University *of* Exeter

Master's Dissertation

By Anand Phadtare

Research Topic:

A study on how to build a Churn prediction model and design a retention strategy based on user preferences

Date: 9th January 2023

DECLARATION

I hereby declare that this Applied Research Project that I have submitted to the University of Exeter, Business School for the award of MSc Business Analytics is the result of my own investigations, except where otherwise stated, it is acknowledged by references. Furthermore, this work has not been submitted for any other degree.

Name: Anand Phadtare

Date: 9th January 2023

Course: MSc Business Analytics

Module: BEMM466J - Business Project

Candidate number: 211433

ACKNOWLEDGEMENT

I would like to take this opportunity to sincerely and wholeheartedly thank to my supervisor Dr Stuart So, for his constant and support advice on every stage of my dissertation journey. His feedback and guidance have been instrumental in the successful completion of the project.

He not only encouraged but also challenged my approaches from the early stages of the project. His teachings have converted my thought process from a master's student to a researcher when it comes to solving a problem from real-world business scenarios. He has always been prompt when it comes to responding to my queries over the email however small they might be. His shared wisdom will stay with me forever and will help me shape my further career in the analytics domain.

I would also like to thank the module Leader, Professor Beth Kewell for conducting insightful classes and resolving all our confusion about making our first dissertation of masters.

Lastly and importantly, I would like to thank some of my classmates, who have always been supportive and encouraging throughout the whole process of making this dissertation. My friends and family has always been supporting pillars for completing the project.

Executive Summary

Introduction

This study focuses on identifying, testing and implementing data-driven approaches which give companies a competitive edge whether they are a start-up, midsize companies, or even huge corporate companies. For businesses whose profit margins are dependent on their consumers or customers, understanding their buying patterns and product/service choices empower businesses to make crucial decisions for flourishing the future in the market. One of the factors which will decide a business's future is being able to keep the customers with them before they start investing their money with other competitors in the market. Hence this study will focus on the topic of Churn prediction, to answer the research question of identifying suitable data-driven techniques available to predict the churning of customers for a telecommunication company. And which of the selected techniques can help design an effective strategy to retain customers, who are predicted to stop doing business with the company. The study analyses publicly available churn data of the business under consideration and proposes delivering set objectives through the knowledge acquired.

Analysis and findings

By keeping the objectives in the mind, I did thorough research on available studies which guided me to design my project by following a Cross-industry standard process for Data mining(CRISP-DM) Framework. The framework not only helped to structure my thoughts into actions but has helped me expand my knowledge and added to my experience of delivering an impactful research project.

For analysing the data, machine-learning models were used to draw insightful inferences from the patterns found in the data relating to user choices, overall expenditure on services, and length of tenure till they were actively using services with the telecommunication business. Since machine-learning statistical algorithms have evolved over the years, the study took advantage of comparing the results of multiple such algorithms to decide on the one with the highest accuracy for predicting a chance of customers churning out of the business. But to be able to generate accurate outcomes, the algorithms need pre-processed data rather than raw data. Hence data preparation phase of the research was most crucial to be able to deploy the model for predicting the probable behaviour of new or existing customers.

One of the concepts used for this is training and running testing on the existing dataset. In this concept, the dataset is split into four fragments with randomly selected data, around 80% of this data is used for training the machine-learning model and the rest 20% is used to test if the model is giving desired output.

Another important concept required for processing data helped in dealing with the limitations of machine-learning algorithms. The dataset had an uneven number of churners to non-churners, to be able to train the models for the highest accuracy this imbalance had to be fixed. To achieve this balance a recently discovered technique called Synthetic Minority Oversampling Technique (SMOTE) was used to match the number of churners with non-churners.

Later the processed data was tested on 4 famously used machine learning models for prediction tasks, and an algorithm named the Light Gradient Boosted model(LightGBM) produced the highest churn prediction accuracy of 96.5%, which was later deployed to create a web application using Flask web framework for taking inputs of either new or existing customer and predict their chances of churning. This tool helped in achieving the second objective of the research.

After analysing the relation of the Churn variable with other features from the dataset such as services subscribed, customer demographics, tenure in months, total and monthly charges etc. I decided to look for customers who can be grouped based on above mentioned choices and build another tool to assign new customers to those identified homogeneous segments. This is a very well-known marketing tactic for understanding a business's customer base and designing strategies to keep customers from respective segments satisfied to prolong their churning. To accomplish this a well-known machine-learning algorithm named K-means clustering was used which produced high accuracy metrics for dividing all customers from the dataset into 4 homogenous segments and retention recommendations are

provided based on the mathematical visual representation of those segments. Lastly, this tool and visualizations of segments helped in solving the third problem of the research.

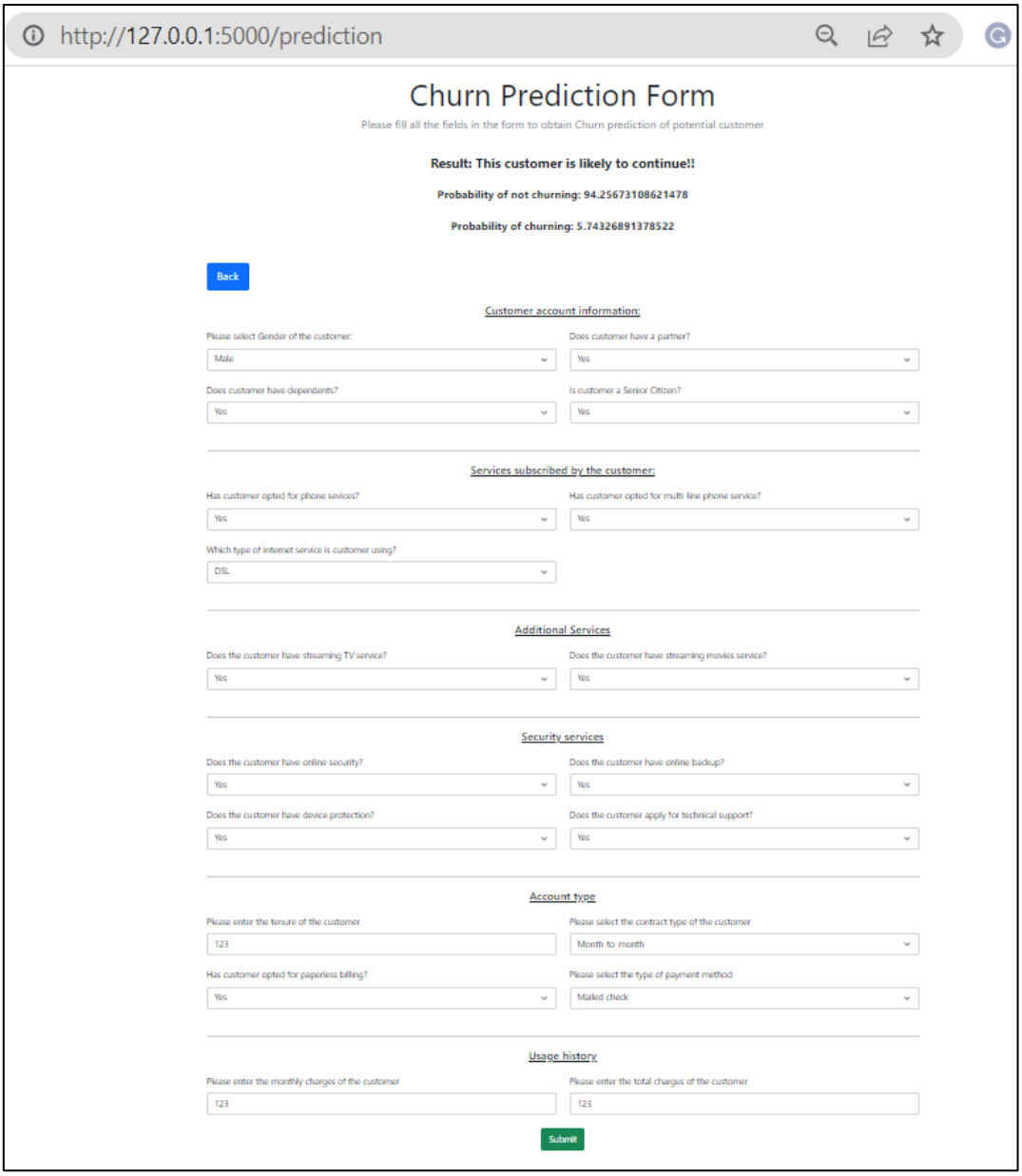
As a master's student, I was able to implement my course's learnings using multiple modules which included knowledge of python programming language, statistical analysis, visualisation tools and ethics to be followed. But most of the knowledge of machine-learning algorithms and the Flask web application framework was learnt during the process of designing the research to deliver the set objectives. Hence to assess the performance of models, I made sure to follow a strong ethical approach by using relevant evaluation metrics wherever relevant in the report.

Recommendations

Since as a student researcher, I only have limited knowledge of the telecommunication business domain and my learnings are based on the available dataset of 7032 customers of a particular business. But the tools built with mentioned accuracies can help business owners and/or marketing managers with expertise in the said field to make sound decisions and design retention strategies to compete in the market by understanding reasons why their customers are leaving or stopped using certain services. The latter part requires gathering customer service preference data through a medium of feedback from existing customers, this will give more insights into which services or processes the telecommunication business need to get fixed, improved and which can be marketed.

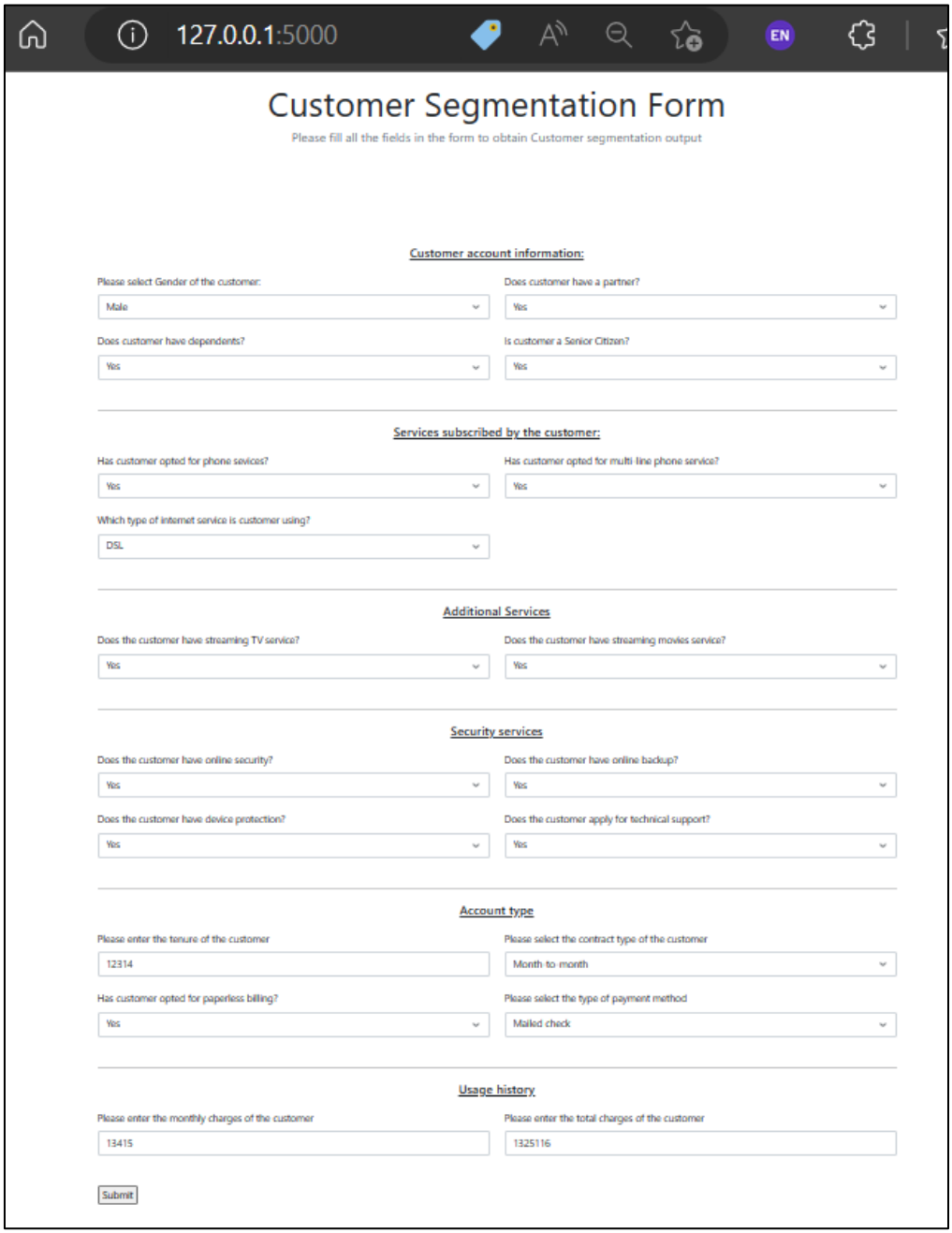
User guide of applications:

The following figure shows an example of Churn prediction output obtained using the Flask application:

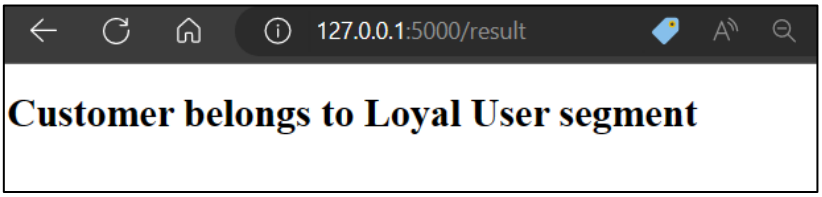


The user of this application simply has to input 19 values which match the customer's account history so far and it will present the output if the customer is going to churn or not, along with the probability percentage of either churning or not churning. The form can be used frequently to understand changes in the probability of a customer churning.

The following diagram shows the output of the Customer segmentation form and its output:



The application form looks the same and also takes the same parameters with 19 values to produce the following output



For a video presentation of user guide, please refer to this link: https://drive.google.com/file/d/17bMbnLWpukR2eccRD6OMDFGj4vL-g7IB/view?usp=sharing

Table of contents

Figures 6
Tables 7
Chapter 1 7
Introduction 8
1.1 Study Background 8
1.2 Research problem 9
1.3 Research Aim 9
1.3 Research Objectives 9
1.4 Research questions 9
1.5 Research motivation and contribution 9
Chapter 2 10
2.0 Literature Review 10
Chapter 3 12
3.0 Research Methodology 12
3.1CRISP-DM 12
3.1.1 Business Understanding 14
3.1.2 Data Description 15
3.1.3 Data Pre-processing 15
3.1.4 Modeling 17
3.1.5 Exploratory Data Analysis 19
3.2 Prerequisites before models' implementation 20
3.2.1 Splitting the dataset 21
3.2.2 Resampling by SMOTE-ENN 21
3.3 Classification Algorithms 23
3.3.1 Decision Tree 23
3.3.2 Random Forest tree 24
3.3.3 Light Gradient Boosted Model(LightGBM) 25
3.3.4 Extreme Gradient Boosted Model(XGB) 25
3.4 Clustering Algorithms 26
3.4.1 K-means Clustering 26
3.4.2 K-prototype clustering 28
Chapter 4 29
4.0 Evaluation 29
4.1 Evaluation metrics for Classification models 29
4.2 Evaluation metrics for clustering models 34
4.3 Results & Comparison 36

Chapter 5 37
5.0 Retention strategy tools 37
5.2 Segment analysis 39
5.3 Retention recommendations 41
Chapter 6 43
6.1 Conclusion 43
6.2 Limitations and future scope 43
References 44
Appendix 47

Figures

Figure 1 The data mining life cycle, IBM (2021) 12
Figure 3 Research design flowchart 17
Figure 4 Churners Vs Non-churners graph 19
Figure 5 Exploratory Data analysis of dataset 20
Figure 6 Understanding train test split, Galarnyk, M. (2022) 21
Figure 7 SMOTE-ENN algorithm, Muntasir Nishat, M. (2021) 22
Figure 8 Decision tree flowchart Chauhan, N.S. (2022) 23
Figure 9 Diagram of a random decision forest, Jagannath, V. (2017) 24
Figure 10 K-means Inertias for clusters 26
Figure 11 K-means Elbow plot 27
Figure 12 K-means Elow location results 27
Figure 13 K-means Clusters graph 27
Figure 14 K-prototype Elbow plot 28
Figure 15 K-prototype Elbow location 28
Figure 16 K-prototype Cluster graph 28
Figure 17 Confusion Matrix 29
Figure 18 Classification report of Decision tree with Imbalanced dataset 30
Figure 19 Confusion matrix of Decision tree with Imbalanced dataset 31
Figure 20 Classification report of Decision tree with SMOTE-ENN 31
Figure 21 Confusion matrix of Decision tree with SMOTE-ENN 31
Figure 22 Classification report of Random forest tree with Imbalanced 31
Figure 23 Confusion matrix of Random forest tree with imbalanced dataset 32
Figure 24 Confusion matrix of Random forest tree with SMOTE-ENN 32
Figure 25 Classification report of Random forest tree with SMOTE-ENN 32
Figure 26 Classification report of LightGBM with Imbalanced dataset 32
Figure 27 Confusion matrix of LightGBM with Imbalanced dataset 33
Figure 28 Classification report of LightGBM with SMOTE-ENN 33

Figure 29 Confusion matrix of LightGBM with SMOTE-ENN 33
Figure 30 Classification report of XGB with Imbalanced dataset 33
Figure 31 Confusion matrix of XGB with Imbalanced dataset 34
Figure 32 Classification report of XGB with SMOTE-ENN 34
Figure 33 Confusion matrix of XGB with SMOTE-ENN 34
Figure 37 Tableau dashboard: Charges, Tenure, Contracts and payments 38
Figure 38 Tableau dashboard Customer Demographics 38
Figure 39 Tableau Dashboard Basic Services 39
Figure 40 Tableau Dashboard additional services 39
Figure 41 XGB model feature importance graph 41

Tables

Table 1 Data description table 15
Table 2 Classification Models results and comparison 36
Table 3 Clustering algorithms results and comparison 36

Chapter 1

Introduction

With an ever-growing and intensely competitive market, every business in the world is always trying to keep their existing customers engrossed in their current products or service. In such situations, a business can stay ahead of its competitors only if they know which consumers are losing interest in its product/services since it is always beneficial for businesses to put efforts into retaining existing customers instead of chasing after new ones. According to an article from Harvard Business School, increasing the customer retention rate by 5% could increase a company's overall profits by 25-95%. As a master' student I am curious to learn about what exactly is involved in the world of businesses when it comes to retaining customers. According to Dan Wolchonok(former Director of Growth and Analytics at HubSpot, 2020), "Retention issues can arise in any industry. Even when retention is increasing and trending in the right direction, businesses still need to keep a close eye on cohorts." However, industry experts have conflicting views and discussions on diverse ways of retaining customers, and one of those ways is predicting customer churn. Customer churn is one of the deciding factors for any business's success or failure rate. Researchers and analysts are always on the hunt for designing strategies and models to better predict the chances of customers churning out of a business.

This research aims to study and identify different techniques to predict customer churn in an organization based on an available dataset of its customer which stores information about user preferences, demographics and charges paid for chosen services. And later focuses on designing a retention strategy based on homogeneous segments identified from the dataset. The research project is presented as a consulting report by analysing customer churn behaviour and helping with a strategy to overcome it. The deliverables of the report will be aligned with the research's aims and objectives.

The background and context will be covered first in this chapter, which will then go on to examine the research problem. The purpose, goals, and questions of the research, as well as its importance and constraints.

1.1 Study Background

In marketing analytics, there are two methods for managing customer retention. The first is the determination of the factor that predicts customer attrition. Another is predicting who will churn and who won't (Kimura et al., 2022). It is not a hidden fact that the success rate of a business is equally driven by the number of active and loyal customers the business has along with the hardworking employees it maintains. Customer loyalty helps create a brand image in minds of customers, saving a lot of resources for the marketing efforts of any business. However, increasing client loyalty is not always possible due to the excessive costs associated with radical quality improvement, service transformation, and extensive marketing campaigns(Kimura et al., 2022).

With the advent of data storage technologies and data analytics in the world, the ways of creating customer loyalty have also evolved to rigorous competitive levels. Predicting chances of customers churning no more requires guesswork, contrary to this big organizations have a set of sophisticated techniques specially designed for retaining customers who are leaving or planning to stop using the business's products or services. Effective churn prediction models are currently being developed by researchers and practitioners(Sharma et al., 1970)

Another crucial step to be taken after identifying which customers are about to churn is to build solid retention strategies. An effective way to design a retention strategy involves understanding your customers' desires and challenges(Bernazzani et at., 2022). Here the concept of customer segmentation comes into the picture, in which, various marketing analytics tools are available to divide available customers into identifiable homogeneous segments. Customer segmentation sometimes referred to as market segmentation, is the categorization of potential customers in each market into distinct groups. Based on customers' shared demands and purchasing habits, that division can be created(Nguyen et al., 2022).

1.2 Research problem

Customer churn is a major concern in the world of businesses, and numerous studies have been conducted by comparing several types of supervised machine-learning algorithms available to researchers to predict the chance of a customer churning. To further improve the precision of predicting such customers, previous studies have used combinations of data-balancing methods for improving the performance of existing classification models. A dataset with skewed class proportions is said to be unbalanced. As a classification model spends most of its training time on the majority class and does not sufficiently learn from the minority class when the class is imbalanced, it is likely to produce poor performance((Kimura et al., 2022). Whereas other studies are focused on identifying customer segments from available churn datasets by using unsupervised machine-learning algorithms; there are limited studies available that are using a combination of churn prediction models and customer segmentation techniques to accomplish a complete version of a retention management strategy. Hence it is crucial to address this opportunity and construct a sequential retention strategy by taking advantage of both approaches which are responsible for achieving a common goal.

1.3 Research Aim

This research aims to write a consulting report for the available churn dataset of a telecommunication business by taking advantage of the identified gaps from earlier research approaches. The findings from analytical techniques will be transformed into application-level deliverables along with retention strategy tools which will help the business make further managerial decisions for retaining its customers.

1.3 Research Objectives

- Research globally recognized analytical techniques for predicting customer churn.

- Testing these techniques on a publicly available dataset of a business and deciding which technique is most suitable for building a prediction model.

- Design a retention strategy using marketing analytics tools for the business based on customers who are predicted to be churned.

1.4 Research questions

- What are currently available machine-learning algorithms for predicting customer churn of a business?
- Are there any measures to decide which of these algorithms would be most suitable for building a prediction model for the business under consideration?
- How to design an effective retention strategy to help marketing managers of the business to make clever decisions on retaining potentially losing customers of a business?

1.5 Research motivation and contribution

The decision to chase after this business domain comes from my personal experience of working as a retention agent in a well-established web hosting company and the inspiration I gained after learning new analytical concepts during my master's journey so far. As a former retention agent, my responsibility is quite self-explanatory, but it only involved direct interaction with a single customer at a time and trying to persuade him/her to continue using the

company's products/services. During my master's studies, I was exposed to so many insightful modules and analytical tools I decided to write a dissertation on the topic which always remained a curious topic to explore on a personal level.

Hence this study will contribute to the knowledge of existing customer retention strategies in the world of business, where marketing tactics are continuously evolving and giving businesses an edge over their market competitors. This will help B2C(Business-to-Consumer) types of businesses reduce customer churning by analysing user preferences and implementing the proposed strategy.

Chapter 2

2.0 Literature Review

In this section of the literature review, we will review and learn from literary work done by various authors who have presented their test results using various classification models in the real world and made comparisons of the results using relevant metrics. The section will also talk about reviewed studies of unsupervised machine learning algorithms, which are beneficial for customer segmentation and designing a retention strategy based on its results. These studies will help in providing structural direction for designing a model which will intern help achieve the research objectives.

If Customer relationship management(CRM) is one of the support pillars for a stable and competitive business, then a good customer retention strategy is the main ingredient required to build this pillar, which helps the business to stand tall against its fierce rivals in the market (Kimura et al., 2022). For telecommunications businesses, client churn is a big issue since it lowers profit. This is especially important given that telecommunications businesses compete in a crowded global market where it is getting harder to keep consumers. Even though many businesses spend a lot of money on marketing to attract new customers, keeping an existing client is typically less expensive than winning a new one. Due to these factors, preventing client turnover has become a top priority for telecom businesses (Zhang et al., 2022). In the long term, retaining customer loyalty and the business' income depends on finding and proposing the best offer that precisely meets the client's demands (Fraihat et al., 2022).

In order to anticipate customer turnover, researchers have used supervised machine learning algorithms(Singh et al.,2018), treating the issue as a binary classification problem(Coussement et al.,2017). The most often utilised algorithms in the earlier research were decision tree, K-Nearest Neighbor, and logistics regression(Hashami et al., 2013). Advanced ensemble learning models(Liang et a., 2019), such as Extreme Gradient Boosting(XGBoost)(Dhaliwal et al., 2018), Light Gradient Boosted Machine (LightGBM)(Tang et al., 2020), and Category Boosting (CatBoost)(Dorogush et al., 2018), have demonstrated good prediction performance in classification issues in recent studies.

The datasets utilised in the customer churn forecast are frequently unbalanced; they contain a disproportionate number of non-churn cases compared to churn cases(Ahmad et at., 2019; Eria & Marikannan, 2018). To balance the data, prior research famously used the Synthetic Minority Oversampling Technique (SMOTE)(Chawla et al., 2002)(Zhang; Chen et al., 2021). Although SMOTE can reduce the overfitting issue that arises from random sampling, it can also produce overfitted models when instances of the majority class invade the minority class space or when the minority class is oversampled and invades the majority class space. Recently, researchers have suggested unique and efficient resampling techniques called hybrid resampling, including Synthetic Minority Oversampling Technique- Edited Nearest Neighbour (SMOTE-ENN) and SMOTE Tomek-Links(Salunkhe & Mali, 2018)(Batista et al., 2004).

The focus of Kimura et al. (2022) work is on creating a prediction model by fusing ensemble learning algorithms with hybrid resampling techniques and evaluating the model's performance against conventional approaches and earlier research. Their study also used the same IBM dataset to test the performances of selected churn prediction techniques. While comparing the performance of the suggested models with that of traditional models and demonstrating the higher performance of the proposed model by Kimura et al. (2022) study's main contribution, in particular, this study demonstrated the higher performance of Boosting algorithms in combination with hybrid resampling techniques. The study provided accuracy scores of algorithms using SMOTE-ENN as follows: Logistic Regression with an accuracy

score(0.697), Random Forest with(0.731), XGBoost with an accuracy score(0.705) and LightGBM with (0.728).To better comprehend and get additional insights into client user preferences for the telecommunications industry, unsupervised machine learning techniques could have been employed to utilise and analyse the current dataset.

To accommodate the various demands of its consumers, telecom firms typically provide a variety of pricing plans or bundles. In the long term, retaining customer loyalty and the business's income depends on finding and proposing the best offer that precisely meets the client's demands. The study done by Fraihat et al.(2022) offers a useful technique for identifying a client base with the possibility to upgrade their telecom plan.

Client segmentation and profiling are popular and regularly utilised methods to comprehend customer behaviour. Several papers have addressed this issue(Tripathi et al., 2018). Using clustering algorithms, Tripathi et al. (2018) investigated the significance of client segmentation in customer relationship management (CRM) data. For CRM data from a mall, they employed hierarchical clustering and k-means. Customer name, gender, age, yearly income, and expenditure score were the data elements. K-means provided higher results in terms of time and accuracy because the dataset was short. K-means, according to Tripathi et al. (2018) may handle bigger datasets more effectively than hierarchical clustering. However, the choice of the number of clusters is k-means' constraint (k).

To ascertain the homogeneity of distinct clusters and the dissimilarity of consumers within a cluster, Ramachandran, S. et al. (2011) examined customer behaviour, preferences, and characteristics in a telecom firm. To understand both global and micro client patterns and behaviour, they employed two-layer clustering. The quantity of consumption for each client is examined in the first layer, called customer value, and consumer behaviour characteristics are used for further categorization in the second layer. According to business expert guidelines, the first layer is divided into several clusters, and then each cluster in the first layer is sub-clustered in the second layer.

Recency, Frequency, and Monetary (RFM) model and k-means clustering were used in Insani and Soemitro's et al.(2016) data mining approach to profile the customers. The data is from an Indonesian telecommunications firm. The RFM model serves as the k-means model's input. K-means is used to build customer profiles based on customer segmentation as well as customer use, invoice, and payment information, allowing the model to identify lucrative clients. The findings demonstrate the viability of the method's adoption and the categorization of consumers as lucrative, devoted, or prone to churn (Fraihat et al.(2022).

K-prototype clustering and a novel feature selection method based on the apriori algorithm were employed in research by Das et al. (2021) to minimise the size of the dataset through the selection of pertinent features. To anticipate lost consumers, an ensemble classifier is created in the second step by mixing KNN, Naive Bayes, SVM, Decision trees, and Logistic Regression. They also used the same telecommunication dataset of IBM to predict the customers who are about to churn from the existing data.

Although all the above-reviewed pieces of literature have used various machine-learning algorithms to build churn prediction and customer segmentation models and presented their findings using a range of metrics, they are missing interface-level applications which will help determine the behaviour of a new customer added to the existing dataset. An application will not only help identify churning chances of existing and/or new customers in a business but will also help decision-makers of the business to produce strategies to focus on retaining them before they plan to stop using products/services. Hence this study will be focusing on delivering a consultancy report based on the results of the models and designing an application to help make marketing decisions for the telecommunication business under consideration.

Chapter 3

3.0 Research Methodology

3.1CRISP-DM

Data mining has become an important need for any business running today it is being widely used in marketing, sales, manufacturing, service-based industries, sports you name it, and it is being used by them somewhere or the other. As a result, there seems to be a standard approach which can be used by companies, and employees so that it can be easier for everyone to understand. Cross Industry Standard Process for Data Mining is known as CRISP-DM. When using analytics to resolve business difficulties, the CRISP-DM technique is practical, adaptable, and helpful (Saltz, J.S. and Hotz, N. 2020).

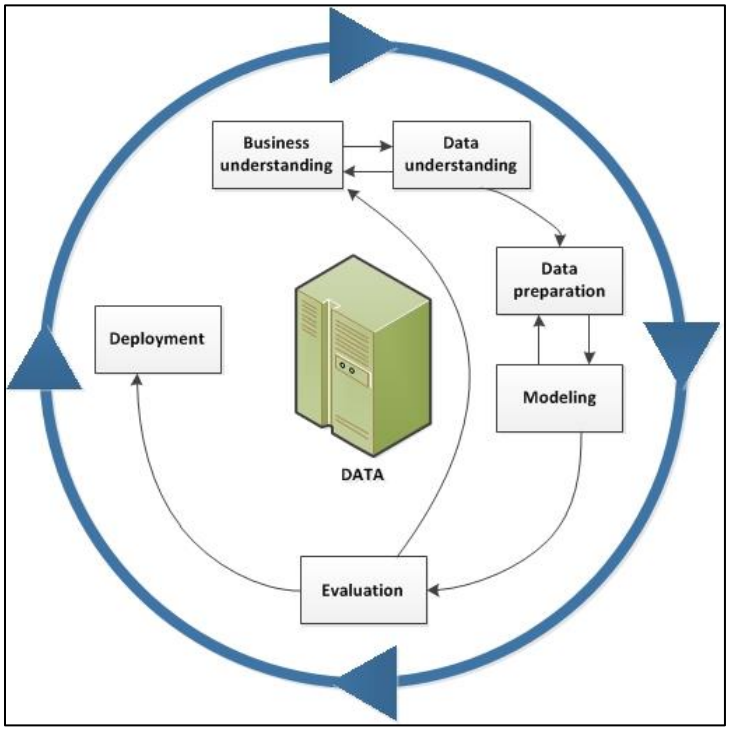


Figure 1 The data mining life cycle, IBM (2021)

CRISP-DM is a data mining technique, methodology, or procedure that aids in or gives instructions on how to carry out a data mining project. Major corporations including Daimler Benz, ISL, NCR, and OHRA created it, and it was first put into operation in 1996. These businesses actually installed 200 data mining tools and users before developing this methodology. Everyone is welcome to use it because it is a documented, open-source procedure that is not owned by anybody(Saltz, J.S. and Hotz, N. 2020).

CRISP-DM offers a roadmap, best practices, and frameworks for leveraging data mining to provide better and quicker outcomes. In this way, it aids the company in planning and executing data mining projects.

Phase 1: Business Understanding:

Business understanding is the first stage where we convert a business objective or understand the project from a business perspective before we convert it to data mining tasks, so we convert a business objective into a data mining objective, or a data mining task where we can apply technologies for modelling into it.

Four major tasks focused on in this phase will be:

1. Determine the business objective

2. Evaluate the situation
3. Determine data mining goals
4. Prepare a project plan

Phase 2: Data Understanding

The second stage of the CRISP-DM process requires us to find the appropriate data required for the project. The initial collection might include a loading of data as well if we are using any specific tool for it. As well as there can be cases where we might need data from multiple sources, so how and when we will integrate all this data will be discussed in this state.

Four major tasks focused on in this phase will be:

1. Data collection
2. Data description
3. Data Exploration
4. Verify data quality

Phase 3: Data Preparation

The third phase is data preparation when we create the final data set that will be used for modelling, which is the following phase. We have the data, we have acquired the data, and we have the quality. To put it simply, the following phase will include using modelling tools, thus it is essential to gather all the data and establish the final data set.

Tasks involved in this phase are:

1. Data selection
2. Data Cleaning
3. Data integration
4. Data formatting

Modelling

When modelling, we will focus on providing several model strategies, choosing, and using them, later we will test whether we can use them, and choose our alternatives.

Four tasks for the modelling phase:

1. Model selection
2. Model testing
3. Creating model
4. Assess the model

Evaluation

In evaluation, we will create and work with our business objectives, then produce evaluation sheets, then produce process reviewing, and then we will see if there is anything that we must determine for the next steps. So here, in evaluation, we summarise the entire result, and then we give it as a business criterion.

Deployment

Here, in the sixth and last phase, we deploy. When we deploy, we either provide the report, decide to move the project forward, or move it to the next stage of the business process.

Four major tasks in this phase are:

1. Plan deployment
2. Plan Monitoring
3. Plan the final report
4. Review the project

3.1.1 Business Understanding

From the secondary data obtained from the IBM community, we have assumed the following situations about the business and its customers. These assumptions are made based on variable and their description mentioned on the IBM Business Analytics Community webpage(Community, I.B.M. 2017).

From this information we understand that the telecommunication business is providing a bundle of services to its customers as follows:

1. Home phone service
2. Internet service
3. Multiple lines: This indicates if the customer subscribed to multiple telephone lines
4. Online Security
5. Online backup
6. Device protection plan
7. Technical support

The telecommunication business offers the following types of contracts for using these services:

1. Month-to-month
2. One year
3. Two years

Payment methods choices used by customers to pay for the above services are:

1. Bank withdrawal
2. Credit card
3. Mailed check

The dataset also stores information about customer demographics as follows:

1. Gender
2. Senior Citizen
3. Dependents

Data collection:

To build a prediction model for a business, gathering suitable quantitative data was a crucial step of the research. Since customer churn prediction and customer segmentation can be derived based on a historical customer dataset of a

business, which is not available to access for the public. A secondary data collection method was used to attain a suitable dataset which was publicly available on the platform www.kaggle.com

The IBM dataset, which is an open-source customer attrition dataset in the telecommunications industry, was used in this study. It was first shared in the IBM community and is now available on the Kaggle website (https://www.kaggle.com/datasets/blastchar/telco-customer-churn). Lalwani et al.(2021) and Pamina et al.(2022), Takuma Kimura et al.,(2022)are three recent research that employed the IBM dataset (2019). There are 7043 instances (customers) and 21 variables in the raw data. The dataset contains information on each customer's demographics, internet connection environment and related support, contract terms, billing and payment methods, and the amount charged.

3.1.2 Data Description

The following table shows the names of variables and their definition from the selected dataset:

Variables in the dataset	
Variable names	Definition
Customer ID	Unique Customer reference number
gender	Whether the customer is a male or a female
SeniorCitizen	Whether the customer is 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)
tenure	Number of months the customer has stayed with the company
PhoneService	Whether the customer has a phone service or not (Yes, No)
MultipleLines	Whether the customer has multiple lines or not (Yes, No, No phone service)
InternetService	Customer's internet service provider (DSL, Fiber optic, No)
OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)
OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)
DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)
TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)
StreamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)
StreamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)
Contract	The contract term of the customer (Month-to-month, One year, Two year)
Paperless	Whether the customer has paperless billing or not (Yes, No)
PaymentMethod	The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
MonthlyCharges	The amount charged to the customer monthly
TotalCharges	The total amount charged to the customer
Churn	Whether the customer churned or not (Yes or No)

Table 1 Data description table

3.1.3 Data Pre-processing

For data pre-processing, Python programming was used on the Jupyternotebook platform. The variable 'Churn' is binary (Yes or No), which was used to find that there are 1,869 churners and 5,174 non-churners in it. Since the percentage of churners is 26.53% and non-churners are 73.46% the dataset is considered uneven or imbalanced.

The analysis will consider all the variables since the goal of this study is to create models with excellent prediction performance rather than to identify the causal connection between predictors and result. To increase the model's parsimony, variables that provide no information that may be used to make predictions will be removed. ‘CustomerID’ variable which is a unique identification number assigned to each customer was removed since it will not be useful for analytical processes. 'SeniorCitizen' and 'tenure' are integer variables, whereas 'MonthlyCharges' is a float variable. The remaining 18 variables are object-type data. The ‘TotalCharges’ variable was also present as object-type data, which was converted to float type.

It was discovered by carefully examining the values included in the variable "TotalCharges" that the variable has 11 instances where a space encased in quotation marks was entered. These 11 examples were eliminated because there are no reliable interpretations of their meaning and because 11 is a small amount in comparison to the overall number of cases.

The categorical variables from the datasets which are: gender, SeniorCitizen, Partner, Dependents, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, Paperless, PaymentMethod were encoded as Dummy variables. Numerical variables known as "dummy variables" are variables that simulate real data. For example, if given the genders of male and female, you may enter 0 for the men and 1 for the females. This provides a realistic representation of the data, which is also presented in numerical form and can be included in the machine learning model(Nyakara, 2021).

The dataset once these pre-processing steps are completed has 7,032 instances, and 46 variables, including the label variable.

3.1.4 Modeling

Research design

The following flowchart is a representation of the research design, which gives an overview of each stage of the CRISP-DM approach for delivering a data mining project.

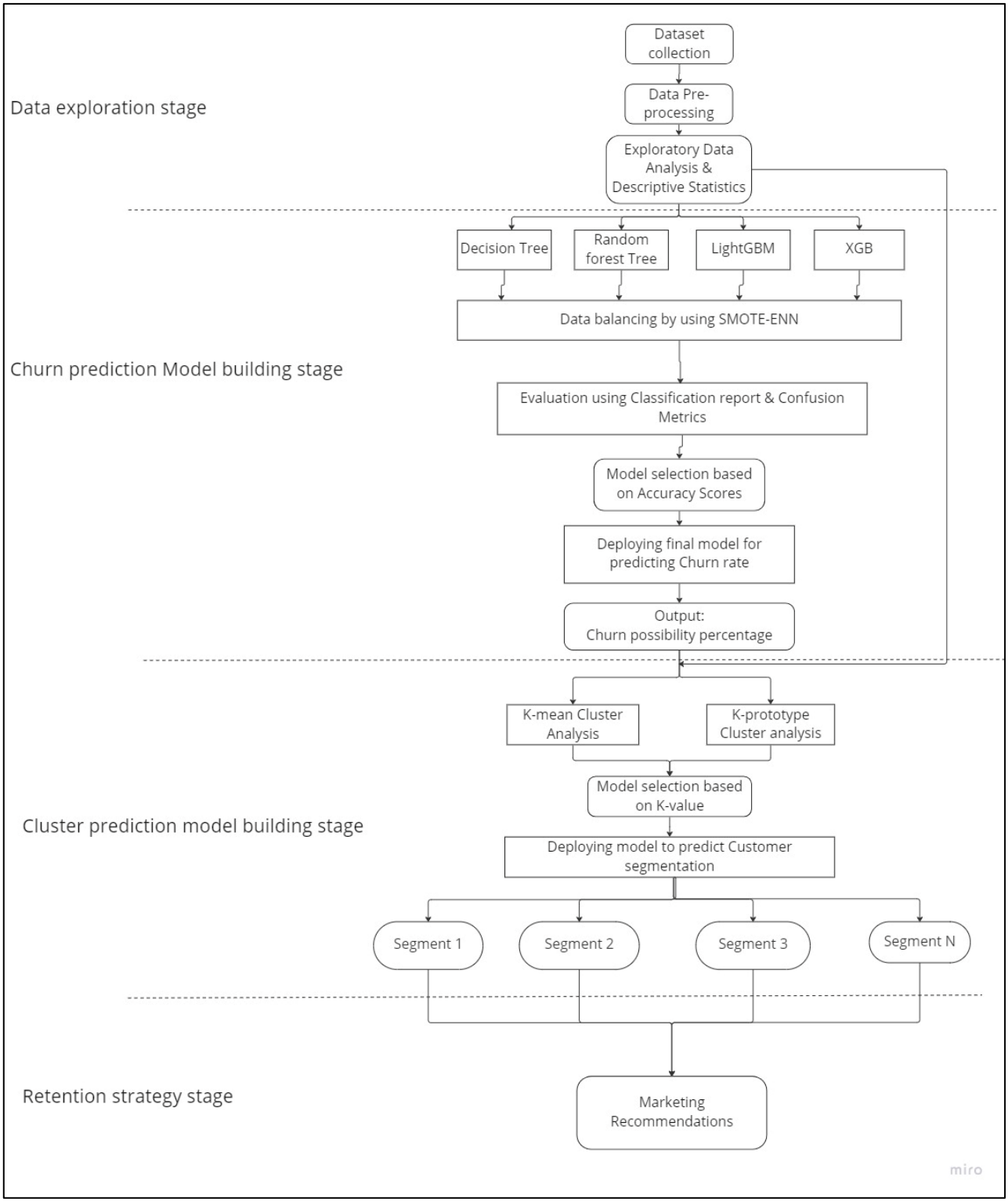


Figure 2 Research design flowchart

In the first stage, we talked about Business understanding and data collection phases. The next phase conducts the data pre-processing procedures for data cleaning, formatting, and corrections. After that data exploration was done by using Exploratory data analysis(EDA) and descriptive statistics(DS) to understand the relationship between the Churn variable, which would be treated as the dependent variable and the rest of the remaining variables will function as independent variables except for the label variable: CustomerID.

In the second stage, we will use the processed data to feed into the choice supervised machine learning classification models to evaluate which of them produces the highest accuracy for predicting customer churn. Based on a thorough study of the literatures following models are selected for this phase with appropriate reasons for choosing them:

1. **Decision tree:** The key advantages of employing a decision tree in machine learning are its simplicity and ease of visualising and comprehending the decision-making process(Seldon 2021). Pruning the tree structure is frequently required because decision trees in machine learning can produce unnecessarily complicated branches by creating highly granular branches(Hashami et al., 2013).
2. **Random forest tree:** The random forest tree method is a user-friendly and adaptable machine learning system. Organizations may overcome regression and classification issues by using ensemble learning(Mbaabu, O. 2020). Developers are recommended to adopt this approach since it addresses the issue of dataset overfitting Kimura et al. (2022). It is a highly useful tool for creating the precise projections required in organisational strategic decision-making.
3. **Gradient Boosted Model(GBM):** For higher efficiency and faster training LightGBM uses a histogram-based approach, which accelerates training by bucketing continuous feature values into discrete bins. Reduced memory utilisation is achieved by switching continuous values to discrete bins. Using a leaf-wise split strategy rather than a level-wise split approach, which is the primary element in getting greater accuracy, creates far more complicated trees. When compared to XGBoost, it can manage huge datasets just as effectively while training takes a lot less time(Surarna, S.U.B.H.A.M. 2020)(Tang et al., 2020).
4. **Extreme Gradient Boosted model(XGBoost):** XGBoost is a tree-based ensemble machine learning technique that improves on the Gradient Boosting framework by incorporating certain precise approximation algorithms. It offers improved prediction power and performance. In data science competitions, XGB is regularly employed and rises to the top of the scoreboard. Greatly Boosted, or XGBoost (Dhaliwal et al., 2018) Kimura et al. (2022)(Krishna, et al, 2020).

Since the dataset at hand is highly imbalanced, meaning there are more cases of non-churners denoted by Churn value: 0, than that of Churners denoted by Churn value: 1, for models to provide better and more accurate results SMOTE-ENN resampling technique is used to balance the dataset. Simply put this technique will produce enough cases of Churners to match the number of cases of non-churners.

Post balancing the dataset, the models are run again to check for improvement in the accuracy scores, which is done by using evaluation metrics: Classification report and Confusion metrics. These matrices allowed us to select an appropriate model to be used for building a churn prediction model, which will provide an accurate value for a customer who is predicted to be churned.

In the next phase, the dataset was assessed for finding homogeneous segments using two clustering algorithms as shown in the figure. The pre-processed dataset which has EDA and DS calculations will be used here to perform clustering.

For identifying customer segments, the following cluster analysis algorithms were used:

1. **K-means:** Implementing K-means is simple. K-Means is computationally quicker than hierarchical clustering when there are many variables (if K is small). Higher clusters can be produced using K-Means than by hierarchical clustering. As soon as the centroids are recomputed, an instance may switch clusters (transfer to a different cluster)(Santini, M. 2016)(Fraihat et al.,2022) (Tripathi et al.,2018).

2. **K-prototype:** A hybrid clustering technique that can manage both categorical and numerical data is the k-prototypes algorithm. This work enhanced the procedure for choosing the initial Cluster Centres and presented a new Hybrid Dissimilarity Coefficient(Jia, Z. and Song, L. 2020)(Das et al.,2021).

Based on comparison of K-values and evaluation metrics used for clustering algorithms, one with an appropriate score was selected for the next stage of the design, which is deploying a model to predict the customer segment of the customer who is under consideration. In the last stage, once the segment for this customer is identified, a relevant marketing strategy is recommended.

3.1.5 Exploratory Data Analysis

Exploratory data analysis is the critical process of doing preliminary analyses of data to find patterns, spot anomalies, and test hypotheses and assumptions using summary statistics. We employ this strategy to enable us to construct a hypothesis based on the facts we have and then attempt to apply various concepts and methodologies to it appropriately. Additionally, we anticipate that it will help us better understand how to approach feature selection.

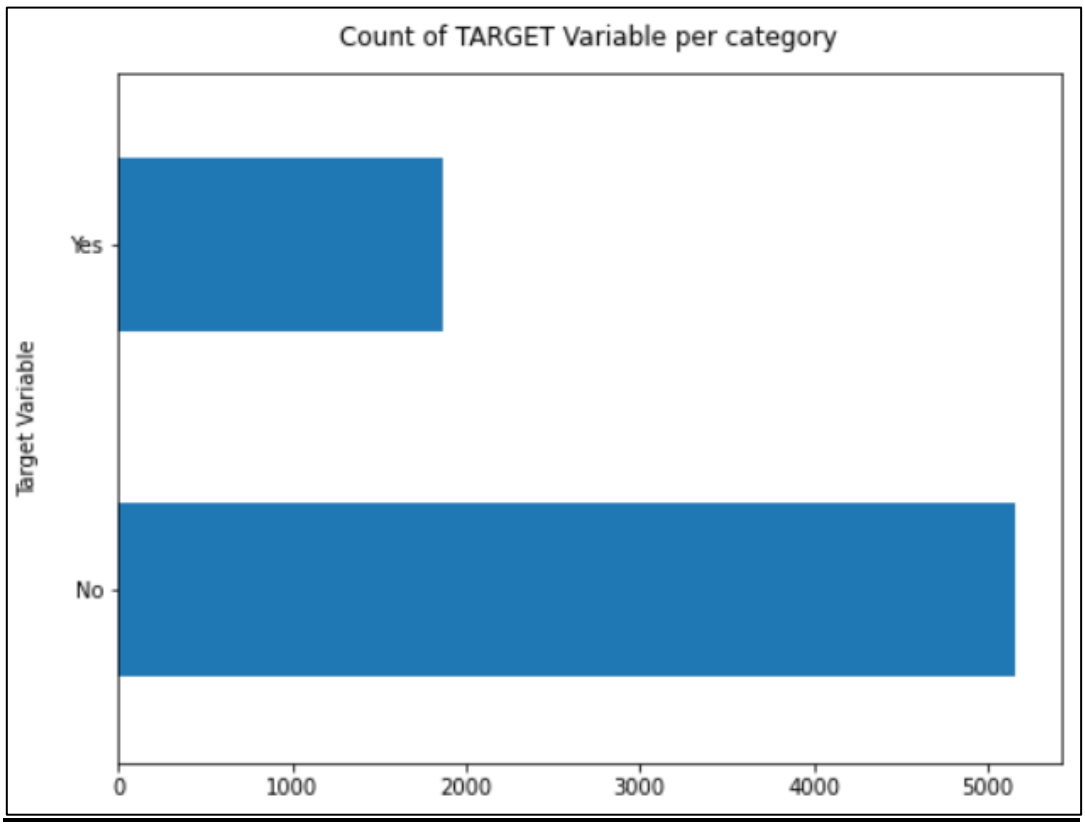


Figure 3 Churners Vs Non-churners graph

From the above figure, we can see that there are two values of the Target variable which is the 'Churn' variable count for 'Yes' indicating customers who churned and 'No' indicating customers who are not churned. This entailed that the dataset is highly imbalanced, where there are many non-churners against the number of churners. For producing accurate results using machine-learning algorithms, classification problems requires a balanced dataset to make sure to

give equal priority to both the classes. This process also helps prevent loss of information and mitigates overfitting which could be caused by oversampling.

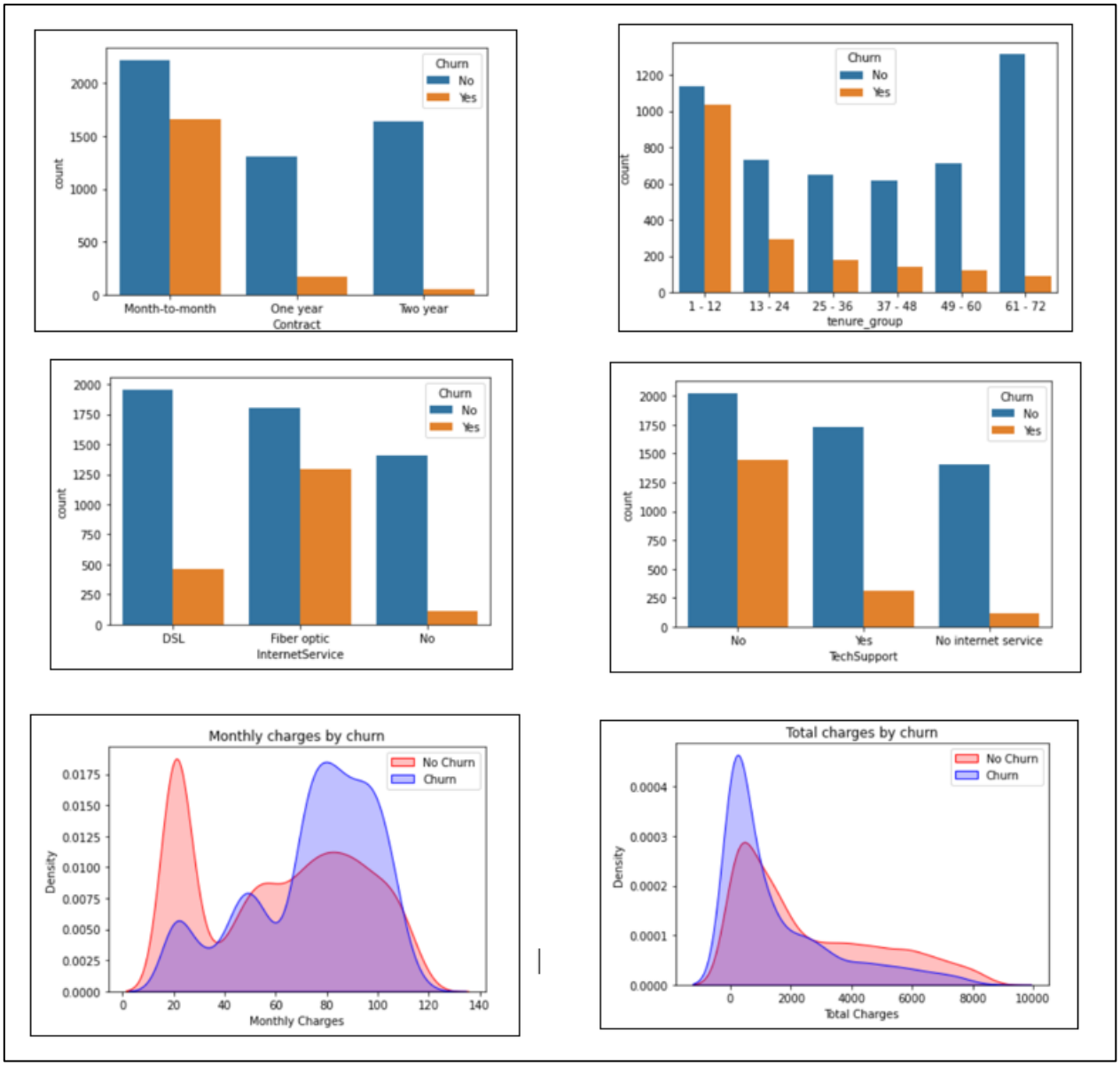


Figure 4 Exploratory Data analysis of data

These insights will be very useful for building an effective strategy in a later chapter of the research.

3.2 Prerequisites before models' implementation

Introduction

Before we start with the actual implementation of our models there are some prerequisites which need to be completed beforehand so that our code gets implemented with high-precision results. All of the models were coded using Python programming on the Jupyternotebook platform using the Anaconda application, which runs on the local machine. The dataset under consideration was pre-processed using the same platform and the newly pre-processed version of the dataset was present under the home directory of Jupyternote, which made it easy to import while implementing each model. Basic libraries such as pandas, NumPy, seaborn and matplotlib were imported and loaded. Pandas are used for importing the dataset. Numpy is used for linear algebra, seaborn and matplotlib; both are needed for doing graphical presentations.

We will use the CSV () function to load the data in Jupternotebook . We will be using other functions of pandas such as drop() which is used to drop/ delete columns from our dataset, and astype() is used to alter the data type of the columns. While importing the dataset a column named: Unnamed is by default created, which created index values for the rows. This column was dropped to avoid any computational errors in the implementation stage.

3.2.1 Splitting the dataset

A dataframe named 'x' was created to separate all the independent variables from the dependent variable, which is the 'Churn' variable. This was achieved by using drop() function of python. Similarly, another dataframe named 'y' was created for the target variable by only choosing the 'Churn' column from the imported dataset. These steps were executed for every machine-learning model's python file to segregate the independent variables from the dependent variable.

Building a model that works well with additional data is one of the objectives of supervised learning. Since are going to build a churn prediction application, it is ideal to test our model before implementing it in the final stage. The application will use new input values for every customer to be predicted for churning, hence a technique like train-test split can be used to imitate this experience(Galarnyk, M. 2022).

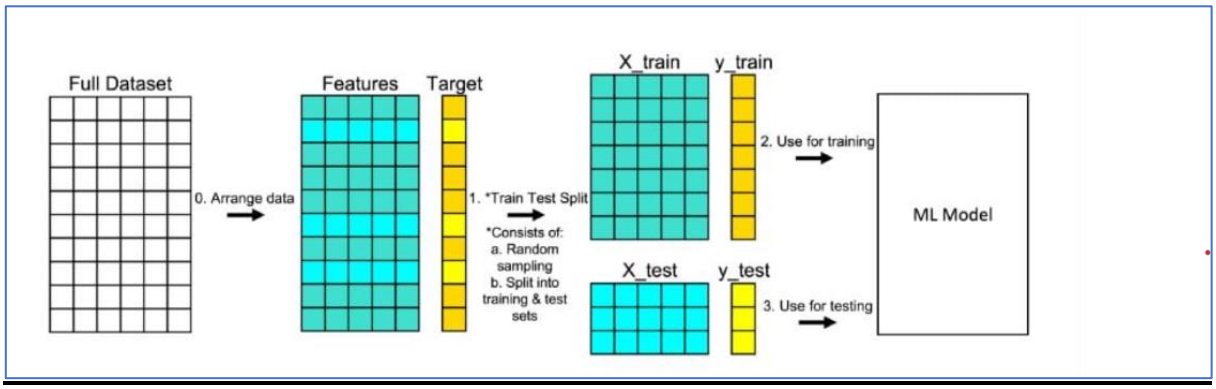


Figure 5 Understanding train test split, Galarnyk, M. (2022)

The ratio of 80:20, which designates 80% of the data as training and 20% as testing, is frequently employed(V, R.J. 2022). In our case, the splitting of the dataset is done based on the proportion of Churners to non-churners as mentioned in the data pre-processing stage. Hence the testing data size is kept as 26.5%(which is the number of churners). For this process, a function named 'train_test_split' was imported from the library 'sklearn.model_selection'. The size of test data is kept the same across all the machine-learning algorithms implementation to avoid any bias in the splitting process of the dataset as well as in the results generated as a cause of it.

3.2.2 Resampling by SMOTE-ENN

Synthetic Minority Oversampling Technique (SMOTE)

For overcoming the issue of the imbalanced dataset for the machine-learning algorithms, this study will be using one of the most popular oversampling techniques developed by Chawla et al. (2002) called SMOTE. SMOTE creates examples based on the distance of each data (often using Euclidean distance) and the minority class's nearest neighbours, thus the created examples are distinct from the original minority class. Random oversampling juscopyes a few random instances from the minority class(Chawla et al., 2002)

.

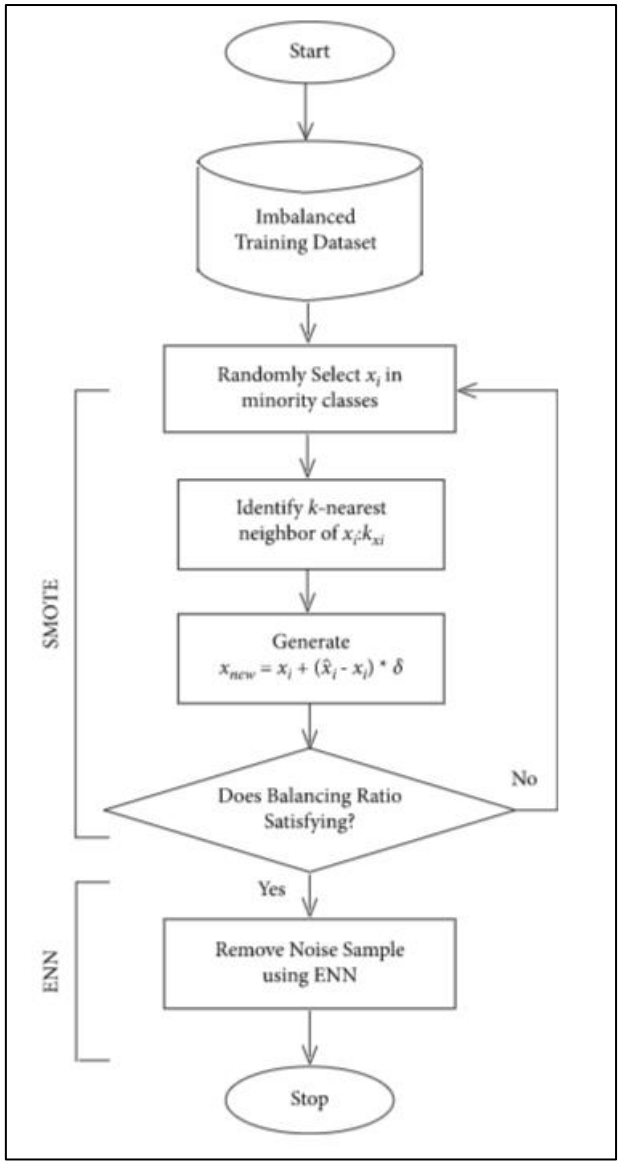


Figure 6 SMOTE-ENN algorithm, Muntasir Nishat, M. (2021)

The procedure to create the synthetic samples is as follows:

1. Select arbitrary information from the minority class.
2. Identify the k closest neighbours of the random data and determine their Euclidean distance.
3. Divide the difference by a number chosen at random between 0 and 1, then add the result to the minority class as a synthetic sample.
4. Continue until the necessary percentage of the minority class is reached.

In contrast to the original oversampling approach, this method adds new "information" to the data since the synthetic data that are created are near the feature space on the minority class.

Edited Nearest Neighbour (ENN)

The ENN technique, created by Wilson (1972), finds each observation's K-nearest neighbour first and then determines whether or not the majority class from that neighbour's k-nearest neighbour matches the observation's class. The observation and its K-nearest neighbour are removed from the dataset if the majority class of the observation's K-nearest neighbour and the observation's class vary. By default, ENN uses K=3 as the number of nearest neighbours(Chawla et al., 2002).

The following will describe how the ENN algorithm works.

1. Determine K as the number of closest neighbours for the dataset with N observations. Unless otherwise specified, K=3.
2. Among the other observations in the dataset, locate the observation's K-nearest neighbour. Then, retrieve the observation's majority class from the K-nearest neighbour.
3. If the class of the observation and the majority class from the observation's K-nearest neighbour are different, the observation and its K-nearest neighbour are eliminated from the dataset.
4. Repeat steps 2 and 3 as necessary to get the required percentage of each class.

This technique, created by Batista et al. (2004), combines the SMOTE ability to create synthetic examples for the minority class and the ENN ability to delete some observations from both classes that are identified as having a different class from the observation's class and its K-nearest neighbour majority class(Viadinugroho, R.A.A. 2021).

For utilizing functionalities of the SMOTE-ENN technique in python, SMOTEENN class was imported the from library imblearn.combine. SMOTEENN() function was used to resample the split data to create a balanced version of it. The model was refitted with new version of the data and a newly created data point. Later the results were compared with the original model.

3.3 Classification Algorithms

3.3.1 Decision Tree

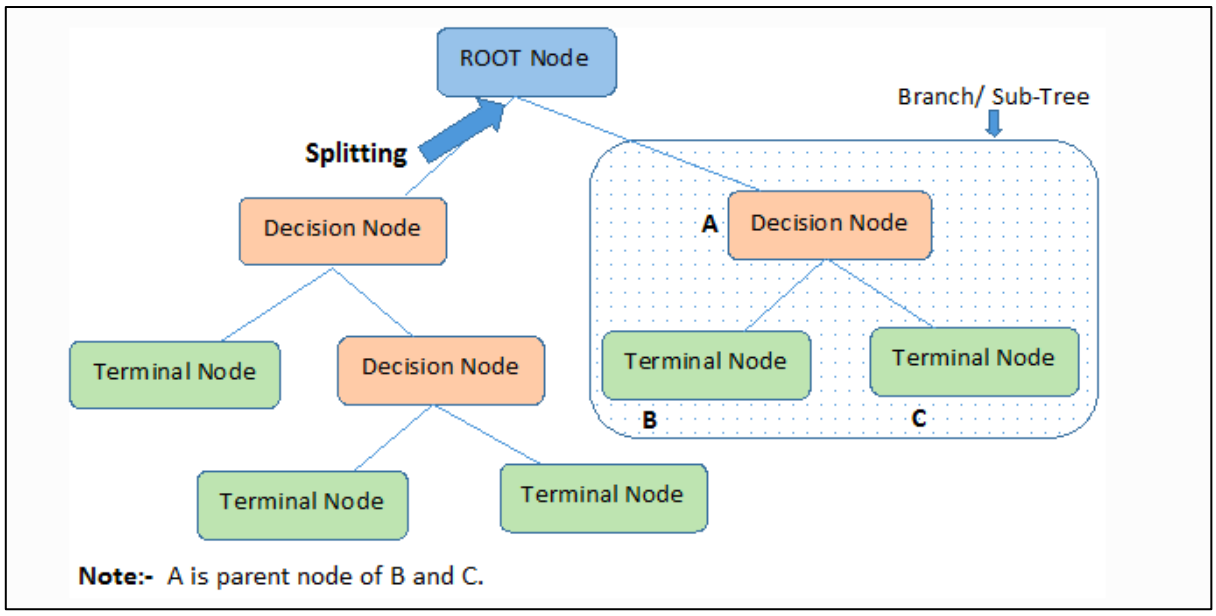


Figure 7 Decision tree flowchart Chauhan, N.S. (2022)

Steps of creating a decision tree:

Step 1: We begin the tree with the complete dataset and call this the root node.

Step 2: We find the best attribute in the dataset using the Attribute Selection Measure

Step 3: Now divide the root node into subsets that contain possible values for the best attributes.

Step 4: Generate the decision tree node, which contains the best attribute.

Step 5: We would recursively make new decision trees based on the subsets created by us in step 3 and we will continue this process until we get to a point where we cannot further divide the points and this will be our leaf node.

Implementation

For implementing Decision tree classification, DecisionTreeClassifier class was imported from the library sklearn.tree. The dataset was split into predictors variables and Target variable according to the decided proportion from splitting the dataset by using the train_test_split() method and by keeping the test_size=0.265, matching the proportion of non-churners.

Decision tree model was built using DecisionTreeClassifier() function, passing parameters by passing appropriate parameters to get the accurate output. Later training data was fit into the model using model_name.fit() method and a prediction variable and test data of independent variables passed into it using predict() method from DecisionTreeClassifier class.

3.3.2 Random Forest tree

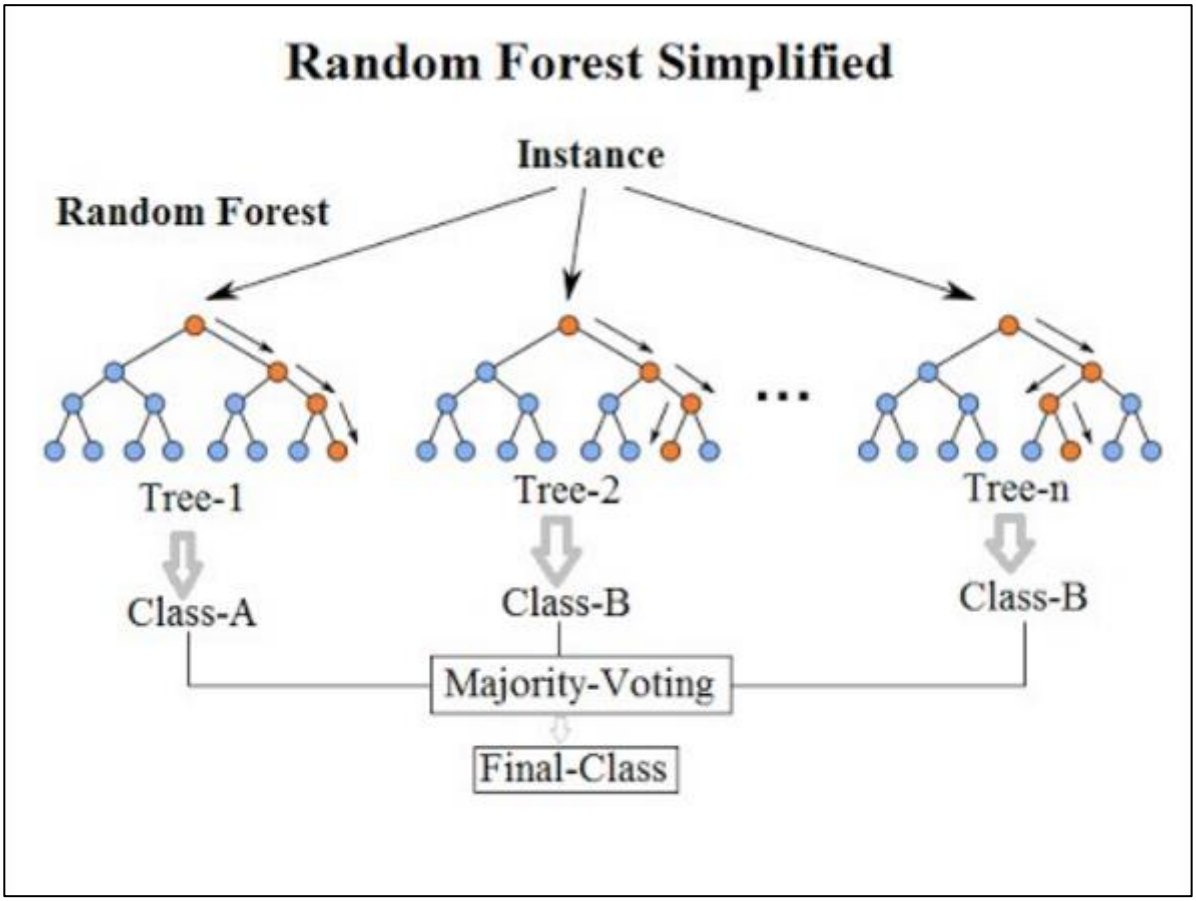


Figure 8 Jagannath, V. (2017) Diagram of a random decision forest

Steps involved in Random Forest:

Step 1: N number of records are taken from the dataset.

Step 2: It creates individual decision trees for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output of the random forest is considered based on majority voting or aggregation depending on classification or regression respectively.

Implementation

For implementing Random Forest tree classification, RandomForestClassifier class was imported from the library sklearn.ensemble. The dataset was split into predictors variables and Target variable according to the decided proportion from splitting the dataset by using the train_test_split() method and by keeping the test_size=0.265, matching the proportion of non-churners.

The random forest tree model was built using RandomForestClassifier() function, passing parameters by passing appropriate parameters to get the accurate output. Later training data was fit into the model using model_name.fit() method and a prediction variable and test data of independent variables passed into it using predict() method from RandomForestClassifier class.

3.3.3 Light Gradient Boosted Model(LightGBM)

In contrast to previous boosting algorithms that develop trees level-by-level, LightGBM divides the tree leaf-wise. It selects the leaf with the greatest delta loss for growth. The leaf-wise algorithm has less loss than the level-wise algorithm since the leaf is fixed. The complexity of the model could rise as a result of leaf-wise tree growth, which could also result in overfitting in limited samples (Singh, S. 2021).

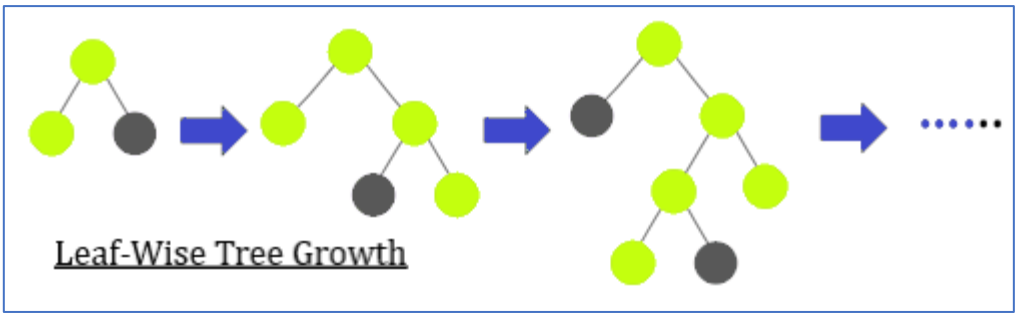


Figure 9 Diagrammatic representation of Leaf-Wise Tree Growth, Singh, S. (2021)

Implementation

For implementing LightGBM classification, the lightgbm library was imported. The dataset was split into predictors variables and Target variable according to the decided proportion from splitting the dataset by using the train_test_split() method and by keeping the test_size=0.265, matching the proportion of non-churners.

LightGBM model was built using the LGBMClassifier() function, passing parameters by passing appropriate parameters to get the accurate output. Later training data was fit into the model using model_name.fit() method and a prediction variable and test data of independent variables passed into it using predict() method from LGBMClassifier class.

3.3.4 Extreme Gradient Boosted Model(XGB)

Extreme Gradient Boosting, or XGBoost, is a distributed gradient boosting toolkit that has been developed to be very effective, adaptable, and portable.. It is a gradient-boosting algorithm that uses decision trees as the base learner. Here are the steps for the XGBoost algorithm:

1. Initialize the model with a set of hyperparameters and an objective function.

2. Use the training data to build a series of decision trees.

3. For each tree, apply the following steps:

   a) Calculate the error between the predicted value and the actual value for each sample in the training data.
   b) Calculate the gradient of the error concerning the predicted value.
   c) Update the prediction by moving in the direction that reduces the error.

4. Repeat steps 2 and 3 until the maximum number of trees is reached or the error is minimized.

5. Make predictions on new data using the trained model.

6. Tune the hyperparameters and repeat the process until the desired model performance is achieved.

Implementation

For implementing LightGBM classification, lightgbm xgboost was imported. The dataset was split into predictors variables and Target variable according to the decided proportion from splitting the dataset by using the train_test_split() method and by keeping the test_size=0.265, matching the proportion of non-churners.

XGB model was built using the XGBClassifier() function, passing parameters by passing appropriate parameters to get the accurate output. Later training data was fit into the model using model_name.fit() method and a prediction variable and test data of independent variables passed into it using predict() method from XGBClassifier class.

3.4 Clustering Algorithms

3.4.1 K-means Clustering

Implementation

For an unsupervised machine-learning algorithm, a Target variable was not considered to make predictions of customer segments, simply because that is the nature of unsupervised algorithms. Hence the clustering was purely done based on all remaining variables from the dataset except for the 'Churn' variable. Thus, the column of this variable was also dropped before implementing the model.

For implementing K-means clustering KMeans class was imported from sklearn.cluster library. By calculating the means and inertias of the variables, the dwere was fitted to Kmeans() functions. This calculation was presented in a form of Elbow plots using range of K values from 1-10. To find the most suitable value of K using inertia, multiple iterations of clusters were created, and each iteration also presented inertia for the number of clusters.

```
The innertia for : 2 Clusters is: 7924788547.314394
The innertia for : 3 Clusters is: 3701554560.7043447
The innertia for : 4 Clusters is: 2115724545.8072972
The innertia for : 5 Clusters is: 1328984361.4124138
The innertia for : 6 Clusters is: 915659259.4881808
The innertia for : 7 Clusters is: 671317049.6595721
The innertia for : 8 Clusters is: 523326634.69356215
The innertia for : 9 Clusters is: 414247889.14460444
The innertia for : 10 Clusters is: 329355938.3071262
The innertia for : 11 Clusters is: 270528194.7821008
The innertia for : 12 Clusters is: 230298848.59726033
The innertia for : 13 Clusters is: 196606169.58120832
The innertia for : 14 Clusters is: 168284373.67920047
The innertia for : 15 Clusters is: 149241783.2236945
The innertia for : 16 Clusters is: 133481682.6076357
The innertia for : 17 Clusters is: 118043294.05673686
The innertia for : 18 Clusters is: 105381393.90132917
The innertia for : 19 Clusters is: 95735188.90807599
```

Figure 10 K-means Inertias for clusters

The following diagram shows an Elbow plot of the above-mentioned inertias against to their respective number of clusters:

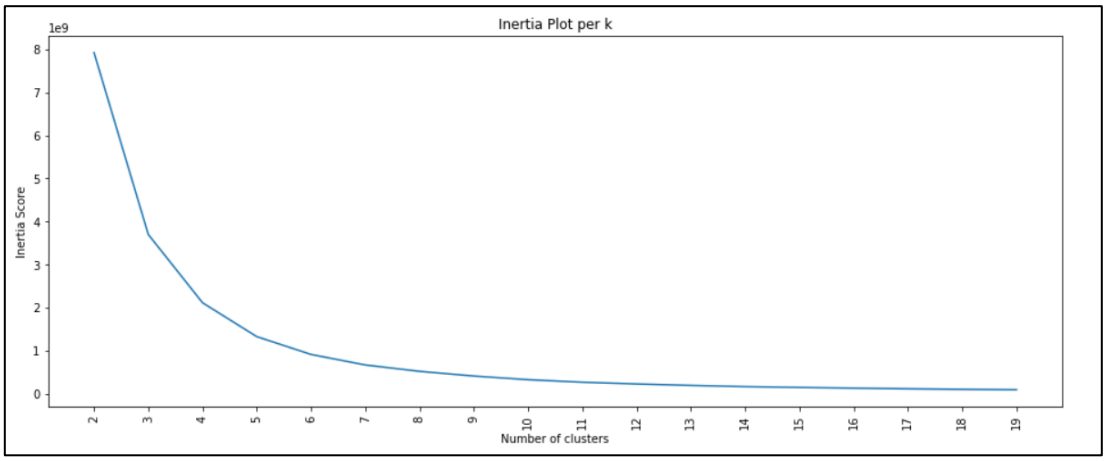


Figure 11 K-means Elbow plot

For precisely deciding the ideal of numbers of clusters to be created, KneeLocator() function from the 'kneed' library was used. The decision was made based on the ideal cost value from the above graph. Cost equals the square of the distance from each location to the nearest cluster centre(Narkhede, S. 2021).

```
Elbow at k  for K-means Clustering= 4 clusters
```

Figure 12 K-means Elow location result

The next phase was to plot a representative graph of identified clusters. Since the dataset has 46 independent variables or components, plotting a 2-dimensional graph of these many variables would not be representative of homogenous clusters. Hence to reduce the dimensionality Principle component analysis(PCA) was used, which transforms the number of possibly correlated variables into smaller number of uncorrelated variables; called principal components. The PCA compresses the data by extracting most important information out of it(Troccoli, E.B. et al. 2022). Result of this technique was able to produce following cluster graph, which represents 4 homogenous clusters of all the customers from the dataset.

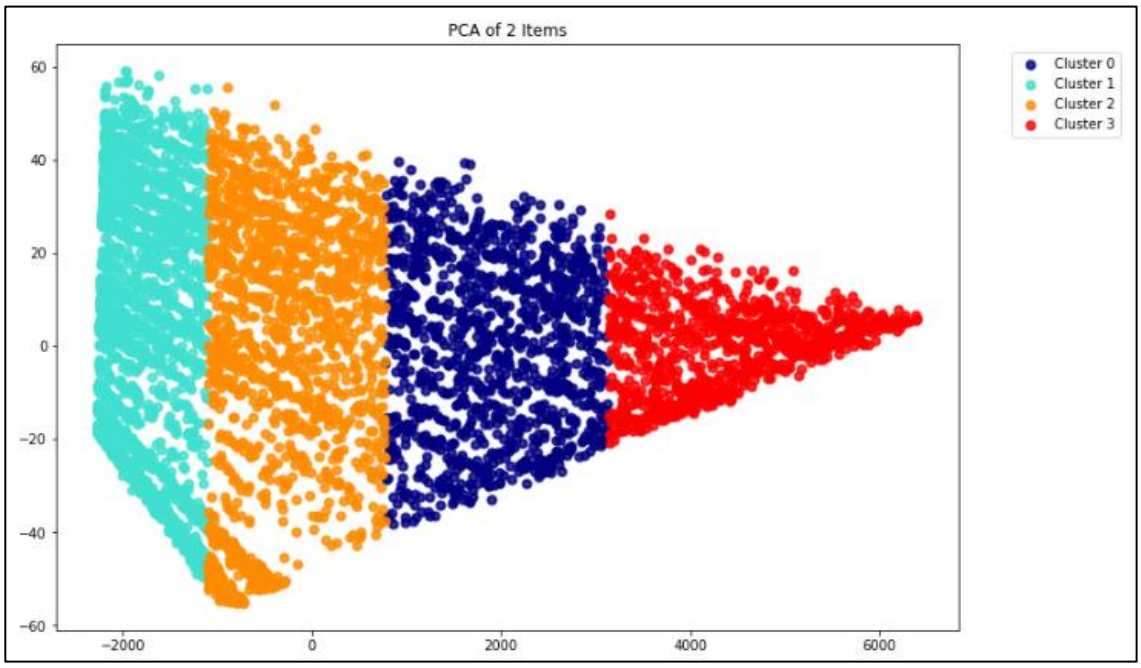


Figure 13 K-means Clusters graph

3.4.2 K-prototype clustering

Implementation

For implementing K-prototype clustering only prerequisite required is pre-processed data without dummy variable transformation. Since numerical data and categorical data was processed separately during implementation of the model itself. After that KPrototypes() function was used from kmodes.kprototypes library. Similar to K-means clustering, being an unsupervised machine-learning algorithm, the Target variable was not included for segmenting the dataset into homogeneous clusters.

In later stage, both numerical data converted into matrix using numpy() function and decoded categorical data was fitted to the K-prototype model using fit_predict() function.

The results of the model were converted into an Elbow method graph, using cost.append() method and iterating number of clusters from 1-10, which is represented as following:

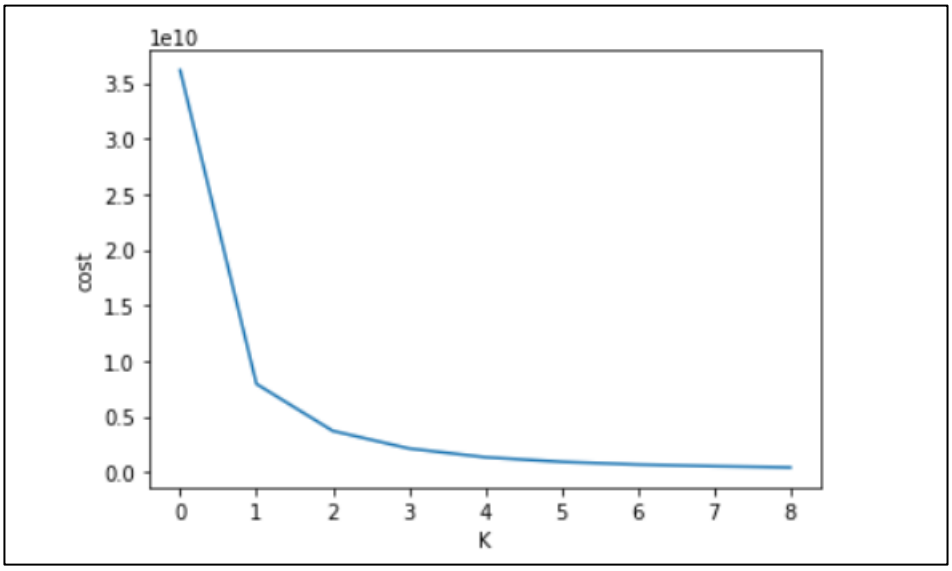


Figure 14 K-prototype Elbow plot

Similar to K-means clustering, to find the exact location of elbow KneeLocator() function from keed library. Which produced following result:

```
Elbow at k for K-prototye Clustering= 2 clusters
```

Figure 15 K-prototype Elbow location

Here also we used PCA to be able to plot a 2 dimensional graph of the identified clusters, which produced following graph:

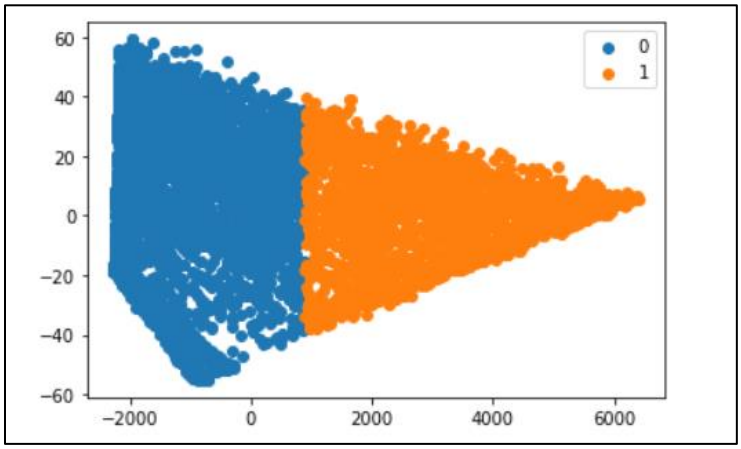


Figure 16 K-prototype Cluster graph

Chapter 4

4.0 Evaluation

In this section, we will discuss about various metrics used to evaluate both Classification algorithms and Clustering algorithms.

4.1 Evaluation metrics for Classification models

Evaluation metrics are a way we would use to evaluate the performance of the classifiers. There are different evaluation metrics used to evaluate the machine learning models. The four main metrics used are Recall, Precision, Accuracy and F1 score. Since our main aim is to predict the type of Churners it would be best to compare the accuracy scores of machine learning models. Our dataset is not a completely balanced dataset. To deal with this situation we will also be considering the results of Confusion matrix with it. A confusion matrix is a combination of predicted and actual values(Narkhede, S. 2021).

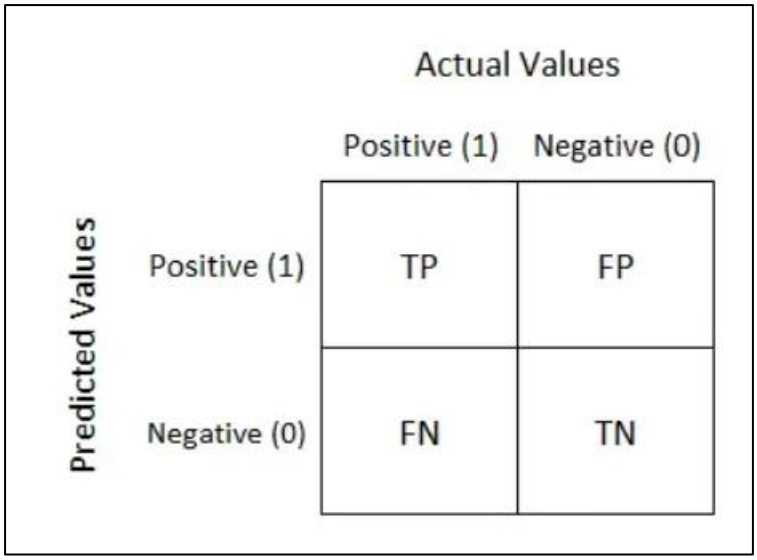


Figure 17 Confusion Matrix

The Confusion Matrix has four parts which are:

1. True Positive: You predicted it as positive and it is true.
2. True Negative: You predicted it as negative and it is true.
3. False Positive: You predicted it as Positive and it is false.
4. False Negative: You predicted it as Negative and it is false.

Accuracy is the most obvious performance metric, it is simply the ratio of correctly predicted over total predicted observations(28)

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

.

Equation 1 Accuracy

Recall represents, how many classes are predicted correctly from all the positive classes(28). The higher the Recall value, the better the predictions will be. And its formula is as following:

$$Recall = \frac{TP}{TP + FN}$$

Equation 2 Recall

Precision represents, how many are actually positive classes from all the those who are predicted as positive(28). Same goes for the Precision score, it should high as possible. Its formula is as follows:

$$Precision = \frac{TP}{TP + FP}$$

Equation 3 Precision

F-score aids in measuring both recall and precision simultaneously. By penalising the extreme values more harshly, it substitutes the harmonic mean for the arithmetic mean(28).

$$F\text{-}measure = \frac{2*Recall*Precision}{Recall + Precision}$$

Equation 4 F-1 Score

Model_name.score() method from the same class was used to compare the result and test the accuracy of the created model. For evaluating the performance of the model using the Classification report metrics, a method called classification_report() from sklearn.metrics library was executed to get results of precision, recall, f1 and support scores. Another metrics to used here was Confusion matrix using confusion_matrix() class from sklearn.metrics library.

Following results were produced post implementing above mentioned metrics:

```
Accuracy Score of Decision tree Model without SMOTE-ENN: 78.3798283261802
6 %
              precision    recall  f1-score   support

           0       0.84      0.88      0.86      1401
           1       0.57      0.50      0.54       463

    accuracy                           0.78      1864
   macro avg       0.71      0.69      0.70      1864
weighted avg       0.78      0.78      0.78      1864
```

Figure 18 Classification report of Decision tree with Imbalanced dataset

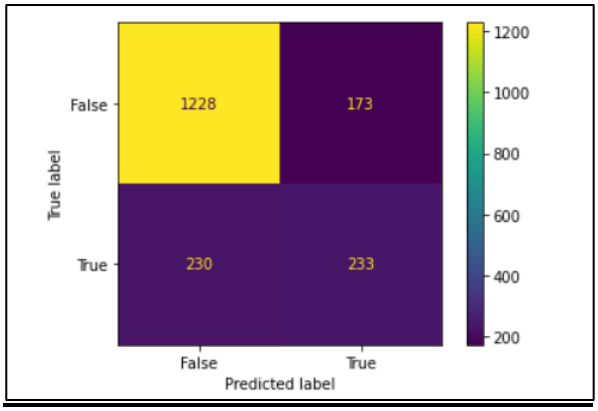


Figure 19 Confusion matrix of Decision tree with Imbalanced dataset

Accuracy Score of Decision Tree with SMOTE-ENN: 93.73801916932908 %

	precision	recall	f1-score	support
0	0.95	0.90	0.92	666
1	0.93	0.97	0.95	899
accuracy			0.94	1565
macro avg	0.94	0.93	0.94	1565
weighted avg	0.94	0.94	0.94	1565

Figure 20 Classification report of Decision tree with SMOTE-ENN

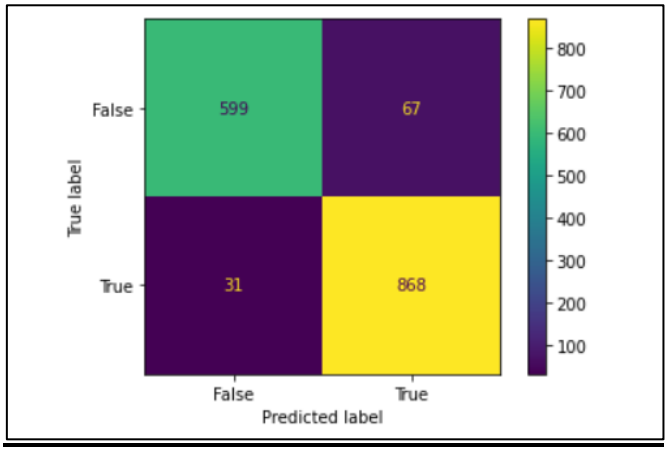


Figure 21 Confusion matrix of Decision tree with SMOTE-ENN

Accuracy Score of Random Forest Model without SMOTE-ENN: 78.75536480686695 %

	precision	recall	f1-score	support
0	0.82	0.91	0.86	1362
1	0.65	0.45	0.54	502
accuracy			0.79	1864
macro avg	0.74	0.68	0.70	1864
weighted avg	0.77	0.79	0.77	1864

Figure 22 Classification report of Random forest tree with Imbalanced

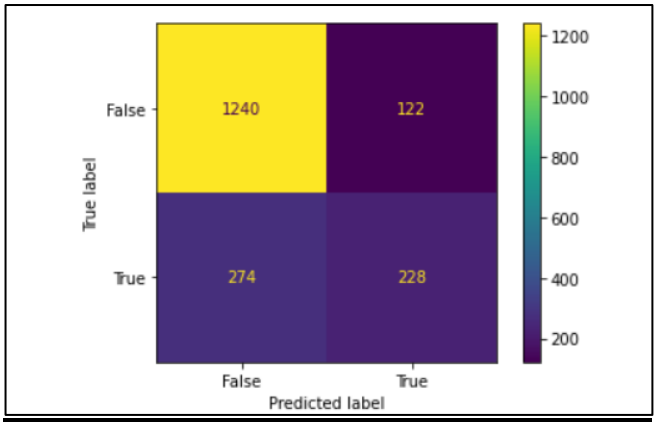


Figure 23 Confusion matrix of Random forest tree with imbalanced dataset

Accuracy Score of Random Forest Model with SMOTE-ENN: 93.34625322997417 %

	precision	recall	f1-score	support
0	0.95	0.90	0.93	711
1	0.92	0.96	0.94	837
accuracy			0.93	1548
macro avg	0.94	0.93	0.93	1548
weighted avg	0.93	0.93	0.93	1548

Figure 24 Confusion matrix of Random forest tree with SMOTE-ENN

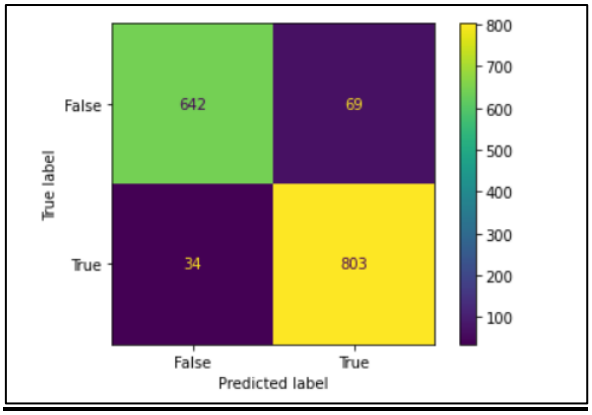


Figure 25 Classification report of Random forest tree with SMOTE-ENN

Accuracy Score of LightGBM without SMOTE-ENN: (78.91630901287554, '%')

	precision	recall	f1-score	support
0	0.83	0.89	0.86	1347
1	0.64	0.54	0.59	517
accuracy			0.79	1864
macro avg	0.74	0.71	0.72	1864
weighted avg	0.78	0.79	0.78	1864

Figure 26 Classification report of LightGBM with Imbalanced dataset

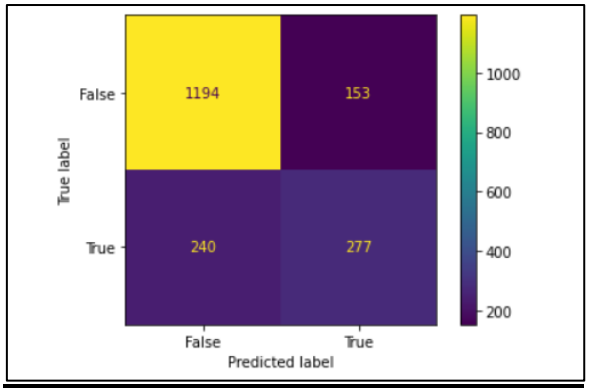


Figure 27 Confusion matrix of LightGBM with Imbalanced dataset

Accuracy Score of LightGBMlgb_sm_score with SMOTE-ENN: 96.53401797175867 %

	precision	recall	f1-score	support
0	0.96	0.96	0.96	714
1	0.97	0.97	0.97	844
accuracy			0.97	1558
macro avg	0.97	0.97	0.97	1558
weighted avg	0.97	0.97	0.97	1558

Figure 28 Classification report of LightGBM with SMOTE-ENN

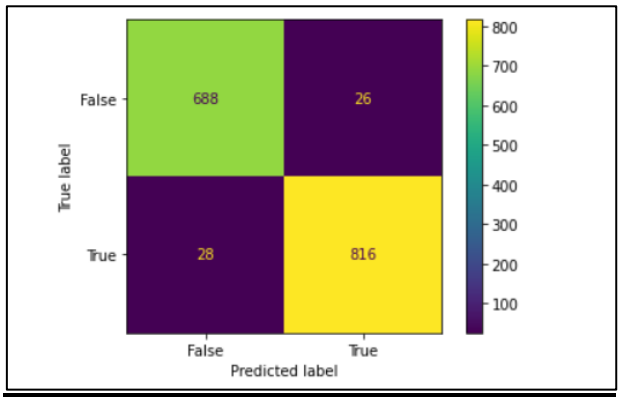


Figure 29 Confusion matrix of LightGBM with SMOTE-ENN

Accuracy Score of XGBoot Model without SMOTE-ENN: 78.54077253218884 %

	precision	recall	f1-score	support
0	0.83	0.88	0.86	1347
1	0.64	0.53	0.58	517
accuracy			0.79	1864
macro avg	0.73	0.71	0.72	1864
weighted avg	0.78	0.79	0.78	1864

Figure 30 Classification report of XGB with Imbalanded dataset

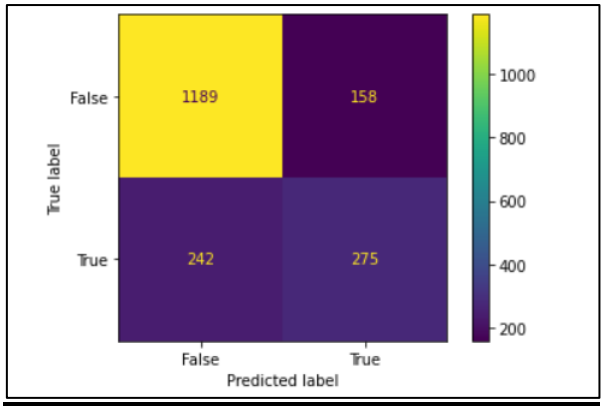


Figure 31 Confusion matrix of XGB with Imbalanced dataset

Accuracy Score of XGBoot with SMOTE-ENN: 96.31849315068493 %

	precision	recall	f1-score	support
0	0.96	0.94	0.95	700
1	0.95	0.97	0.96	839
accuracy			0.95	1539
macro avg	0.95	0.95	0.95	1539
weighted avg	0.95	0.95	0.95	1539

Figure 32 Classification report of XGB with SMOTE-ENN

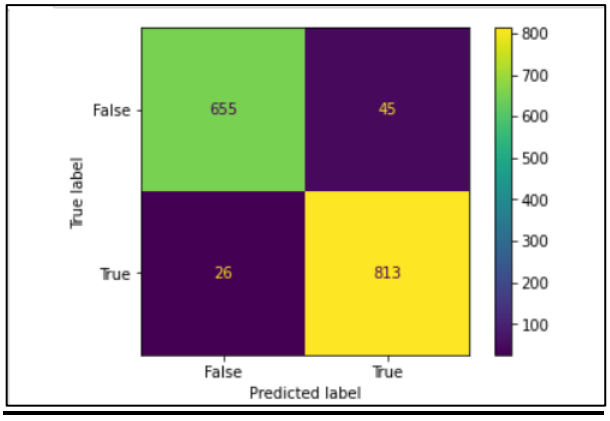


Figure 33 Confusion matrix of XGB with SMOTE-ENN

4.2 Evaluation metrics for clustering models

1. **Silhouette Score**

   The silhouette score is a measure for calculating a clustering algorithm's goodness of fit, but it may also be used to find the ideal value of k. The mean distance between the closest and intra-cluster samples is used to

compute it. Its value is between -1 and 1.When the value is 0, it means that clusters overlap and that the data or the calculation of the value of k is flawed. The optimum value, 1, denotes extremely dense and well-separated clusters. A negative number means that the components were probably assigned to the incorrect clusters. The clusters are more clearly distinguished the closer the silhouette score's value is to 1(Parashar, A. 2022).

In python, silhouette_score() function was used to calculate the Silhouette score from sklearn.metrics library. The function takes the input of predictor variables and the Target variable.

2. **Calinski Harabaz Index**

It is also known as the Variance Ratio Criterion. The ratio of the total of the between-cluster and within-cluster dispersion is known as the Calinski Harabaz Index. The clusters are more easily distinguished the higher the index(Parashar, A. 2022).

In python, calinski_harabasz_score() function was used to calculate the Calinski Harabaz Index from sklearn.metrics library. The function takes input of predictor variables and the Target variable.

3. **Davies Bouldin index**

Another measure for assessing clustering algorithms is the Davies-Bouldin index (DBI), which was developed by David L. Davies and Donald W. Bouldin and published in 1979. The Davies Bouldin index is the average distance between each cluster and its closest neighbour, where similarity is determined by the proportion of within-cluster to between-cluster distances.The DB Index has a minimum value of 0, and a smaller value (near to 0) indicates a better model that generates better clusters(Parashar, A. 2022).

In python, davies_bouldin_score() function was used to calculate the davies_bouldin_score from sklearn.metrics library. The function takes input of predictor variables and Target variable.

The following results were produced after implementing the above-mentioned evaluation metrics on both K-means and K-prototype Clustering:

K-means:

```
Silhouette score for K-mean Clustering: 0.601825244208776
```

```
Calinski Harabasz Score for K-mean Clustering: 37671.351048928256
```

```
Davies Bouldin Score for K-mean Clustering: 0.5168214248404494
```

K-prototype:

```
Silhouette score for K-prototype Clustering: 0.7029853221037297
```

```
Calinski Harabasz Score for K-prototype Clustering: 25029.456703735337
```

```
Davies Bouldin Score for K-prototype Clustering: 0.4411894170330021
```

4.3 Results & comparison

This is one of the most pivotal parts of the research. In this section, we will be analysing the results of all Classification models and Clustering models.

The following table presents the outputs of the classification models and also a comparison of each model to its previous imbalanced version before applying the SMOTE-ENN technique:

	Imbalanced				**SMOTE-ENN**			
	Accuracy	Precision	Recall	F-1	Accuracy	Precision	Recall	F-1
Decision Tree	0.783	0.57	0.50	0.54	0.937	0.93	0.97	0.95
Random Forest tree	0.787	0.65	0.45	0.54	0.933	0.92	0.96	0.94
LightGBM	0.789	0.64	0.54	0.59	0.965	0.97	0.97	0.97
XGB	0.785	0.64	0.53	0.58	0.963	0.95	0.97	0.96

Table 2 Classification Models results and comparison

Since as per the objectives of the research, the models were built to predict the churn probability of customers based on the available dataset, above table represents the same. From the results, it is evident that the accuracy scores of all the models with an imbalanced dataset are around 0.78. Whereas applying SMOTE-ENN technique all the models show noticeable improvement in their accuracy scores. LighGBM has scored highest across all the metrics with an accuracy score of 0.965, which will be deciding score for selecting this model to build a Churn prediction application as a part of retention strategy. Combined effect of SMOTE-ENN and XGB model comes second highest with an accuracy score of 0.963, following this Decision tree scored 0.937 and the last one is Random Forest tree with score of 0.933.

The following table represents evaluations metrics for Clustering algorithms tested namely K-means and K-prototype:

	Silhouette score	**Calinski Harabasz score**	**Davies Bouldin score**	**K-value**
K-means	0.601	37671.35	0.516	4
K-prototype	0.702	25029.45	0.441	2

Table 3 Clustering algorithms results and comparison

From the above table, although the Silhouette score is high for K-prototype (0.702) compared to K-means(0.601), Calinski Harabasz score, and Davies Bouldin score are highest for K-means clustering with values of 37671.35 and 0.516 respectively. Another deciding factor to consider here is K-value. Since the dataset has 7032 instances, to divide customers into simply 2 segments based on the K-value of K-prototype algorithm, would create segments which cannot be identified as homogenous in nature. Whereas K-value 4 for K-means will help in analysing the dataset and look for insights on 4 identified clusters, which will help marketing managers make better retention strategies.

Based on the above observations, the decision was made to build a churn prediction application using LightGBM model and a Customer segmentation application using K-means model.

Chapter 5

5.0 Retention strategy tools

After achieving the first two objectives of the proposed research, this section will produce a working application for predicting both the Churn probability and segment of new customers to help marketing managers of the telecommunication business make strategic decisions on retaining those customers. Mathematical results will be converted into insightful visualizations by analysing data of Clustered dataset.

To build a Churn prediction application, we have used Flask web framework, which integrates Python code as a backend program and HTML documented form to take inputs of customers matching all the variables' respective values from the original dataset except for the Churn variable. The coding for the application was done using Visual Studio code by creating a virtual environment to produce results on a local server. LightGBM model was loaded in the backend program which will help in making churn predictions with an accuracy of 96.5% for every new record of a customer. Another part of the output will provide a probability percentage of that specific customer churning or not churning.

Similarly, for building a customer segmentation application Flask web framework was used combining python programming in the back-end and HTML document for creating input web form. Visual studio code platform combined the effect of the backend python program and the prediction is made on the basis of K-means clustering model built in the earlier phase of the research.

For making an effective retention strategy marketing managers requires easy-to-learn and useful tools. In the era of data-driven businesses, visualisation is one of the most powerful rather essential tools for assessing the performance of a business. Hence this consultancy report delivers the following Interactive visualisation dashboard to empower the telecommunication business in making better retention decisions.

Based on business understanding, the dashboard is designed to provide information about the demographics of customers from homogeneous segments, distribution of services under these segments, payment method used, scatterplot showing monthly charges versus total charges of customers from all segments, distribution of contract types etc.

Another important graph which can help in making sound decision is the correlation graph of Churn variable against all the independent variables from the dataset. To represent a better version of correlation, we used the feature importance method in the XGB model to understand which features or variables are highly related to the churn variable. The higher the value of a feature against the churn variable, the higher the importance needs to be given to improving service related to that feature or design a strategy which will decrease its effect and ultimately help retaining more customers.

Figure 34 Tableau dashboard: Charges, Tenure, Contracts and payments

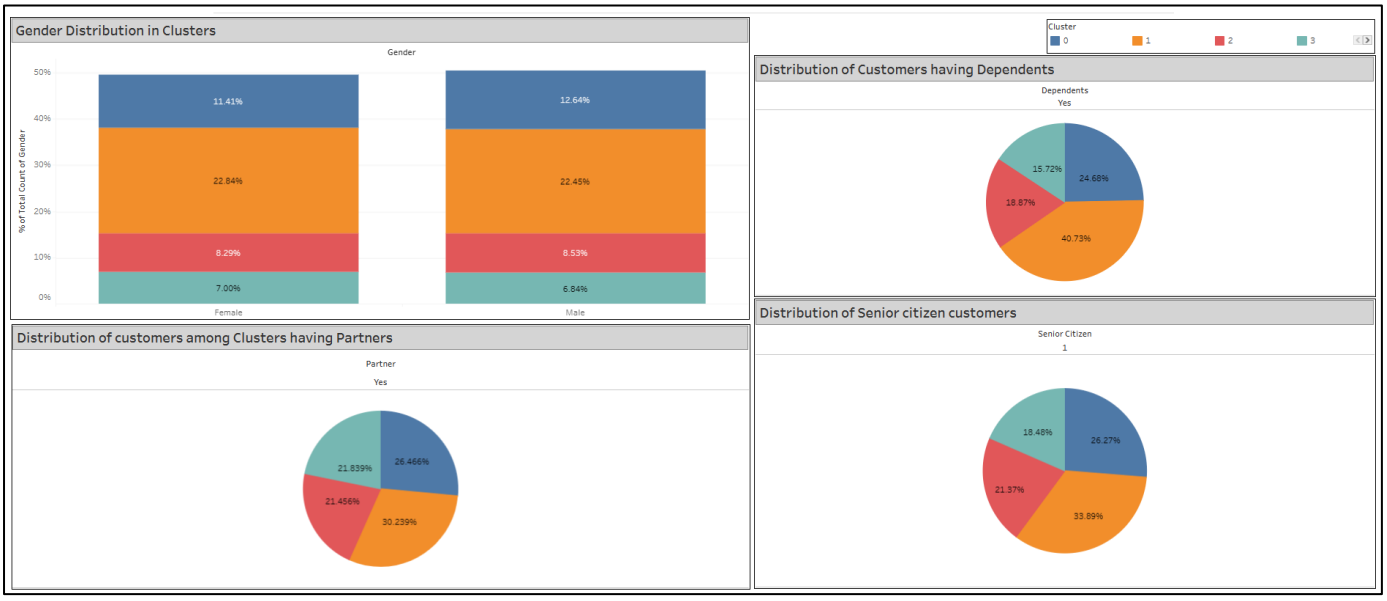


Figure 35 Tableau dashboard Customer Demographics

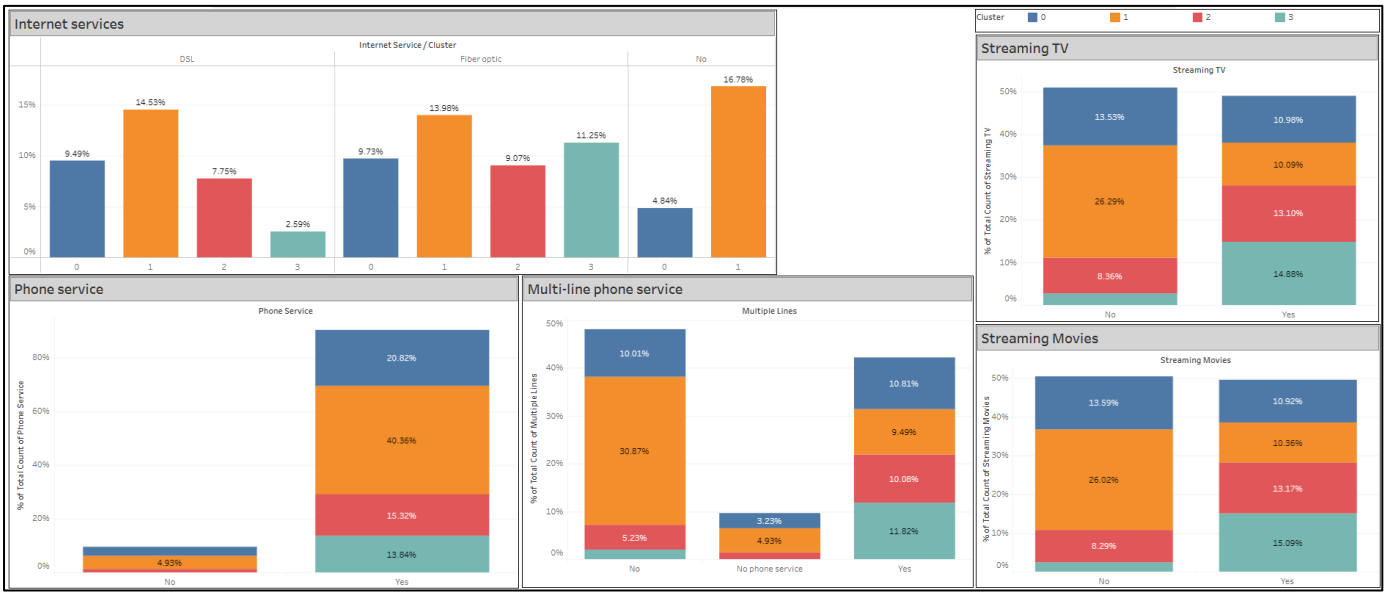


Figure 36 Tableau Dashboard Basic Services

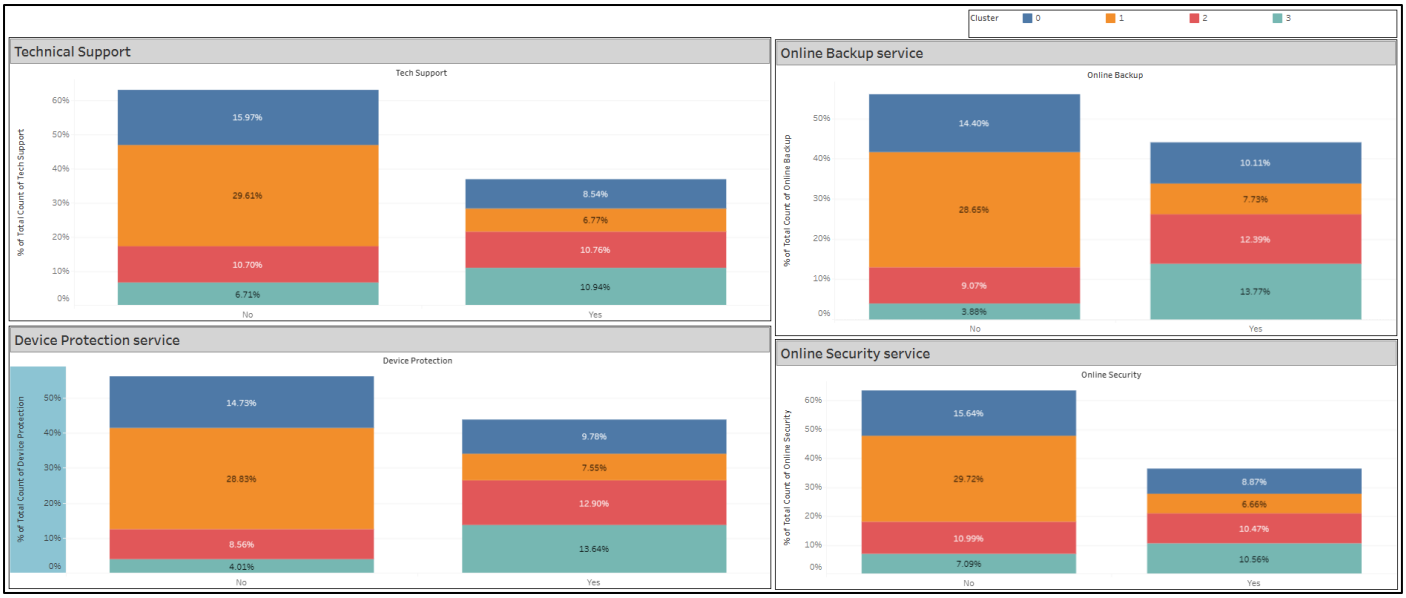


Figure 37 Tableau Dashboard additional services

Advantages of using these tools:

1. The tools can be used iteratively for the customers who are predicted to churn and also for customers who are already churned, to understand the reason for stopping the usage of products/services. This will help determine, how the churn probabilities of customers have changed over the period they are active with the company and also to make improvements in on-going retention strategies of the business.
2. Analysis and performance of ongoing products/services provided by the company. It will help understand which services are most subscribed to by customers from a specific segment. Marketing managers can take advantage of this knowledge for strategizing their next marketing campaign.
3. Clusters will help understand the customer base of the company and how their user preferences are impacting the business's profit. It will help in profit comparisons between clusters, which will help in deciding which type of customers are worth retaining based on the marketing budget of a quarter or an year.

5.2 Segment analysis

Based on Cluster analysis of the dataset, this section will briefly describe information about each identified segment of customers.

Cluster 0: Of 7032 customers of the telecommunication business, 24% of customers are under this cluster, which determines its population compared to other clusters. The average monthly expenditure of customers from this segment is between $20 to just above $100. Whereas the average total expenditure is between $70-$3000. And average tenure value of this segment is 38 months. Most of the customers from this segment prefer using electronic check(7.91%) method, 6.24% uses Bank transfer and rest are using Credit card and least number of customers uses mailed check. 12.8% of the whole population uses, month-to-month contract type and only 11% are on one and year contracts

Female customers are 11.41% and Males are 12.64% of the data. 26.2% of these customers are Senior citizens. 26.4% are with a partner and 24.6% have dependents.

Most of the customers from this segment are using Fiber optic as well as DSL internet service contributing around 19% of the population. Where around 27% of customers are combinedly using Movies and TV services, as compared

to around 22% are not using them. Around 45% of customers from these segments are not using security services like online backup, device protection etc only 19% of are using them, and 4.8% are not using internet services at all.

Based on above analysis, we have named this segment as **'Basic users'.**

Cluster 1: Around 45% of the total population of customers are under this segment, making it the segment with largest population among all identified segments. On the other side, this segment has the lowest average tenure of 12 months. 34% of customers are on month-to-month contract and spending least average monthly charges ranging from $18-111$. Average total expenditures are also lowest in this segment.

Around 40% of customers have dependents and gender percentage is same for both genders around 23%

Around 15% of customers are subscribed to DSL and 14% are using Fiber optic, but 17% have not opted for any internet services. Customers from this segment are responsible for covering 40% proportion of phone services. And around 52% of customers are observed to be not using both TV and movies subscriptions, whereas only around 21% are using them. Contribution of this segment for using security services is only around 7% for each type of security.

Based on above analysis, we have named this segment as **'Short-term subscribers.**

Cluster 2: Around only 17% of all customers belong to this segment, making it the second smallest segment among other. But when it comes to average tenure this segment stands second highest with 52 months. Average monthly and total expenditure also stand second highest here. Contract types of distribution is around same figures between 5%-6% for each contract type and 5.42% customers are using electronic check payment method.

Gender distribution is even for both genders around 9% each. 19% of customers with dependents come under this segment and 22% of them are senior citizens and 21.37% have partners.

Every customer from this segment subscribed either DSL or Fiber optic service. And around 26% combined observed to be using both Tv and movie services. 15% are using phone line, out of that 10% are using multi-line service as well.

Based on above analysis, we have named this segment as **'Standard users'**

Cluster 3: Only around 14% of overall customers come under this segment, making it the least populated of all segments. But it scores highest in average tenure number with 66 months with around 8% of customers subscribed to two year contract, 4% subscribed to one year. And again contributing highest for average monthly and total charges among all other segments.

Gender ratio is close to even in this segment, with 16% have dependents, 18% are senior citizens and around 22% have partners.

The segment stands second highest for using Fiber optic service and all customers using one of the other internet as well as phone services. Out of all customers around 30% from this segment are using Tv and movie services which is highest among all other segments.

Based on above analysis, we have named this segment as **'Loyal users'**

5.3 Retention recommendations

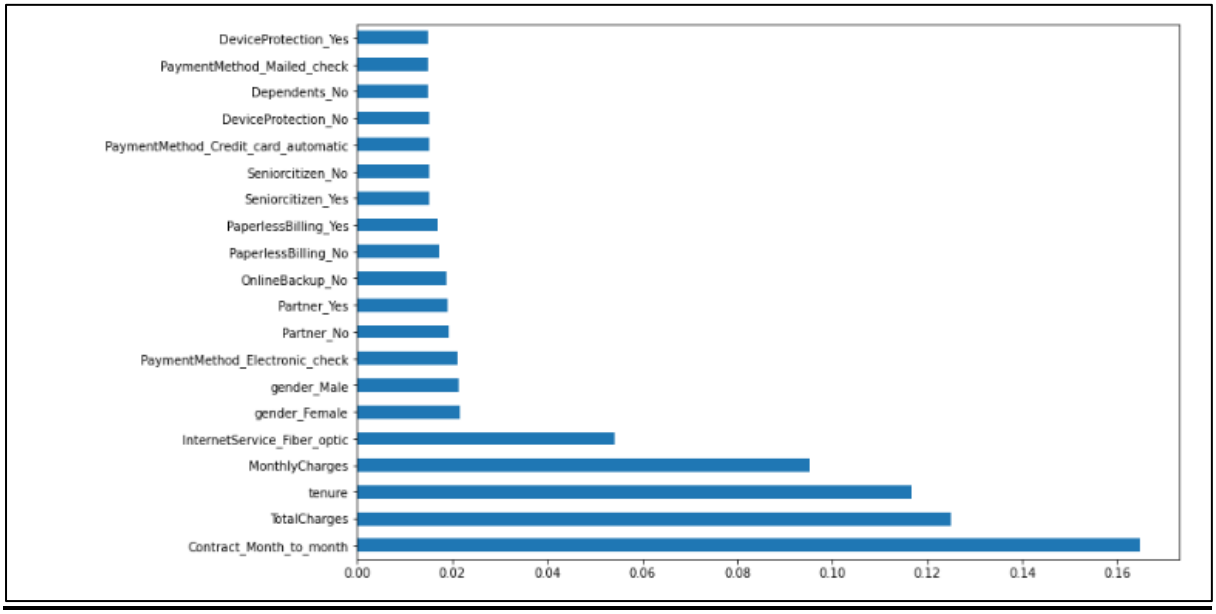


Figure 38 XGB model feature importance graph

From above figure we can see that driving factors for generating most of the churners based on customers who are on month-to-month contract Based on segment analysis and feature importance criteria given above, we have segment-specific retention recommendations as follows:

Recommendations for Basic users:

1. Try to convert these Basic users into Standard users by offering discounted prices for longer monthly cycles.
2. Reached out to customers who are not subscribed to any internet services and pitch them one month free subscription along with free access to TV and movie services as a luxury offer.
3. Make them aware about benefit of using security services for protecting their devices.

Recommendations for Short-term users

1. Reach out to these customers to collect feedback of their experiences of using the services with the business considering this segment has a large population responsible for generating monthly sales targets.
2. Since around 45% of customers from all are under this segment, and their churning could affect the performance of business in the market. Try to convert these customers into one year contract to keep them engaged with the services.
3. Considering the percentage of dependents, business can try to attractive offer users bundle offers to increase keep the customers retained even before they leave.
4. Attract customers with limited time offers on special occasions, festivals and holidays.

Recommendations for Standard users

1. Customers with second highest average tenure should be treated as loyal customers to improve relationship between the business and its customers, since their churn rate is relatively less compare to Basic and short-term users.

2. Since customers from this segment seems to be enjoying internet service along with TV, movies as well as phone services, good feedback from these customers could be used to attract new customers in the market.
3. For customers who are not subscribed to security services, try to make them aware about usefulness of those services, this might increase the trust factor of customers with the business.
4. Offer customers with referral codes to increase number of subscribers by using basic marketing tactic through word of mouth.

Recommendations for loyal users

1. Continue to deliver excellent product and services to loyal customers.
2. Provided personalized services and offers to make them feel special and as a part of the business family.
3. Try saving the cost of the marketing budget on this segment since the customers are predicted to stay longer with the business, hence trying to retain them would be excessive use of retention efforts.

Chapter 6

6.1 Conclusion

This chapter will conclude the research study by addressing how the aims and related objectives were achieved. The discussion will also include a review of the project's research cycle, to give a brief of the challenges faced, what went according to the project plan, the usability of concepts in the real world and their ethical implications on businesses.

This study aimed to write a consultancy report for a telecommunication business, whose churn dataset of 7043 customers was used to test a few well-known and industry-standard churn prediction models. Based on the results of these tests, the model with the highest precision metrics (Accuracy = 96.5%) named LightGBM was selected to build a churn prediction application and the K-means model to design an interactive visual dashboard representing customer segments. Combining outcomes was used to make retention recommendations. The challenge set for the project was to build useful tools which will help in retaining customers and designing a retention strategy. At the beginning of the research, the main objective was to test churn prediction models on a specific dataset, but considering the popularity of this topic there was enough existing research even on the chosen dataset. After reviewing more literature, it was found that a combination of supervised and unsupervised machine learning was not used together to present a consultancy report based on a business's data. Hence the framework of the problem evolved and hence the goals as well.

CRISP-DM was used to deliver a data mining project, which helped in delivering objectives in a structured approach. During the model selection phase, alongside the K-means algorithm, K-prototype clustering was decided to be tested in the later stage for comparing the performance of at least two clustering algorithms.

The study ensured to comply to ethical approaches by evaluating the results of all algorithms using relevant evaluation metrics.

Clustering analysis helped in making retention recommendations based on an available dataset of the telecommunication business.

6.2 Limitations and future scope

Despite its innovative and distinctive contribution, this study raised difficulties that need to be addressed in follow-up studies. In this section, we will discuss opportunities for future researchers.

1. Considering the timeline, the prediction models were built to predict churning and segment of a single customer at a time, but future researchers can try to build a dynamic application for records of multiple customers by taking advantage of SQL databases.
2. The retention tools and strategy recommendations are based on a single dataset. But an impactful retention strategy can be designed with complete knowledge of existing business strategy, market demand and competitor analysis.
3. Since the study is conducted on a dataset of a telecommunication business, by using models tested in this study, future studies can implement them on other datasets to test the generalizability of the findings of this study.
4. The applications can only be tested manually using the unit test method, considering the knowledge gap of the research automated tests could be performed to test further limitations of the applications.

References

Ahmad, A.K., Jafar, A. and Aljoumaa, K. (2019) "Customer churn prediction in telecom using machine learning in Big Data Platform," Journal of Big Data, 6(1). Available at: https://doi.org/10.1186/s40537-019-0191-6.

Baheti, P. (2023) *Train test validation split: How to & best practices [2023], V7*. Available at: https://www.v7labs.com/blog/train-validation-test-set#:~:text=In%20general%2C%20putting%2080%25%20of,dimension%20of%20the%20data%2C%20etc. (Accessed: January 8, 2023).

Batista, G.E., Prati, R.C. and Monard, M.C. (2004) "A study of the behaviour of several methods for balancing machine learning training data," ACM SIGKDD Explorations Newsletter, 6(1), pp. 20–29. Available at: https://doi.org/10.1145/1007730.1007735.

Bernazzani, S. (2022) *22 examples of customer retention strategies that work, HubSpot Blog*. HubSpot. Available at: https://blog.hubspot.com/service/customer-retention-strategies (Accessed: January 8, 2023).

Chawla et al., 2002 https://www.jair.org/index.php/jair/article/view/10302/24590, View of smote: Synthetic minority over-sampling technique. Available at: https://www.jair.org/index.php/jair/article/view/10302/24590 (Accessed: January 8, 2023).

Chauhan, N.S. (2022) kdnuggets.com. Available at: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html.

Community, I.B.M. (2017) "IBM Business Analytics Community." Available at: https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113 (Accessed: 2022).

Coussement, K., Lessmann, S. and 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, pp. 27–36. Available at: https://doi.org/10.1016/j.dss.2016.11.007.

Das, C. *et al.* (2021) "A new hybrid feature selection-classification method to identify churned customers," *Lecture Notes in Electrical Engineering*, pp. 193–204. Available at: https://doi.org/10.1007/978-981-16-0275-7_16.

Dhaliwal, S., Nahid, A.-A. and Abbas, R. (2018) "Effective intrusion detection system using XGBoost," Information, 9(7), p. 149. Available at: https://doi.org/10.3390/info9070149.

Dorogush, A.V., Ershov, V. and Gulin, A. (2018) CatBoost: Gradient boosting with categorical features support, arXiv.org. Available at: https://arxiv.org/abs/1810.11363 (Accessed: December 16, 2022).

Eria, K. and Marikannan, B.P. (2018) Systematic review of customer churn prediction in the telecom sector, https://jati.sites.apiit.edu.my/files/2018/07/2018_Issue1_Paper2.pdf. Available at: https://jati.sites.apiit.edu.my/files/2018/07/2018_Issue1_Paper2.pdf (Accessed: December 16, 2022).

Fraihat, M. *et al.* (2022) "An efficient enhanced K-means clustering algorithm for best offer prediction in telecom," *International Journal of Electrical and Computer Engineering (IJECE)*, 12(3), p. 2931. Available at: https://doi.org/10.11591/ijece.v12i3.pp2931-2943.

Galarnyk, M. (2022) *Understanding train test split, Built In*. Available at: https://builtin.com/data-science/train-test-split (Accessed: January 8, 2023).

Hashmi, N., Butt, N.A. and Iqbal, M. (2013) Customer churn prediction in Telecommunication, researchgate.net. Available at: https://www.researchgate.net/profile/Nabgha-Hashmi/publication/257920014_Customer_Churn_Prediction_in_Telecommunication_A_Decade_Review_and_Classification/links/00b495261475ba6758000000/Customer-Churn-Prediction-in-Telecommunication-A-Decade-Review-and-Classification.pdf?origin=publication_detail (Accessed: January 9, 2023).

IBM (2021) The data mining life cycle, IBM. Available at: https://www.ibm.com/docs/it/spss-modeler/saas?topic=dm-crisp-help-overview.

Insani, R. and Soemitro, H.L. (2016) *Business intelligence for profiling of telecommunication customers - APIAR*, *apiar*. Available at: https://apiar.org.au/wp-content/uploads/2016/05/APCAR_BRR7120_ICT-289-298.pdf (Accessed: January 3, 2023).

Jagannath, V. (2017) Diagram of a random decision forest, en.wikipedia.org. Available at: https://en.wikipedia.org/wiki/Random_forest#/media/File:Random_forest_diagram_complete.png.

Jia, Z. and Song, L. (2020) "Weighted K-prototypes clustering algorithm based on the hybrid dissimilarity coefficient," *Mathematical Problems in Engineering*, 2020, pp. 1–13. Available at: https://doi.org/10.1155/2020/5143797.

Kimura, T. et al. (2022) Customer churn prediction with hybrid resampling and Ensemble Learning. Available at: https://www.researchgate.net/publication/360287935_Customer_Churn_Prediction_with_Hybrid_Resampling_and_Ensemble_Learning (Accessed: December 16, 2022).

Krishna, et al. (2020) *XGBoost: What it is, and when to use it*, *KDnuggets*. Available at: https://www.kdnuggets.com/2020/12/xgboost-what-when.html (Accessed: January 8, 2023).

Lalwani, P. et al. (2021) "Customer churn prediction system: A machine learning approach," Computing, 104(2), pp. 271–294. Available at: https://doi.org/10.1007/s00607-021-00908-y

Mbaabu, O. (2020) *Introduction to random forest in machine learning*, *Section*. Available at: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/#:~:text=Advantages%20of%20random%20forest,over%20the%20decision%20tree%20algorithm. (Accessed: January 8, 2023).

Muntasir Nishat, M. (2021) SMOTE-ENN algorithm, hindawi. Available at: https://www.hindawi.com/journals/sp/2022/3649406/fig5/ (Accessed: March 2022).

Narkhede, S. (2021) *Understanding confusion matrix*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 (Accessed: January 8, 2023).

Nguyen et al., 2022 *Customer segmentation: A step by step guide for growth*, *OpenView*. Available at: https://openviewpartners.com/blog/customer-segmentation/#What-IsCustomer-Segmentation (Accessed: January 8, 2023).

Nyakara, A. (2021) *Dummy variables in machine learning.*, *Medium*. Medium. Available at: https://abbynyakara.medium.com/dummy-variables-in-machine-learning-b3991367bd59 (Accessed: January 8, 2023).

Pamina, J. et al. (2019) An effective classifier for predicting churn in Telecommunication, SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3399937 (Accessed: December 16, 2022).

Parashar, A. (2022) *How to evaluate clustering based models in python?*, *Medium*. Heartbeat. Available at: https://heartbeat.comet.ml/how-to-evaluate-clustering-based-models-in-python-503343816db2 (Accessed: January 8, 2023).

Ramachandran, S. et al. (2011) "Redd: Redundancy eliminated data dissemination in cluster based mobile sinks," 2011 International Conference on Recent Trends in Information Technology (ICRTIT) [Preprint]. Available at: https://doi.org/10.1109/icrtit.2011.5972478.

Saltz, J.S. and Hotz, N. (2020) "Identifying the most common frameworks data science teams use to structure and coordinate their projects," 2020 IEEE International Conference on Big Data (Big Data) [Preprint]. Available at: https://doi.org/10.1109/bigdata50022.2020.9377813.

Salunkhe, U.R. and Mali, S.N. (2018) "A hybrid approach for class imbalance problem in customer churn prediction: A novel extension to under-sampling," International Journal of Intelligent Systems and Applications, 10(5), pp. 71–81. Available at: https://doi.org/10.5815/ijisa.2018.05.08.

Santini, M. (2016) Advantages Disadvantages of k-Means and Hierarchical clustering (Unsupervised Learning), santini.se. Available at: http://santini.se/teaching/ml/2016/Lect_10/10c_UnsupervisedMethods.pdf (Accessed: January 8, 2023).

Seldon (2021) *Decision trees in machine learning explained*, *Seldon*. Available at: https://www.seldon.io/decision-trees-in-machine-learning#:~:text=The%20main%20benefits%20of%20using,structure%20is%20often%20a%20necessity. (Accessed: January 8, 2023).

Sharma, T. *et al.* (1970) *Customer churn prediction in telecommunications using gradient boosted trees*, *SpringerLink*. Springer Singapore. Available at: https://link.springer.com/chapter/10.1007/978-981-15-0324-5_20 (Accessed: January 8, 2023).

Singh, M. et al. (2018) "Comparison of learning techniques for prediction of customer churn in Telecommunication," 2018 28th International Telecommunication Networks and Applications Conference (ITNAC) [Preprint]. Available at: https://doi.org/10.1109/atnac.2018.8615326

Soria, J., Chen, Y. and Stathopoulos, A. (2020) "K-prototypes segmentation analysis on large-scale ridesourcing trip data," Transportation Research Record: Journal of the Transportation Research Board, 2674(9), pp. 383–394. Available at: https://doi.org/10.1177/0361198120929338.

SURANA, S.U.B.H.A.M. (2020) *What is light GBM? advantages & disadvantages? Light GBM vs XGBoost?: Data Science and Machine Learning*, *Kaggle*. Available at: https://www.kaggle.com/general/264327 (Accessed: January 8, 2023).

Tang, C., Luktarhan, N. and Zhao, Y. (2020) "An efficient intrusion detection method based on LIGHTGBM and autoencoder," Symmetry, 12(9), p. 1458. Available at: https://doi.org/10.3390/sym12091458.

Team, T.E. (2021) *4 simple & powerful customer retention strategies*, *Delighted*. Available at: https://delighted.com/blog/improving-customer-retention-strategies (Accessed: January 8, 2023).

Tripathi, S., Bhardwaj, A. and E, P. (2018) "Approaches to clustering in customer segmentation," *International Journal of Engineering & Technology*, 7(3.12), p. 802. Available at: https://doi.org/10.14419/ijet.v7i3.12.16505.

Troccoli, E.B. *et al.* (2022) "K-means clustering using principal component analysis to Automate label organization in multi-attribute seismic facies analysis," *Journal of Applied Geophysics*, 198, p. 104555. Available at: https://doi.org/10.1016/j.jappgeo.2022.104555.

Santini, M. (2016) *Advantages Disadvantages of k-Means and Hierarchical clustering (Unsupervised Learning)*, *santini.se*. Available at: http://santini.se/teaching/ml/2016/Lect_10/10c_UnsupervisedMethods.pdf (Accessed: January 8, 2023).

Singh, S. (2021) Diagrammatic representation of Leaf-Wise Tree Growth, geeksforgeeks. Available at: https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/.

Soria, J., Chen, Y. and Stathopoulos, A. (2020) "K-prototypes segmentation analysis on large-scale ridesourcing trip data," *Transportation Research Record: Journal of the Transportation Research Board*, 2674(9), pp. 383–394. Available at: https://doi.org/10.1177/0361198120929338.

Tyagi, N. (2020) *Understanding the gini index and information gain in decision trees*, *Medium*. Analytics Steps. Available at: https://medium.com/analytics-steps/understanding-the-gini-index-and-information-gain-in-decision-trees-ab4720518ba8 (Accessed: January 8, 2023).

V, R.J. (2022) *Optimal ratio for data splitting - Joseph - Wiley Online Library*, *onlinelibrary.wiley*. Available at: https://onlinelibrary.wiley.com/doi/full/10.1002/sam.11583 (Accessed: January 8, 2023).

V. Chawla, N. (2002) *https://www.jair.org/index.php/jair/article/view/10302/24590*, *View of smote: Synthetic minority over-sampling technique*. Available at: https://www.jair.org/index.php/jair/article/view/10302/24590 (Accessed: January 8, 2023).

Viadinugroho, R.A.A. (2021) *Imbalanced classification in Python: Smote-tomek links method*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/imbalanced-classification-in-python-smote-tomek-links-method-6e48dfe69bbc (Accessed: January 8, 2023).

Zhang, T., Moro, S. and Ramos, R.F. (2022) "A data-driven approach to improve customer churn prediction based on Telecom Customer Segmentation," Future Internet, 14(3), p. 94. Available at: https://doi.org/10.3390/fi14030094.

Zhang, Y. and Chen, L. (2021) "A study on forecasting the default risk of bond based on xgboost algorithm and over-sampling method," Theoretical Economics Letters, 11(02), pp. 258–267. Available at: https://doi.org/10.4236/tel.2021.112019.

Appendix

Introduction

With an ever-growing and intensely competitive market, every business in the world is always trying to keep their existing customers engrossed in their current products or services with the respective business. In such situations, a business can stay ahead of its competitors only if they know which consumers are losing interest in its product/services since it is always beneficial for businesses to put efforts into retaining existing customers instead of chasing after new ones. According to an article from Harvard Business School, increasing the customer retention rate by 5% could increase a company's overall profits by 25-95%. As a master' student I am curious to learn about what exactly is involved in the world of businesses when it comes to retaining their customers. Hence, I proposed to write a research study about a well-known business concept, which is customer retention using data-driven approaches.

Background

The decision to chase after this business domain comes from my personal experience of working as a small part of a well-established company and the inspiration I gained after learning new analytical concepts during my master's journey so far. When I started working as a technical support agent for a well-known web hosting company in India in the year 2017, I was responsible for assisting the company's customers with technical doubts and/or troubleshooting issues of products/services provided by the business. After showing good work progress in technical assistance, expert-level product knowledge, and persuasive customer interactions, I was put into a team called the 'Retention team', and the role I played is quite self-explanatory. But it was rather fascinating to understand the impact it had on the business of the company. Fast-forwarding to the entire journey of the year 2022, I got an opportunity to learn various concepts of business analytics, to name a few which got my attention are 'Statistics and Mathematics for Business Analytics', 'Analytics and Visualization for Managers and Consultants' and 'Marketing Analytics'. These modules have enlightened me about the power of data analytics and in which business situations can we use it to sustain a competitive space in the market and thrive in the business using data of any kind.

Aim and objectives

-Research on globally recognized analytical techniques for predicting customer churn

-Testing these techniques on a publicly available dataset of a business and deciding which technique is most suitable for building a prediction model

- Design a retention strategy using marketing analytics for the customers who are predicted to be churned

Literature Review

If Customer relationship management(CRM) is one of the support pillars for a stable and competitive business, then a good customer retention strategy is the main ingredient required to build this pillar, which helps the business to stand tall against its fierce rivals in the market (Kimura, 2022). For telecommunications businesses, client churn is a big issue since it lowers profit. This is especially important given that telecommunications businesses compete in a crowded global market where it is getting harder to keep consumers. Even though many businesses spend a lot of

money on marketing to attract new customers, keeping an existing client is typically less expensive than winning a new one. Due to these factors, preventing client turnover has become a top priority for telecom businesses (Zhang et al., 2022). In the long term, retaining customer loyalty and the business' income depends on finding and proposing the best offer that precisely meets the client's demands (Fraihat et al., 2022).

In order to anticipate customer turnover, researchers have used supervised machine learning algorithms(Singh et al.,2018), treating the issue as a binary classification problem(Coussement et al.,2017). The most often utilised algorithms in the earlier research were decision tree, K-Nearest Neighbor, and logistics regression(Hashami et al., 2013). Advanced ensemble learning models(Liang et a., 2019), such as Extreme Gradient Boosting(XGBoost)(Dhaliwal et al., 2018), Light Gradient Boosted Machine (LightGBM)(Tang et al., 2020), and Category Boosting (CatBoost)(Dorogush et al., 2018), have demonstrated good prediction performance in classification issues in recent studies. The datasets utilised in the customer churn forecast are frequently unbalanced; they contain a disproportionate number of non-churn cases compared to churn cases(Ahmad et at., 2019; Eric & Marikannan, 2018). To balance the data, prior research largely used the Synthetic Minority Oversampling Technique (SMOTE)(Chawla et al., 2022)(Zhang; Chen et al., 2021). Although SMOTE can reduce the overfitting issue that arises from random sampling, it can also produce overfitted models when instances of the majority class invade the minority class space or when the minority class is oversampled and invades the majority class space. Recently, researchers have suggested unique and efficient resampling techniques called hybrid resampling, including Synthetic Minority Oversampling Technique- Edited Nearest Neighbour (SMOTE-ENN) and SMOTE Tomek-Links(Salunkhe & Mali, 2018)(Batista et al., 2004). K-mean clustering and segmentation can be used to divide customers into homogeneous clusters based on their user preferences and demographics. Segmenting customers into clusters can help design a marketing plan about what campaign or which offer to run(Fraihat et al., 2022). Prior to segmenting customers, profiling is used, in which groups of customers are separated according to established business principles. The k-means clustering technique is then used to apply consumer segmentation to the resulting profiles(Fraihat et al., 2022).

The above studies are all focused on identifying which customers might churn out of a business and presenting their findings using evaluation matrices.

Research Problem

We gained comparative information about numerous techniques to develop churn prediction models for any B2C type (Business to Consumer) business out there and identified how consumers may be segmented into groups to design a marketing plan for building a sustainable market position from the aforementioned research. However, none of the studies has used a combination of churn prediction models with cluster analysis for designing a suitable retention strategy as a goal of their research. Therefore, this study will treat this gap as an opportunity to explore an executable retention method for a telecommunication company. **The study will design a sequential strategy by using a combination of recommended classification models and will consult the business with suitable marketing tactics depending on the outputs of the selected churn prediction model. Since the study is dependent on a single dataset of a telecommunication company, the outcomes will only be relevant and practical for the company and not necessarily apply to all companies in the telecommunication industry.**

Research Methodology

Dataset

To build a prediction model for a business, gathering suitable quantitative data was a crucial step of the research. Since customer churn prediction and customer segmentation can be derived based on a historical customer dataset of a business, which is not available to access for the public. A secondary data collection method was used to attain a suitable dataset which was publicly available on the platform www.kaggle.com

The IBM dataset, which is an open-source customer attrition dataset in the telecommunications industry, was used in this study. It was first shared in the IBM community and is now available on the Kaggle website (https://www.kaggle.com/datasets/blastchar/telco-customer-churn). Lalwani et al.(2021) and Pamina et al.(2022), Takuma Kimura et al.,(2022)are three recent research that employed the IBM dataset (2019). There are 7043 instances (customers) and 21 variables in the raw data. The dataset contains information on each customer's demographics, internet connection environment and related support, contract terms, billing and payment methods, and the amount charged.

Data Analysis

For data preparation and analysis, Python programming was used on the Jupyternotebook platform. The variable 'Churn' is binary (Yes or No), which was used to find that there are 1,869 churners and 5,174 non-churners in it. Since the percentage of churners is 26.53% and non-churners are 73.46% the dataset might be considered uneven. After performing an Exploratory data analysis, it was found that there are no null values for any variables in the dataset. As a result, we don't need to take any steps to deal with missing values. The dataset does not have any major outliers which might cause any issues in the model-building stage. 'SeniorCitizen' and 'tenure' are integer variables, whereas 'MonthlyCharges' is a float variable. The remaining 18 variables are object-type data. The ‘TotalCharges’ variable was also present as object-type data, which was converted to float type. ‘CustomerID’ variable which is a unique identification number assigned to each customer was removed since it will not be useful for analytical processes.

For checking the distribution Univariate analysis was performed for the ‘Churn’ variable against variables which represented user preferences like PhoneService, MultipleLines , InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod and variables which represented user demographics as gender, SeniorCitizen, Partner, Dependents etc.

For similar variables, Bivariate analysis was also performed to look for more insights relating to the Churn rate of the customers from the dataset. ‘tenure’ variable represented the number of months the customer was active with the company, had the highest observed variance, which was not suitable for comparing with Churn, hence it was divided into groups for getting better insights out of it. Distribution of Churn was also checked against ‘MonthlyCharges’ and ‘TotalCharges’

Descriptive statistics like, mean, median, mode, standard deviation and variance were checked for continuous variables to get more insights into the dataset.

A correlation function from python was used to check the relationship of all the variables from the dataset against the ‘Churn’ variable to identify which variables have the strongest relationship with it, which will help build a prediction model.

The dataset has 16 categorical variables, which were converted to dummy variables for further regression analysis. We can use dummy coding to convert categories into something that a regression can treat as having a high (1) and low (0) score. Any binary variable has directionality because if it is higher, it is category 1, and if it is lower, it is category 0. Instead of expecting each unit to correlate with some form of increment, this allows the regression to look at directionality by comparing two sides (Moran, 2021).

Research Design

Based on the literature review and data analysis, for binary classification of the Churn variable the study will be using the following machine learning algorithms for developing a preliminary model for churn prediction:

1. Decision Tree
2. Random forest tree
3. Gradient Boosted Model(GBM)
4. Extreme Gradient Boosted model(XGBoost)

For dealing with imbalanced Churn variables of Churners and non-churners, where churners are a minority class, the Synthetic Minority Oversampling Technique- Edited Nearest Neighbour (SMOTEENN) is a suitable ensemble method to be used for balancing the dataset to obtain higher accuracy for predicting churn.

The accuracy of each model will be tested with its associated report using its functions from python programming. After balancing the dataset and testing the accuracy of each model, the model with the highest score will be used to build the final prediction model.

The processed data will also be used to perform K-mean Cluster analysis for creating homogeneous segments of customers and based on input values used in the model and output of the Churn prediction model, which will be Churn possibility percentage will be used to decide which segment the customer belongs and a relative marketing strategy will be implemented to retain him/her.

The following diagram summarizes a possible approach to derive a retention strategy for trying to retain customers who are identified to be churned:

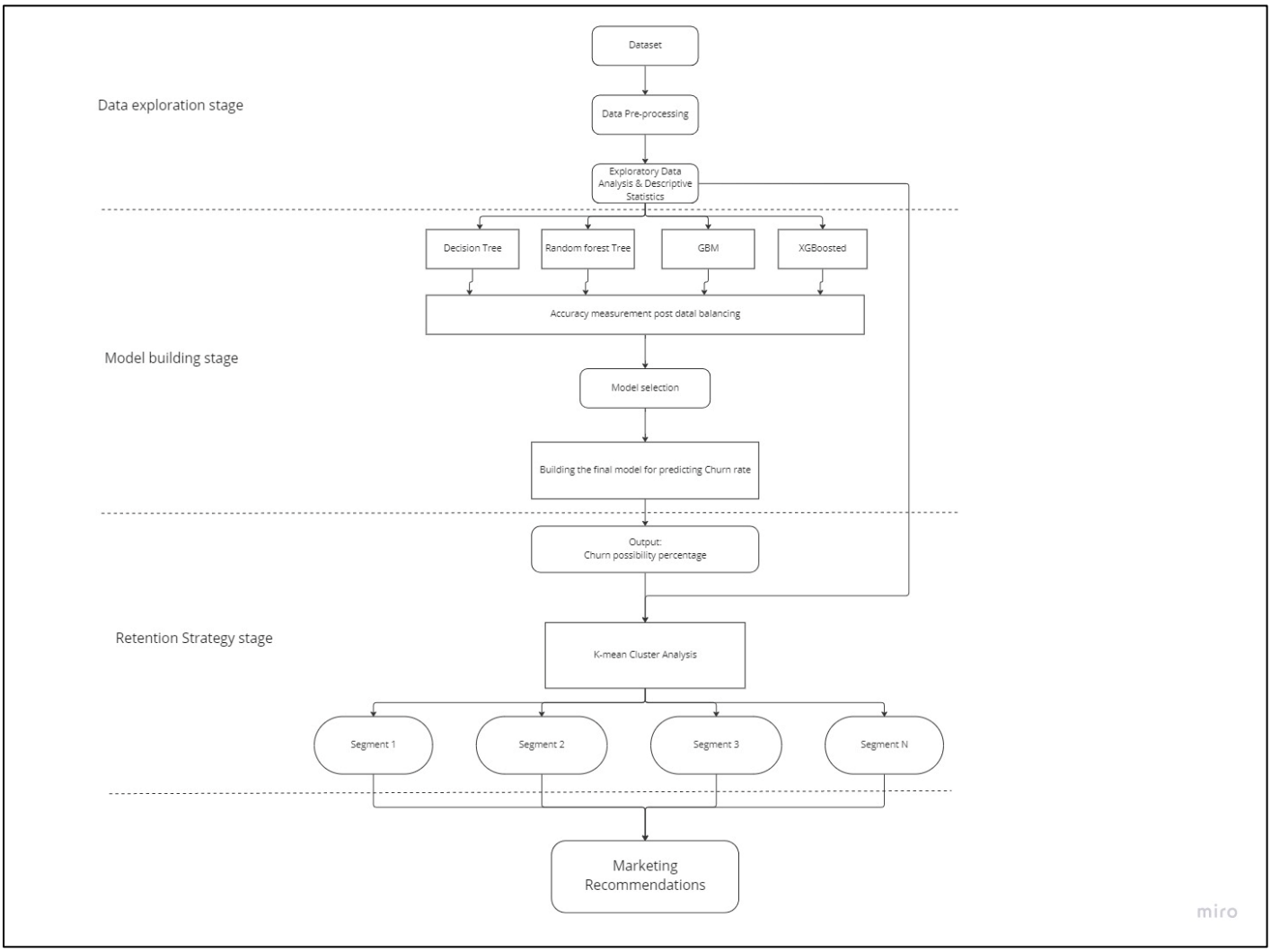


Figure 39: Flowchart of the research plan

Limitation

The outcomes of the prediction models and retention strategy can only be relevant to the company of the dataset which is under consideration since the strategies are designed based on variables from its dataset.

Although with in-depth knowledge, experience and research effective marketing strategies could be implemented, considering the time constraint and limited resources like the company's ongoing strategy, subscription styles

discounts and offers etc, the current study is solely focused on designing a retention strategy, but insufficient to measure its success rate due to absence stakeholders.

Ethical Considerations

1. Because the original dataset source anonymized the personal details of consumers such as their name and address details, accountability for not releasing any sensitive information through the report was assured.
2. The data was gathered from an IBM community site, where datasets are freely available for business and/or non-commercial usage.
3. While working on and analysing consumer demographics, a fair study of the customers' gender, age, and geography is done throughout the report.
4. During the data modification procedure, it was made sure that there is no departure in the information from the original datasets, and the operations were only done for computational flexibility.