysis that aids in the creation of analytical models. It's a subset of AI (Artificial Intelligence). Machine learning algorithms learn from data, detect broad patterns, and make decisions with little or no human interaction. When we have a difficult problem or job that requires a large quantity of data, machine learning is implemented. It's an excellent choice for more complicated data since it produces faster and more accurate results. It assists a company in discovering profitable possibilities as well as any unexpected hazards (Sayed, H., et. al., 2018).

Figure 2: Machine Learning Techniques



Source: Google Images

2.3.1 Supervised Machine Learning Algorithms

There are two types of predictive analytics algorithms: supervised learning techniques and unsupervised learning methods. However, because this is a classification problem, only supervised learning approaches will be employed in this study.

According to Fabris, et. al., (2017), Supervised Machine Learning is the computational job of learning correlations between variables in a training dataset and then using this information to create a prediction model capable of inferring labels for incoming data. When examples are provided with known labels, the learning is referred to as supervised learning. The characteristics might be continuous, categorical, or binary in nature (Kotsiantis, S., et. al., 2006). Supervised learning models try to predict a target variable, which is represented by a single column in the dataset, using the other variables or columns in the dataset as an input data (Singh et. al., 2016). Supervised Learning is also known as Predictive Modelling. Supervised learning methods used in this study are Decision Tree, Random forest, Neural Network, Naïve Bayes, Logistic Regression, XGBoosted Tree.

2.4 Machine Learning Models

2.4.1 Decision Tree and Random Forest

The Decision Tree (DT) model creates a tree-like structure that reflects a series of decisions. The Random Forest method is comparable to the Decision Tree algorithm, however it offers several distinct benefits over the Decision Tree technique. Because decision trees are not affected by feature size and can accept both quantitative and qualitative qualities, they would be a good fit for supervised machine learning with the data we have (Jha, 2017). Random forests, on the other hand, build decision trees from randomly selected data columns, get predictions from each tree, and vote with more accuracy on the optimal option. Also, Random Forest algorithm is unaffected by nonlinear features and performs well with missing data and outliers.

Random Forest algorithms were used for groundwater potential mapping as one of the predictive machine learning algorithms (Naghibi, et. al., 2017).

Whereas, Decision Trees are widely used in finance for choice pricing, and banks use them to identify loan applicants based on the likelihood of default (Jha, 2017).

2.4.2 Naïve Bayes

The Bayes algorithm calculates the likelihood of an event occurring based on prior knowledge of the variables involved. A classification approach based on Bayes' theorem is known as naive Bayesian (NB). The Bayes theorem combined with strong (naive) independence assumptions are used in the Naive Bayes classifier, which is a basic probabilistic classifier. In terms of the output (class) vector, this approach requires that input variables are independent of one another (Nettleton, 2014). A big training set is not required for this algorithm. Because of its independent assumption, it nevertheless performs well with tiny data sets. As the dataset used in this study is about 7000 records, this algorithm may have an advantage over other algorithms.

According to Lowd, et. al. (2007), As Naive Bayes is a fast-learning classifier algorithm it is generally used in making real-time forecasts. This algorithm can estimate the likelihood of several target variable groups here i.e., it is a multi-class classifier algorithm.

"To create a recommendation system, the naive bayes and collaborative filtering work together which helps to evaluate unknown data and determine if a user prefers a specific feature using the ML and data processing techniques." (Ray, 2017).

2.4.3 K- Nearest Neighbour (KNN)

K-Nearest Neighbour is a machine learning method that is based on instances. It is also known as memory based learning algorithm (Sabbeh, S. F., 2018). The instances formed by the KNN algorithm on the basis of the training data are stored in the feature spaces which are then used to class each instance based on the majority votes of its neighbours (Deng, Z., et. al., 2016). The fact that KNN can be utilised for multi-class prediction problems is a benefit of utilising it for predictive modelling.

2.4.4 Logistic Regression

One can forecast the chance of a churn, or a client cancelling their subscription, using logistic regression. A supervised learning classification technique is logistic regression. A threshold in logistic regression is established, and only logistic regression is utilised for classification based on the limit (Jain, H., et. al., 2020). This machine learning model is being used in many types of analysis to develop a predictive model for areas like medicine, financial sector and many more. To define Logistic regression in a mathematical sense, Let the conditional probability P(z)=p be based on the probability of an observation relative to an event, then the logistic regression model can be expressed as:

$$p(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

As Logistic Regression is considered to be one of the best classification algorithm for predicting binary classes, it will be an excellent option to implement this algorithm in this study as the churn class variable is binary variable with values “Yes” and “No”.

2.4.5 Artificial Neural Network

According to Abiodun, O. I., et.al. (2019), Similar to the working concept of biological neural network in a brain of a human, the artificial neural network algorithms are developed. These machine learning algorithms are very adaptive in learning patterns in the data and are fault tolerant. These neural networks can be either hardware-based in which neurons are represented by physical components, or software-based which are computer algorithms (Asthana, P., 2018). In this study one such software-based neural network known as Multi-layer Perceptron (MLP) algorithm is used to solve this supervised machine learning problem.

2.4.6 XGBoosted Tree

Extreme Gradient Boosting is abbreviated as XG Boost. XG Boost is a high-speed and high-performance implementation of gradient boosted decision trees (Pamina, J., et. al., 2019). It's a supervised and ensemble learning approach that integrates trees to get a more generalizable result. In general, a Machine Learning model boost classifier builds a large number of trees and averages the results for better prediction. While generating the tree structure, it has the ability to reduce time consumption by using the best memory resources, parallel execution, and handling missing values (Chen, T., et. al., 2016).

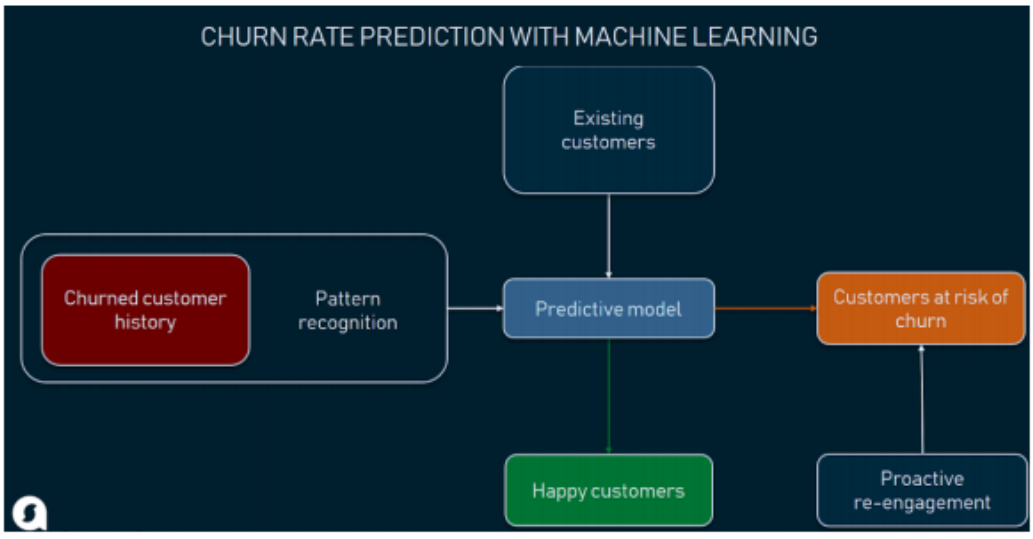
2.5 Churn Prediction

The goal of customer churn prediction is to forecast upcoming churners based on a predetermined time span and data associated with each network user. Customer churn prediction aids customer relationship management (CRM) in avoiding clients who are likely to leave in the future by recommending retention policies and better incentives or packages to lure probable churners to stay. The researched data based customer churn prediction can be split into two categories; customer informational data and customer behavioural data. This helps in analysts to form prediction model suitable for their end goal.

According to (Vafeiadis, et. al., 2015), Random forest, Decision tree and Naïve Bayes algorithms used for churn prediction which were among the top three ML models with best performance.

Customer churn prediction is a ever growing trend in the market analysis and strategic planning in today's world. This technique is being used in many field such as financial, medical, gaming industry (gambling) and also in telecom industry. Customer churn modelling focuses on identifying customers who are likely to depart and taking actions to avoid churn (Oyeniyi & Adeyemo, 2015). This technique is basically a proactive approach towards finding customers that are likely to churn and therefore manage those customers by providing offers and incentives.

Figure 3: Churn Prediction Flow Diagram (Source: Beker, 2019)



2.6 Customer Relationship Management (CRM)

CRM stands for customer relationship management, and it is a method or technique for understanding more about customers' wants and requirements in order to develop stronger bonds with them (Farquad, et. al., 2014). CRM is generally a process of continuously using improved information about existing and future customers to predict and respond to their requirements. This uses the behavioural customer data in order to predict their future moves.

According to (Buttle, 2009), CRM consists of three levels, on which it is based. These are:

Strategic CRM: The type of CRM is concerned with the development of a customer-centric company culture in which a competitive advantage is gained by making decisions on where to best spend the organization's resources.

Operational CRM: This type is concerned with the automation of customer processes, such as marketing, sales force, and service. “Operational CRM is mainly concerned with automating and simplifying workflow in the front office, including data collection, transaction processing, and workflow control in the sales, marketing, and services departments” (Zhang, et. al., 2008).

Analytical CRM: Customer data serves as the foundation for analytical CRM. Sales data (purchase history), financial data (payment history and credit score), marketing data (campaign response, loyalty scheme data), and service data can all be found in enterprise-wide repositories. This data is then used to gain insightful knowledge about the customer using various data mining techniques.

2.7 Data Mining

Data mining is a method of obtaining previously undiscovered, valid, and actionable patterns or information from huge databases for the purpose of supporting critical business decisions (Khalid, L. F., et. al., 2021). The new data mining techniques that are being created are proving to be useful for business owners in solving inquiries that were previously too time-consuming to manage. According to the (SAS Institute,2000), Data mining is "the process of choosing, examining, and modelling huge amounts of data in order to identify previously undiscovered data patterns for economic gain."

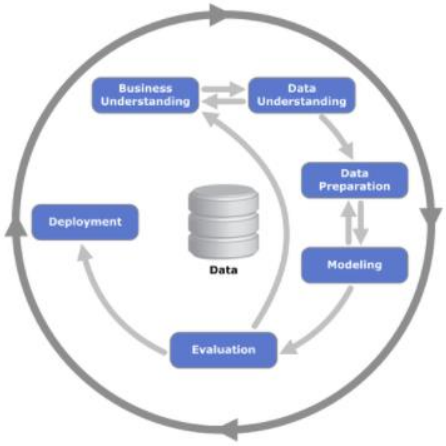
2.7.1 CRISP-DM

The CRISP-DM (Cross-Industry Standard Process for Data Mining) is a commonly used method for data mining projects. The CRISPDM method is an iterative methodology that divides a project's life cycle into six sections (Gončarovs, P., & Grabis, J., 2017). CRISP-DM is a non-proprietary, open-source standard approach for integrating data mining into a company or research unit's overall problem-solving strategy. This framework was developed jointly by Daimler Chrysler, SPSS and NCR in the 1996 (Karvana, K. G. M., et. al., 2019). According to Simsek Gursoy (2010), CRISPM DM technique consists of six stages;

- **Business Understanding:** The first stage of CRISP-DM is to get a thorough understanding of the company and to define the organization's unique needs or goals. Understanding a business entails determining the issues the company want to address. The main goal of this step is to get a clear idea of the problem and the expected outcome.
- **Data Understanding:** Clear understating of the data is the second most important aim of any problem at hand. As the data will be used as an input for further analysis, it is important to know what the data consists of and what part of the data is important to achieve the desired goal.

- **Data Preparation:** Exploratory analysis is one of the first steps towards successful development of a project. This step is very extensive in nature and can take up to 80% of the total time taken for a project to complete.
- **Modelling and Artificial Intelligence:** Different Techniques for modelling are chosen and implemented in this step. Because certain approaches, such as neural networks, have specific data format requirements, there may be a loop back to data preparation here.
- **Evaluation:** Once the modelling applied on the training data produce high quality results, the same model is then implemented on the test dataset and evaluated using different metrics considering all the desired goals are met.
- **Deployment:** This is basically the final step in development. When all the customers' desired results are achieved the model is then deployed as a final product to the client.

Figure 4: CRISP-DM Life Cycle



2.7.2 Data Mining Platform

For the purpose of this study, the Knime Data Analytics Platform is chosen as it is an open source data mining platform. As one of the main goals of this study was to enable developers with significantly less experience in programming to develop a predictive model using different machine algorithm, this platform seemed to be perfect. Knime is very intuitive and has constant developments in its data mining features. It provided various functions for all data mining steps and methods.

2.8 Related Work

Despite the fact that customer churn prediction is not a new study subject, experts in this field continue to utilise and analyse data mining approaches. The research community has developed and utilised a variety of approaches to tackle the customer churn prediction problem. Artificial Neural Networks, Decision Trees Learning, Regression Analysis, and Support Vector Machines are some of the most prominent approaches implemented. Furthermore, the feature section and how to deal with the problem of class imbalance have been carefully researched. Following works performed by various researchers is not all that have been conducted in this field.

Idris, A., et al. (2012) used the Random Forest and KNN machine learning algorithms to forecast churn in the telecom industry. Particle Swarm Optimization, as well as feature reduction techniques such as PCA, F-score, Fisher's ratio, and Minimum Redundancy Maximum Relevance, are utilised to manage the dataset's unbalanced distribution. Their research was based on genetic programming using AdaBoost method to further boost the algorithm. The Area Under the Curve (AUC), sensitivity, and specificity techniques were used to assess the modelling methods' performance. These simulations were shown to be successful in predicting churn in a favourable way. The researcher achieved an overall accuracy of 89%.

Faris, H. (2018) proposed a hybrid approach based on the over-sampling method that combines Particle Swarm Optimization (PSO) with a Random Weight Network to solve the churn problem in telecoms data. To balance the dataset used in his research, the ADASYN over sampling technique was used. Balancing the data proved to be useful as the result from this method was significantly better.

Huang, Y., et. al., (2015), In the big data platform, investigated the issue of client turnover. The researchers wanted to show that, depending on the volume, diversity, and velocity of the data, big data may substantially improve the process of forecasting churn. AUC was used to evaluate the Random Forest method.

The study done by Azeem et. al. (2017) proposed a fuzzy based churn prediction model and validated using original data from telecom sector in South Asia. Several predominant classifiers namely Linear Regression, Neural Network, SVM, c4.5, Gradient Boosting, Random Forest and AdaBoost have been compared with fuzzy classifiers to mention the fuzzy classifiers superiority in finding the exact churners set.

The study conducted by Ahmad et. al. (2019) research is to create a churn prediction model that will help telecom carriers identify consumers who are likely to churn. The model used in this work employs machine learning techniques on a big data platform to create a novel approach to feature selection and engineering. The standard metric of AUC (area under curve) is used to assess model performance, and the value of AUC obtained is 93.3 percent. Another significant addition of this research is the utilisation of a customer's social network in a prediction model by extracting characteristics from Social Network Analysis.

On the basis of historical data, Qureshi, S. A., et. al., (2013) discussed the use of under sampling and oversampling techniques for solving the class imbalance problem and identifying consumers who are likely to leave. Burez, J., & Van den Poel, D. (2009) investigated the problem of data imbalance in churn prediction models and compared the performance of Advanced Under-Sampling, Random Sampling, Weighted Random Forests, and the Gradient Boosting Model. The AUC and Lift evaluation metrics were used as model performance measures. The results of the examination indicated that the under-sampling strategy outperformed the other strategies tested.

Most of the research studies performed in the field of customer churn prediction in telecom industries were evaluated on the basis of overall accuracy obtained by the implemented machine learning algorithm or by observing the Area Under the Curve (AUC) performance evaluation metric. Also, the research carried out was done by using different platforms. Two of the most popular used platforms were Python and RStudio programming languages.

In this paper, the agenda of carrying out this study was to help researchers with significantly less programming experience to develop a predictive modelling tool and achieve satisfactory results. Also, the evaluation of the performance of the predictive models was based on Recall

and Precision values obtained for the churning class of the dataset. The reason behind this decision was to minimize the false prediction of classifying non-churning customers as churning. As recall value only states the accuracy percent of predicting churning customers among the actual churners, it was the best evaluation metric. Also, the decision of not considering the prediction results of non-churning customers was taken as it won't affect the firms business and also won't result in any loss.

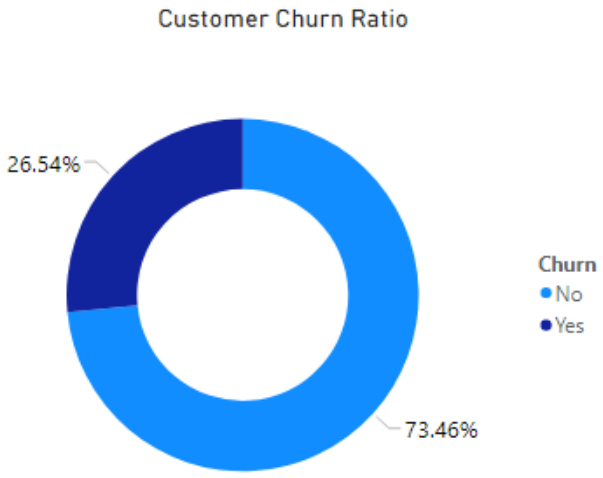
SECTION III.

3. Methodology

3.1 Data Overview

For the purpose of this study, the Customer Churn data is acquired from the IBM official website. As per the information provided on the website, this data is totally fictional and hence does not disclose any private or personal information of the users. This dataset contains information related to 7043 subscribers of which 5174 are non-churner and remaining 1869 are churners and 22 variables including 21 feature variables and 1 class variable.

Figure 5: Churn Statistics



This dataset includes information relating to many continuous and categorical variables as follows,

- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range (if senior citizen or not), and if they have partners and dependents
- Column describing churning of customers – Churn

Following is a brief description of each variable present in the dataset:

• customerID - Customer ID

• gender - Whether the customer is classified as a male or a female

• SeniorCitizen - Whether the customer is classified as a senior citizen or not (1, 0)

• Partner - Whether the customer has a partner or not (Yes, No)

• Dependents - Whether the customer has dependents or not (Yes, No); e.g (children)

• Tenure - The length of time a client has been with the firm. (in months)

• PhoneService - Whether the customer has subscribed to phone service or not (Yes, No)

• MultipleLines - Whether the customer has subscribed to multiple lines or not (Yes, No, No phone service)

• InternetService – Type of internet service opted by the customer (DSL, Fiber optic, No)

• OnlineSecurity - Whether the customer has subscribed to online security or not (Yes, No, No internet service)

• OnlineBackup - Whether the customer has subscribed to an online backup or not (Yes, No, No internet service)

• DeviceProtection - Whether the customer has subscribed to device protection or not (Yes, No, No internet service)

• TechSupport - Whether the customer has subscribed to tech support or not (Yes, No, No internet service)

• StreamingTV - Whether the customer has subscribed to streaming TV or not (Yes, No, No internet service)

• StreamingMovies - Whether the customer has subscribed to streaming movies or not (Yes, No, No internet service)

• Contract – The basis of the contract opted by the customer (Month-to-month, One year, Two years)

• PaperlessBilling - Whether the customer has opted for paperless billing or not (Yes, No)

• PaymentMethod – The method by which the customer prefers to pay for the subscription (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))

• MonthlyCharges - The monthly charges incurred by the customer as per the contract of service.

• TotalCharges - The total amount charged to the customer

• Churn - Whether the customer churned or not (Yes or No)

3.2 Ethics overview

The data used in this study has no risks based on the ethics stand point. IBM Watson Analytics Community gathered the data collection, which is now publicly utilised as an input data in machine learning algorithms for developing a churn prediction model for telecommunication industry. Ethical problems in research are one of the topics that researchers use to safeguard study participants' rights while creating research techniques and establishing a trusting connection with the themselves (Rana, J., et. al., 2021). This secondary data does not contain any sensitive data regarding politics, culture and religion. Personal information related to the telecom firm's customers is anonymized and are differentiated using unique customer ID. Other non-sensitive data such as demographic data is used for the study purpose to correlate it with the churning of the customers. Hence, the data used in this study is risk free and have no ethical issues involved.

3.3 Data Preparation

Various different tasks such as data cleaning which includes handling missing values and handling outliers, creating new informative variables from already existing one's, normalizing data, selecting features that are useful and removing the other variables are carried out as a preparation to convert the raw data into meaningful and useful data to further provide it as an input to various models.

3.3.1 Handling Missing Values

The first step in the life cycle of a model development is always to handle any missing values in the dataset. As having missing values in the dataset can cause the model to abruptly fail or give wrong output results, it is better to handle this missing data. If the number of missing data values are small in number, then those instances are generally removed from the dataset. But on the other hand, if the missing data is in large quantity, then statistical methods such as mean, median or mode can be applied to assume this data (Coussement, K., et. al., 2017). In the

case of smaller quantity of missing data, if a relation is found with other variables, then it can be handled differently based on the specific situation. Similar situation was handled during this study.

During the exploratory analysis phase of this study, it was found that the column "TotalCharges" had 11 missing values and this was the only variable in the dataset with missing data values. After some more analysis it was discovered that the "Tenure" corresponding to these missing values was "0". This meant that these customers were new to the firm and hence the charges accounted to nil. Therefore, to handle these missing values it was decided to input "0" as the values in place of the missing data. This process was carried using the "Missing Value" node of the Knime Data Analytics Platform.

3.3.2 Data Manipulation

During the data cleaning process, it was observed that the most categorical columns consisted of 3 different categories. Example: MultipleLines :- (Yes, No, No phone service). The categories "No" and "No phone service" seemed to be redundant in nature. Hence, columns with such redundant categorical values were handled by altering all the rows with such redundant categories with a common value "No". To implement this, "Column Expression" node is used. Further in the manipulation stage, the column "CustomerID" was removed from the telecom dataset as it was of no specific importance in predicting the churning customers. This step was carried out using the "Column Filter" node.

3.3.3 Feature Engineering

Some of the method used in Feature Engineering process are label encoding and one-hot encoding. The dataset used for the purpose of this study consists of many categorical variables. It is found that some machine learning algorithms do not work well with categorical variables as they accept only numerical values as an input data. Hence one of the feature engineering processes mentioned above needed to be implemented on the data. The main problem with label encoding is, as it assigns a unique value to each category in the categorical column, the machine learning algorithm may prioritize the categories based on the numbers assigned to them and hence give a biased result. Whereas, using one-hot encoding helps to divide a categorical column into many columns based on the total number of categories it consists.

Hence each of these newly created columns become a binary variable. Basically, One – to – Many operation in performed on each categorical variable in the dataset. This process was carried out using the “One to Many” node in Knime.

Another feature engineering process is carried out on the “tenure” data column. As this column consists of the number of months the customer has been with the telecom firm, it has a range of values from 0 to 72. This may lead the algorithm to produce biased results as it may prioritize based on the number. To eliminate this possibility, it was converted into a categorical variable called “Tenure”. The five categories created are ("< 12 months", "13-24 months", "25-36 months","37-48 months", “> 48 months"). The “Rule Engine” node is used to create this new variable. The former “tenure” variable is then discarded from the dataset.

3.3.4 Sampling Data

In many real-life circumstances, such as telecommunication churn data, the problem of class imbalance is a serious concern (Haixiang, G., et. al., 2017). When the occurrences of one class outweigh the occurrences of other classes, this issue arises. In a dataset, the majority class may be far more numerous than the minority class, while the minority class is relatively uncommon since it does not occur as frequently as the majority class (Hanif, A., & Azhar, N., 2017). The classification column of churn in the dataset used in this study has two possible values (Yes, No). But, based on the “Figure 5” of Data overview, it can be seen that 73.46% of the customers in the dataset are classified as non-churners and the remaining 26.54% as churners.

The most frequent approach for dealing with unbalanced classes is sampling, which involves changing the distribution of training samples (Burez, J., & Van den Poel, D., 2009). For the purpose of this research two sampling techniques were implemented; Synthetic Minority Over Sampling Technique (SMOTE) and Random Under Sampling.

- **SMOTE Technique:** This sampling approach enriches the data by oversampling it by adding fictional or synthetic rows instead of replacement or randomized sampling. Rather than using data space, this approach uses feature space to create fake samples (Chawla, N. V., 2009). In this study the "SMOTE" node was implemented to apply this oversampling technique on the telecom dataset. In this node the developer needs to configure the class column based on which the imbedded algorithm categorizes the majority and minority class from the input data. As the class column for this study is the churn column, it is set accordingly. To produce synthetic data samples of the minority class the value based on which they will be generated is set to "3" for the purpose of this study. Selecting the number of nearest neighbours is an option that controls how many of your closest neighbours will be taken into account to generate the synthetic data samples for the minority class. The algorithm takes one item from the target class ("No" in churn column), chooses one of its neighbours at random, and creates a new synthetic example along the line between the sample and the neighbour.

- **Random Under Sampling:** The Random Under Sampling method is the exact opposite of the SMOTE method. Essentially, instead of adding data from the minority class synthetically, this approach eliminates data samples from the majority class at random to balance the dataset on the basis of the selected target column. This technique has a complimentary advantage as it excels in shortening run times and lowering storage needs for huge datasets. But on the other hand, it has a major disadvantage as it discards data samples which may have a potential to be a contributing factor in the decision making process of the algorithms implemented. This may in turn reduce the overall accuracy result acquired by the predictive model developed. In this study an "Equal Size Sampling" node of Knime Platform is used to carry out the random under sampling technique on the telecom churn data.

3.3.5 Normalizing Data

Data normalisation is a pre-processing method that involves scaling or transforming data to ensure that each characteristic contributes equally. According to Singh, D., & Singh, B. (2020), to develop a generalised prediction model of the classification issue, machine learning algorithms rely on the quality of the data. From the dataset used for this study, there were two columns namely “MonthlyCharges” and “TotalCharges” which consisted of continuous data and needed to be normalized before applying it to the machine learning models. The “Normalizer” node is used to carry out this operation on the telecom dataset used in this study. The configuration used for normalizing the data is the “Min-Max” technique in which the data values are transformed into a range of 0 to 1. This process uses the following formula to compute the normalized value of the selected data values.

Where, A = Original Data Value.

[C,D]= [min, max] range to transform to.

B = value after normalizing

$$B = \left(\frac{(A - Min(A))}{Max(A) - Min(A)} \right) * (D - C) + C$$

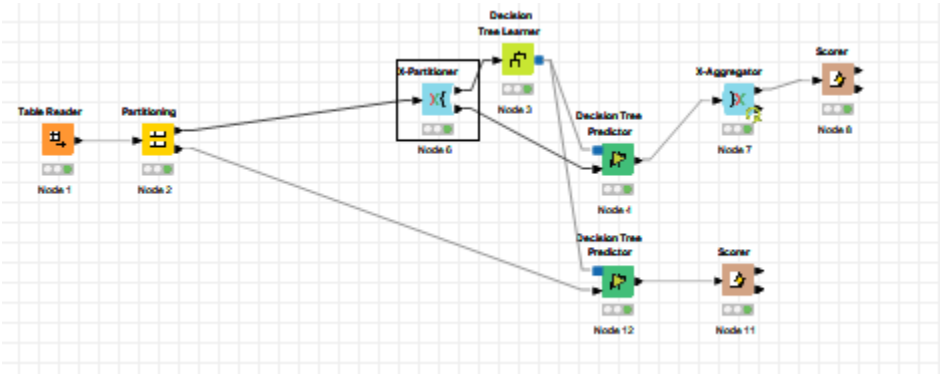
3.4 Data Modelling

Knime Data Analytics Platform is used to carry out the machine learning tasks required to develop a predictive model based on telecom industry dataset.To predict the probability of whether a customer would end their contract with the current telecom firm and go on to a competitor firm, we use a number of different machine learning algorithms to test and compare their performance. Machine Learning algorithms like Decision Tree (DT), Random Forest (RF), Naïve Bayes, Logistic Regression, XGBoosted tree, and Multi-Layer Perceptron (ANN).

The first step to developing a predictive model for this study is to split the data into training and test set. The dataset is divided into 80% training and 20% test data. To carry out this process a "Partitioner" node is used. This helps to train the machine learning algorithms and then use this model to further predict the churning possibility of the customer data in the test set.

The next step implemented in the process is to apply a 10-fold cross validation which helps on figuring how the model reacts on new data provided in each fold. This is done with the help of "X-aggregator" node. In this process the training data is further divided into 10 parts. This data is further provided to the learner nodes of the selected machine learning algorithms and then to the respective predictor node. The model developed on this training data is then applied on the test data through the second predictor node of the same algorithm.

Figure 6: Work Flow of Predictive model



This process is carried out throughout all the machine learning algorithms used in the development of the predictive model for telecom churn data.

The development of a predictive model for telecom churn data is divided into four parts so as to observe the differently sampled datasets into the selected machine learning models and to compare the achieved results in predicting the possible churning customers. The three datasets used in this study are the i) The Original Telecom Dataset, ii) SMOTE sampled Dataset, iii) Random Under Sampled Dataset, and iv) Boosted Dataset. For the fourth dataset, whichever dataset from the other three types produces the best results is then applied to a boosted iterative machine learning process to check if the prediction results can be further increased.

This is to check which sampling technique implemented on the dataset or the original data produces the best possible prediction results for the churning customers.

To model the boosted algorithm four different boosting nodes are used. “Boosting Learner Loop start” and “Boosting Learner Loop End” nodes are implemented in which the model learns from the training data during the 100 iterations and then the probabilities are weighted based on the classification errors. This algorithm uses the AdaBoost.SAMME method to further boost the machine learning algorithms applied. The advantage of AdaBoost.SAMME is that it can also handle multi-class problem. AdaBoost is an iterative method that works with certain weak classifiers to solve classification issues by constructing a strong classifier. AdaBoost boosts the performance of machine learning classifiers by several orders of magnitude (Saghir, M., et.al., 2019). AdaBoost is a machine learning algorithm that operates by assigning weights to each occurrence in the training dataset.

Further in the process, the model developed using the training dataset is applied to the test set with the respective predictor node. Following figures shows the work flow of the “Boosting Learner” and “Boosting Predictor” meta nodes.

Each machine learning model is evaluated using the “Scorer” node which gives detailed information about the various results achieved by the predictive model. The scorer node provides information like the Overall accuracy, Recall, precision and F-measure values for each class involved in this classification problem.

Figure 7:Work Flow for Boosted Algorithm

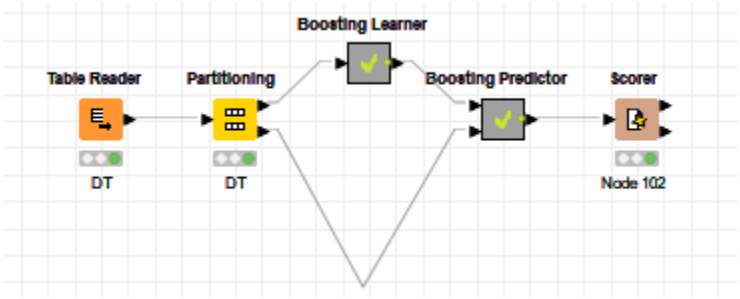


Figure 8: Inside Boosting Learner node

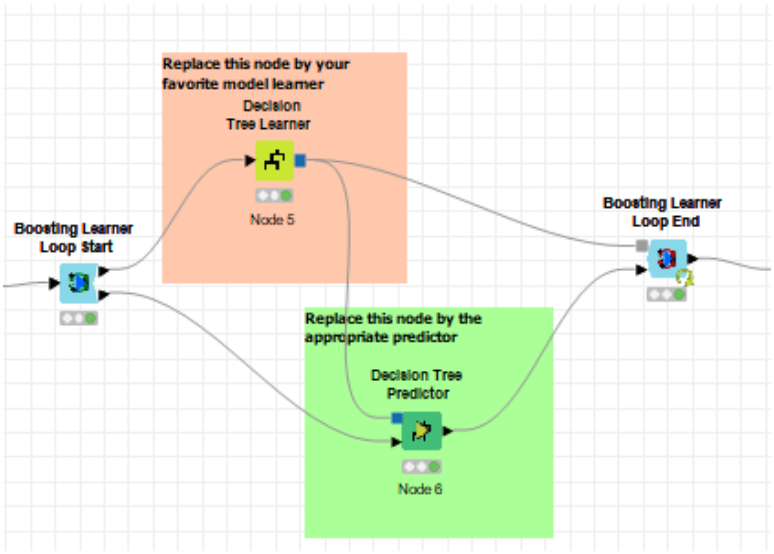
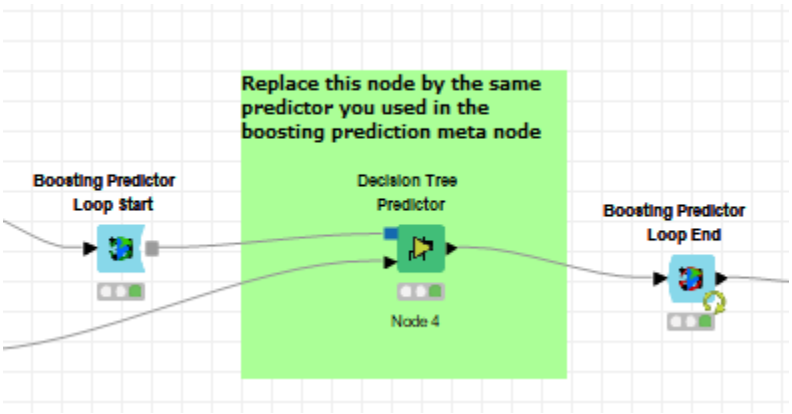


Figure 9: Inside Boosting Predictor node



3.4.1 Configuration Setup

3.4.1.1 Decision Tree

The configuration setup done for the decision tree algorithm is such that it maximizes prediction accuracy and avoid the problem of overfitting. This is implemented by selecting the Minimal Description Length (MDL) in the learner node. The quality measure by which the splits in the trees are made is set to Gini Index. The class column here is set to “churn” data column as this is the main target to be predicted.

3.4.1.2 Random Forest

Knime Data Analytics Platform does not allow great variation in configuration setup for this machine learning algorithm. The two useful configuration setups that can be implemented are the feature selection and type of split criterion based on which the model learns. The information gain split criterion is implemented in this machine learning model as it usually is a good choice for classification and regression machine learning models (Breiman, L., 2001). The target column is set to “churn” column.

3.4.1.3 Naïve Bayes

The configuration setup for this machine learning model is set to the default values provided by the Knime platform. The disadvantage of this algorithm is that it is one of the least configurable algorithmic model that Knime platform provides. The classification column is set to “churn” column.

3.4.1.4 Logistic Regression

The classification column is set to “churn” as this is the target column to be predicted. Further the reference column here is set to “No” as in people not churning. As this is a binary classification problem it takes into consideration 1-P(No) to calculate the probability. As the main focus for this study is to calculate the churning customers, we select the “No” class as reference to calculate the probabilities of the “Yes” class. The method used to carry the logistic regression analysis; the iterative reweighted least squares is implemented. This method is used to minimize the effect of the outliers in the dataset.

3.4.1.5 K-Nearest Neighbours

This is another algorithm which provides very less configurable setting on the Knime platform. The “churn” column is set as the column with class labels. For the K-Nearest Neighbour machine learning algorithm the value of “k” has to be set up as it used that value to classify new instances based on that number of neighbours around each data sample. Here it was set to k=3.

3.4.1.6 XGBoosted Tree

The “churn” column is selected as the target column for this machine learning model. The boosting rounds were configured to 200 which is a default number set for the algorithm. The configuration for objective setting is set to “softprob” as it produced the best results. The feature selector and ordering method was select as cyclic method as it selects the features one at a time in a deterministic way.

3.4.1.7 Multi-Layer Perceptron (MLP)

Similar to all other algorithms the class column was set to “churn” column. There is setting which allows developer to configure the number of iterations the MLP should loop and learn from the training dataset. The number of hidden layers is set to 1 and the number of hidden neurons per layer is set to 10.

3.5 Evaluation Methods

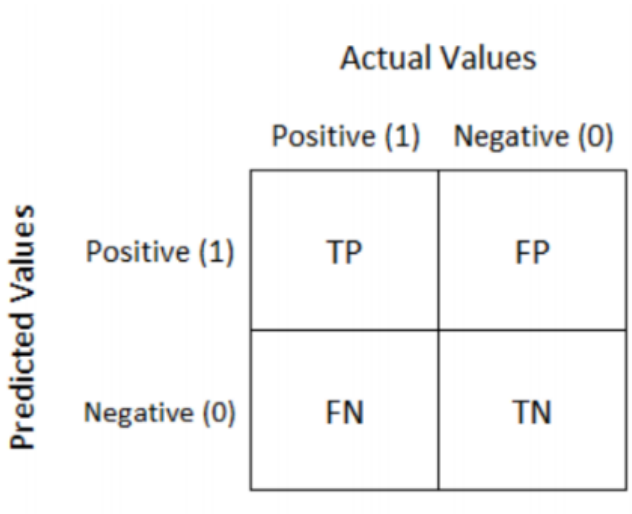
To assess the efficacy of classifiers in predicting churn for various machine learning algorithms using the proper parameters is an important step in developing a robust predictive model. For the purpose of evaluating the quality of prediction in this study, four metrics were utilised. Precision, Recall, F-measure score, and overall accuracy are the four metrics. To calculate these evaluation metrics, it is important for the predictive algorithms to first calculate the four parameters required to further apply them in different performance evaluation formulas. These four parameters are True Positives, True Negatives, False Positives, and False negatives which are represented as (TP), (TN), (FP), and (FN) respectively in the confusion matrix.

In this study, as the main goal is to successfully predict the churning customers from the telecom dataset, it is important to evaluate the prediction results achieved by the various machine learning algorithms related to the "Yes" class in the "churn" data variable which is related to the churning customers. Hence, evaluating the prediction results solely based on overall accuracy is not enough. Focusing on the results achieved for the churning customers will enable the telecom firms to maximize customer retention and hence minimize loss incurred. To evaluate the results for churning customers the evaluation metrics like Precision, Recall and F-measure are used, which gives a correct representation of the results achieved for predicting churning customers. Precision and Recall values are the metrics that actually determine how accurate a churn prediction model is performing (Azeem, M., et. al., 2017).

Various standard performance criteria have been devised to analyse the efficiency of different classifiers for churn prediction. These metrics may be used to evaluate the performance of any model created using both balanced and unbalanced data (Awoyele, T., 2020).

- **Confusion Matrix:** In a table with two rows and two columns, the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), is displayed. It provides the information needed to assess the accuracy of churning and non-churning class. In this study, the confusion matrix is obtained by implementing the "Scorer" node at the end of the machine learning model.

Figure 10: Confusion Matrix



Where,

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

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

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

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

- **Precision score:** Precision score is the factor that determines what percentage of the positively predicted cases are actually positive. Which means that this measures the percentage of the customers which were classified as churners whether actually churning or not against the customers that actually churned. Precision score refers to the specificity of the algorithm that has declared customers as positive churners out of actual positive churners. The formula used to calculate Precision is as follows:

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP)\ +\ False\ Positive(FP)}$$

- **Recall score:** Recall score is a factor that determines the ability of the machine learning model to identify or predict the positive cases out of the actual positive cases. Which means that it measures the accuracy percentage of algorithms to identify or predict churning customers which are actually churning. Recall score refers to the sensitivity of the machine learning algorithms to successfully predicting the churning customers out the actual churning customers. The formula used to calculate Recall is as follows:

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP)\ +\ False\ Negative(FN)}:$$

- **F- measure score:** Precision and recall alone are not sufficient to characterise a predictive model's efficiency as excellent performance alone does not always imply high performance. As a result, F-measure, a popular combination of Precision score and Recall score, is frequently employed as a single metric for machine learning classifier performance evaluation. The harmonic mean of precision and recall is defined as F-measure score.

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}.$$

- **Overall Accuracy Score:** The basic purpose of this method is to calculate the accuracy percentage of the predicted values compared to the actual class column.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

SECTION IV.

4 Results and Discussion

4.1 Introduction

This chapter analyses and discusses upon the achieved results from the ML algorithms applied to the telecom dataset with comparison between different sampling techniques that are applied on the data. In addition, this chapter will also provide insights on how customer loyalty and other behavioural patterns proves to be an important factor in the churn prediction through data visualisation.

4.2 Results

This section will provide the results achieved by implementing different Machine Learning algorithms on the original, oversampled and under-sampled dataset. As the main aim of this study is to predict the churning customers, the precision, recall and f-measures scores of those customers are considered to evaluate the performance. Taking into consideration the imbalanced dataset, evaluating the performance based only on accuracy would result in a biased predictor model. Predicting the non-churning customers correctly is of not great importance as this will not result in any losses incurred by the firm. Hence, the recall values for the churning customers classed as “Yes” will be evaluated. In addition, the dataset producing the best results will be further boosted using the AdaBoost.SAMME boosting node from the Knime Data Analytics platform to achieve higher accuracy in forecasting churning customers.

4.2.1 Experiment 1: Original Dataset

ML Algorithm	Overall Accuracy (%)	Precision (Yes class)	Recall (Yes class)	F-Measure (Yes class)
Decision Tree (DT)	78.9	0.597	0.522	0.557
Random Forest (RF)	79.5	0.652	0.487	0.557
Naïve Bayes	76.8	0.543	0.797	0.646
Logistic Reg	81.5	0.699	0.535	0.606
XGBoost	80.3	0.678	0.489	0.568
KNN	76.6	0.565	0.52	0.542
MLP	79.9	0.661	0.5	0.569

Table 1: Accuracy Stats- Original Dataset

From Table 1, We can conclude that, in terms of accuracy (81.5%) and precision (69.9%), Logistic Regression outperformed all other algorithms, but the recall value, which indicates the real proportion of accurately recognised churning customers, is comparatively low. If we consider all three metrics then Naïve Bayes turns out to be the best performing model among others on the imbalanced dataset with an accuracy of 79.7% (Recall Value) of correctly identifying churning customers from those that are actually churning, and overall accuracy of 76.8%. Other algorithms performance was as follows DT (78.9%), RF (79.5%), XGB (80.3%), KNN (76.6), MLP (79.9%).

4.2.2 Experiment 2: SMOTE Oversampled Dataset

ML Algorithm	Overall Accuracy (%)	Precision (Yes class)	Recall (Yes class)	F-Measure (Yes class)
Decision Tree (DT)	77.3	0.772	0.792	0.782
Random Forest (RF)	84.7	0.809	0.907	0.856
Naïve Bayes	76.5	0.742	0.813	0.775
Logistic Reg	77.7	0.759	0.813	0.785
XGBoost	78.6	0.77	0.815	0.792
KNN	76.5	0.735	0.831	0.78
MLP	78.4	0.773	0.804	0.788

Table 2: Accuracy Stats - Oversampled Dataset

As the original telecom churn dataset was imbalanced based on the classification which is the predictor class column, the machine learning model developed on such data may give biased results against the minority class. Therefore, an oversampling technique is implemented on the original dataset. This technique is known as Synthetic Minority Oversampling Technique (SMOTE). This is a statistical method for evenly increasing the number of instances in your dataset. The module generates new instances based on current minority cases provide as input. The number of majority cases does not change as a result of SMOTE implementation. The main objective of this study is to properly forecast churning customers so that the firm can better manage and retain them. Hence, in addition to evaluating the overall accuracy, the precision and recall is also evaluated. Table 2 clearly states that the Random Forest (RF) Machine Learning model outperforms every other algorithm used in this study by a significantly large margin. It achieves an accuracy of 84.7%. Also, the accuracy of correctly predicted churning customers from the actually churned ones is 90.7% i.e. the recall value. Making the dataset balanced using the SMOTE oversampling method proves to be an effective way to achieve the desired results. Apart from Random Forest the nearest competitors were the KNN, XGBoost, Naïve Bayes and Logistic Regression in correctly predicting the churning customers with 83.1%, 81.5%, 81.3% and 81.3% based on their respective recall values. With some improvement in the recall value compared to those achieved on original dataset for Decision Tree and MLP algorithm it performed less accurately in terms of overall accuracy.

4.2.3 Experiment 3: Under-sampled Dataset

ML Algorithm	Overall Accuracy (%)	Precision (Yes class)	Recall (Yes class)	F-Measure (Yes class)
Decision Tree	77.5	0.796	0.772	0.784
Random Forest	76.3	0.774	0.743	0.741
Naïve Bayes	75.5	0.736	0.797	0.765
Logistic Reg	74.6	0.734	0.773	0.753
XGBoost	74.2	0.733	0.762	0.747
KNN	70.3	0.682	0.759	0.719
MLP	73	0.76	0.789	0.745

Table 3: Accuracy Stats - Under sampled Dataset

Similar to the technique used for oversampling the dataset where the minority class data was synthesized to create an equal and balanced data, here the majority class data was randomly reduced or downsized to create a balanced data. The highest overall accuracy achieved was by the Decision Tree model (77.5%) and the highest recall value was recorded by Naïve Bayes (79.7%) with overall accuracy of (75.5%). Hence, Naïve Bayes was better in predicting the churning customers. But compared to the results achieved using oversampled data, this experiment performed poorly in terms of overall accuracy as well as in successfully predicting the churning customers. But this does not mean that this technique isn't useful in prediction models. As under-sampling randomly downsizes the majority class the results may vary. The techniques to achieve higher accuracy varies from problem to problem.

In this case, the oversampling technique worked the best and hence, the oversampled dataset will be used to implement further boosting techniques to achieve higher accuracy in predicting churning customers.

4.2.4 Experiment 4: Boosting Oversampled Dataset

ML Algorithm	Overall Accuracy (%)	Precision (Yes class)	Recall (Yes class)	F-Measure (Yes class)
Decision Tree	83	0.815	0.855	0.835
Random Forest	87.4	0.839	0.931	0.88
Naïve Bayes	77.1	0.755	0.805	0.779
Logistic Reg	77.1	0.749	0.814	0.78
XGBoost	82.7	0.803	0.868	0.834
MLP	84.6	0.812	0.9	0.854

Table 4: Accuracy stats - Boosted Oversampled Dataset

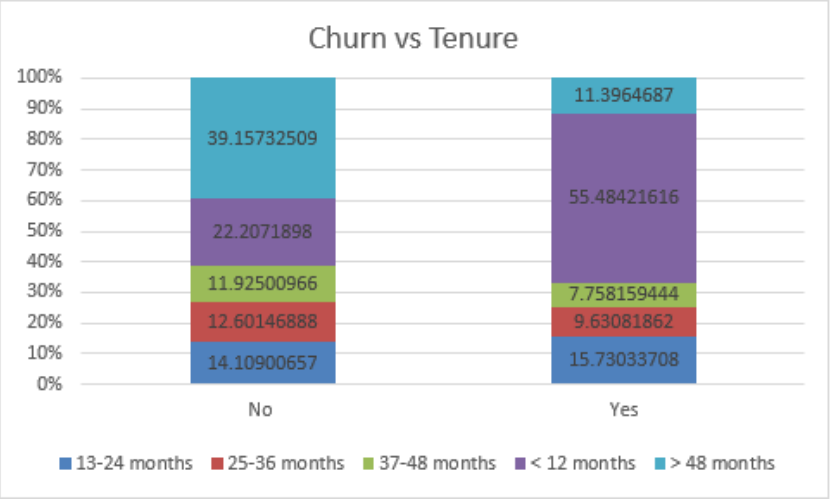
As data which was oversampled using the SMOTE technique performed the best among the three types of datasets, it was further used to boost the same Machine Learning algorithms applied before. The Boosting Learning and Predictor node developed by the Knime Data Analytics Platform is used to iterate the machine learning models up to 100 times to learn and reduce the error rates of the predictions. The AdaBoost.SAMME boosting trains the model

based on classification error and weighs them accordingly. The learning continues to go on till the number of iterations are completed. This helps in boosting the overall accuracy of the prediction model. Table 4 clearly states the increased accuracy as compared to all other experiments. Also, it has a significant increase in results from the non-boosted oversampled dataset experiment. Major improvements are recorded in most of the machine learning models used in this study. Random Forest was the one which performed the best and gave the highest overall accuracy of 87% among all the experiments carried out in this study. Also, it recorded the highest accuracy of correctly predicting the churning customers at 93.1% (recall value). The only limitation occurred during the experiment of booting the algorithms was for the KNN model. Due to knime software's limitations the KNN machine learning algorithm was not able to be boosted and hence was omitted from this experiment.

4.2.5 Churn vs Tenure

From the figure 11, it can be concluded that the clients among lower end of the tenure category have the highest churn rate among all the categories combined. It consists of more than 50% of the total number of churners in the dataset. This suggests that the customer relationship management needs to improve the services and incentives or offers provided to the them to gain loyalty from the new customers joining their firms. This will in turn increase the customer lifetime value and the firm will profit by retaining these outgoing customers.

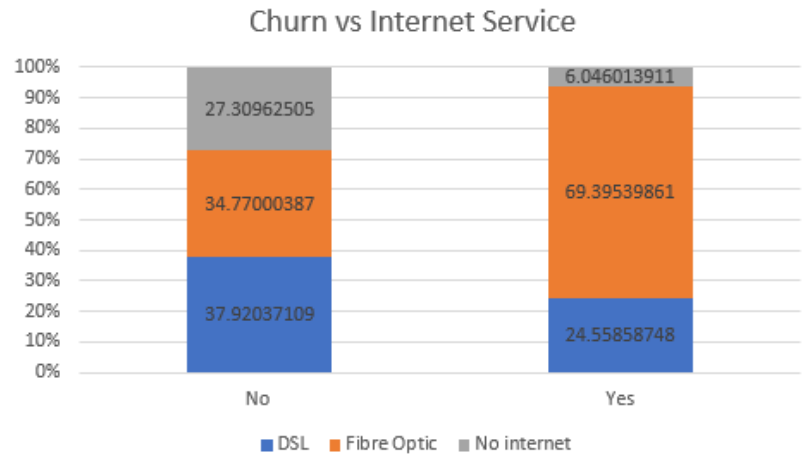
Figure 11: Churn vs Tenure



4.2.6 Churn vs Internet Service

The finding from the figure 12 were surprising and not expected. As fibre optic internet service is the fastest internet service available in today's world the people using this service must be the most satisfied customers. But from this graph it can be concluded that the customers subscribing to use the fibre optic service are most likely to be churned. This suggests that the internet service provided by the firm through fibre optic cable is not satisfactory or the prices for it are higher than its competitors. Analysing such data helps the telecom firm to improve their services or lower their service cost to retain their outgoing customers.

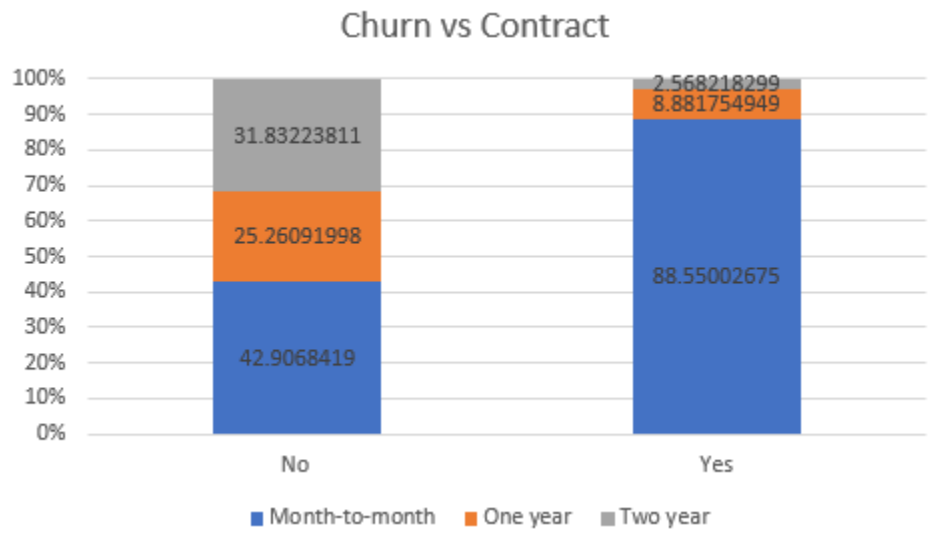
Figure 12: Churn vs Internet Service



4.2.7 Churn vs Contract

The contract type was another variable that showed a substantial difference between churning and non-churning consumers. Consumers on a 'month-to-month' contract are far more likely to abandon the firm than customers on lengthier contracts. The graph suggests that almost all the churners from the dataset were on the monthly contract. From this it can be concluded that the customer loyalty is the biggest factor in the churning of the customers in the telecom industry.

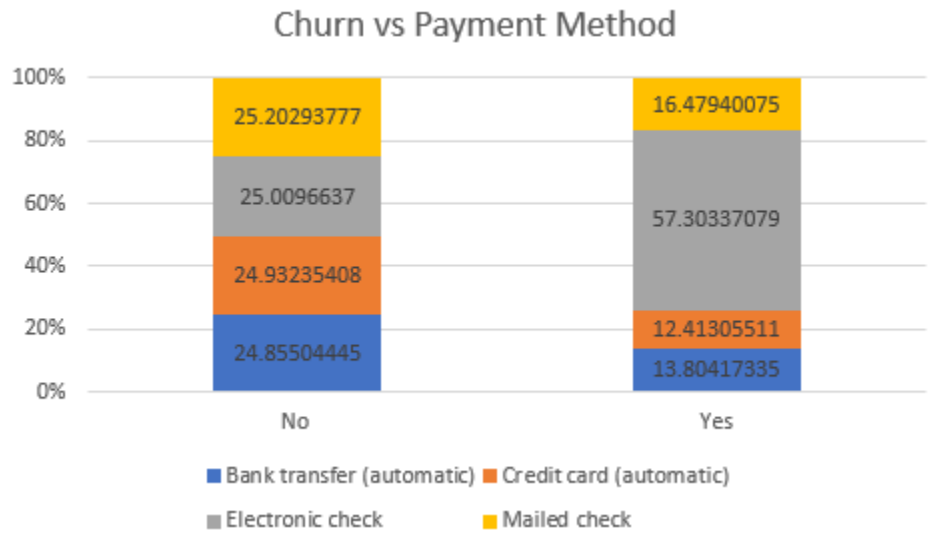
Figure 13: Churn vs Contract



4.2.8 Churn vs Payment method

Figure 14, shows the churn distribution based on the payment method selected by the customers. It can be observed that clients opting for automatic payment methods are less likely to churn than the ones which pay their bills in a conventional manner such in electronic check or mailed check.

Figure 14: Churn vs Payment method



4.3 Discussion and Conclusion

The labelling, encoding and normalising techniques used to pre-process the dataset are the same for original, oversampled and under-sampled dataset. Initially, following pre-processing, the dataset is used to deploy the various combinations of feature extraction and classification methods. Then, similar experimentation is conducted upon differently sampled datasets. The performance of the various techniques applied into the study are compared based on the performance they display on original as well as sampled dataset. The training is done using 10-fold cross validation, following with Accuracy, Recall, Precision and F-measure as performance metrics on the results obtained from test set.

From the experiments conducted, the results achieved from experiment 2 which was carried out on the SMOTE oversampled dataset were very promising and hence the oversampled data was further used in boosted algorithm to check if the overall accuracy as well as the accuracy of correctly predicting the churning customers was increased as compared to predictive learning models in experiment 2. The Random Forest (RF) machine learning algorithm produced the best results among all other algorithms in the experiment 2. The experiment conducted with

boosted algorithms resulted in a significant increase in the accuracy which proved to be an effective technique to achieve the desired results. Based on findings of this research, the random forest with AdaBppst.SAMMA boosting technique produces the most accurate churn prediction results.

Other data variables like “Tenure”, “Internet Service”, “Contract”, and “Payment Method” were a contributing factor towards churning of the customers in the telecom industry data used in this study. This suggests that customer behavioural data plays a vital role in customer relationship management in taking strategic steps to retain clientele.

4.4 Future Work

To extend this study further, hybrid machine learning models can be developed with results from multiple algorithms can be merged together to reduce the probability error and further increase the prediction accuracy of the predictive models. Also, deep learning techniques can enhance the accuracy and is an extensive area still to be researched.

More detailed data, such as call logs, call drops, customer complaints, coverage issues etc can help developers to find patterns and insights and predict an outgoing customer.

Geographic information may be helpful in correlating other variables in churn prediction.

Telcom customer data over a longer period of time can be used to compare the after and before customer behaviour in relation to the provided offers or other incentives.

4.5 Limitations

The main limitations experienced during the research process was that while using the Knime Data Analytics platform the use of SVM machine learning algorithm was producing faulty results. The results produced by this algorithm were with 50% accuracy for predicting the non-churning customers in the dataset. No matter which sampled dataset was used as an input the prediction results were either very low or they were blocked at 50% accuracy.

During boosting process, some machine learning algorithms were producing higher prediction errors hence the boosting loop was unable to run for 100 loops and so had to lower it significantly.

During the feature selection process, the “Feature Selection Loop” node of knime platform suggested to use all the feature in the dataset as it was producing higher accuracy results but the observed results of the correlation matrix were suggesting a different scenario.

References

Abbott, D. (2014). *Applied predictive analytics : Principles and techniques for the professional data analyst*. ProQuest Ebook Central https://ebookcentral.proquest.com

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Umar, A. M., Linus, O. U., & Kiru, M. U. (2019). Comprehensive review of artificial neural network applications to pattern recognition. IEEE Access, 7, 158820-158846.

Ahmed, A. A., & Maheswari, D. (2017). Churn prediction on huge telecom data using hybrid firefly based classification. Egyptian Informatics Journal, 18(3), 215-220.

Ahn, J., Hwang, J., Kim, D., Choi, H., & Kang, S. (2020). A Survey on Churn Analysis in Various Business Domains. IEEE Access, 8, 220816-220839.

Arivazhagan, B., & Sankara, S. D. R. S. (2020). Customer Churn Prediction Model Using Regression with Bayesian Boosting Technique in Data Mining.

Asthana, P. (2018). A comparison of machine learning techniques for customer churn prediction. International Journal of Pure and Applied Mathematics, 119(10), 1149-1169.

Awoyele, T. (2020). Confusion Matrix, Precision , Recall and F1-Score. retrieved on 30th August 2021 from https://medium.com/analytics-vidhya/confusion-matrix-precision-recall-and-f1-score-d5f340e38cca

Azeem, M., Usman, M., & Fong, A. C. M. (2017). A churn prediction model for prepaid customers in telecom using fuzzy classifiers. telecommunication Systems, 66(4), 603-614.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636.

Buttle, F. (2009). Customer relationship management [electronic resource] : concepts and technologies (2nd ed.). Oxford: Butterworth-Heinemann.

Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. Data mining and knowledge discovery handbook, 875-886.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decision Support Systems, 95, 27-36.

Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016). Efficient kNN classification algorithm for big data. Neurocomputing, 195, 143-148.

Devriendt, F., Berrevoets, J., & Verbeke, W. (2021). Why you should stop predicting customer churn and start using uplift models. Information Sciences, 548, 497-515.

Fabris, F., De Magalhães, J. P., & Freitas, A. A. (2017). A review of supervised machine learning applied to ageing research. Biogerontology, 18(2), 171-188.

Faris, H. (2018). A hybrid swarm intelligent neural network model for customer churn prediction and identifying the influencing factors. Information, 9(11), 288.

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

Gončarovs, P., & Grabis, J. (2017, September). Using data analytics for continuous improvement of CRM processes: case of financial institution. In European Conference on Advances in Databases and Information Systems (pp. 313-323). Springer, Cham.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, 220-239.

Hanif, A., & Azhar, N. (2017, December). Resolving class imbalance and feature selection in customer churn dataset. In 2017 International Conference on Frontiers of Information Technology (FIT) (pp. 82-86). IEEE.

Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., ... & Zeng, J. (2015, May). Telco churn prediction with big data.

Idris, A., Rizwan, M., & Khan, A. (2012). Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies. Computers & Electrical Engineering, 38(6), 1808-1819.

Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost.

Jain, H., Khunteta, A., & Srivastava, S. (2021). Telecom churn prediction and used techniques, datasets and performance measures: a review.

Jha, Vishakha (2017). Decision Tree Algorithm for a Predictive Model. retrieved on 30st May 2021 from https://www.techleer.com/articles/120-decision-tree-algorithm-for-a-predictive-model/#:~:text=The%20decision%20tree%20is%20an,outcomes%20based%20on%20certain%20conditions.

Joolfoo, M. B., Jugumauth, R. A., & Joolfoo, K. M. (2020, November). A Systematic Review of Algorithms applied for Telecom Churn Prediction.

Karvana, K. G. M., Yazid, S., Syalim, A., & Mursanto, P. (2019, October). Customer churn analysis and prediction using data mining models in banking industry. In 2019 International Workshop on Big Data and Information Security (IWBIS) (pp. 33-38). IEEE.

Khalid, L. F., Abdulazeez, A. M., Zeebaree, D. Q., Ahmed, F. Y., & Zebari, D. A. (2021, July). Customer Churn Prediction in Telecommunications Industry Based on Data Mining

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering, 30(1), 25-36.

Larose, D.T., & Larose, C.D. (2015). *Data mining and predictive analytics*. ProQuest Ebook Central.
https://ebookcentral.proquest.com

Li, F., & Whalley, J. (2002). Deconstruction of the telecommunications industry: from value chains to value networks.

Naghibi, Seyed Amir, Ahmadi, Kourosh, & Daneshi, Alireza. (2017). Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping. Water Resources Management, 31(9), 2761–2775. https://doi.org/10.1007/s11269-017-1660-3

Naz, N. A., Shoaib, U., & Sarfraz, M. S. (2018). A review on customer churn prediction data mining modeling techniques. Indian Journal of Science and Technology, 11(27), 1-27.

Nettleton, D. (2014). Commercial data mining : Processing, analysis and modeling for predictive analytics projects. ProQuest Ebook Central https://ebookcentral.proquest.com

Oh, Y., Lee, J., & Kim, N. (2018). The contingency value of the partner firm's customer assets in a business-to-business relationship.

Oyeniyi, A. O., & Adeyemo, A. B. (2015). Customer churn analysis in banking sector using data mining techniques. Afr J Comput ICT, 8(3), 165-174.

Pamina, J., Raja, B., SathyaBama, S., Sruthi, M. S., & VJ, A. (2019). An effective classifier for predicting churn in telecommunication. Jour of Adv Research in Dynamical & Control Systems, 11.

Qureshi, S. A., Rehman, A. S., Qamar, A. M., Kamal, A., & Rehman, A. (2013, September). Telecommunication subscribers' churn prediction model using machine learning.

Rababah, K., Mohd, H., & Ibrahim, H. (2011). Customer relationship management (CRM) processes from theory to practice: The pre-implementation plan of CRM system.

Rana, J., Dilshad, S., & Ahsan, M. A. (2021). Ethical Issues in Research.

Ray, S., (2017). 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R. retrieved on 30st May 2021 , from https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/#:~:text=Naive%20Bayes%20uses%20a%20similar,with%20problems%20having%20multiple%20classes.

Ray, S., (2017). Understanding Support Vector Machine(SVM) algorithm from examples (along with code). retrieved on 30st May 2021 , from https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/#:~:text=%E2%80%9CSupport%20Vector%20Machine%E2%80%9D%20(SVM,mostly%20used%20in%20classification%20problems.

Richter, Y., Yom-Tov, E., & Slonim, N. (2010, April). Predicting customer churn in mobile networks through analysis of social groups. Society for Industrial and Applied Mathematics.

Sabbeh, S. F. (2018). Machine-learning techniques for customer retention: A comparative study. International Journal of Advanced Computer Science and Applications, 9(2).

Saghir, M., Bibi, Z., Bashir, S., & Khan, F. H. (2019, January). Churn prediction using neural network based individual and ensemble models.

SAS Institute, (2000). Best Price in Churn Prediction, SAS Institute White Paper.

Sayed, H., Abdel-Fattah, M. A., & Kholief, S. (2018). Predicting potential banking customer churn using apache spark ML and MLlib packages: a comparative study

Shaaban, E., Helmy, Y., Khedr, A., & Nasr, M. (2012). A proposed churn prediction model. International Journal of Engineering Research and Applications, 2(4), 693-697.

Shabankareh, M. J., Shabankareh, M. A., Nazarian, A., Ranjbaran, A., & Seyyedamiri, N. (2021). A Stacking-Based Data Mining Solution to Customer Churn Prediction.

Şimşek Gürsoy, U.T. (2010), Customer churn analysis in telecommunication sector Istanbul University Journal of the School of Business Administration.

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

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. Applied Soft Computing, 97, 105524.

Umayaparvathi, V., & Iyakutti, K. (2012). Applications of data mining techniques in telecom churn prediction. International Journal of Computer Applications, 42(20), 5-9.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, 1-9.

Xia, G. E., Wang, H., & Jiang, Y. (2016, November). Application of customer churn prediction based on weighted selective ensembles.

Xu, T., Ma, Y., & Kim, K. (2021). Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping.

Zhang, L., Li, J., & Wang, Y. (2008, October). Customer relationship management system framework design of Beijing Rural Commercial Bank. In 2008 IEEE International Conference on Service Operations and Logistics, and Informatics (Vol. 1, pp. 97-101). IEEE.

Zhang, X., Zhu, J., Xu, S., & Wan, Y. (2012). Predicting customer churn through interpersonal influence. Knowledge-Based Systems, 28, 97-104.