PREDICTING CUSTOMER CHURN IN TELECOM COMPANIES THROUGH A MACHINE-LEARNING APPROACH

BY

HLAYISANI RESULT KHOZA

DISSERTATION

Submitted in fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

E-SCIENCE

in the

FACULTY OF SCIENCE AND AGRICULTURE
(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

SUPERVISOR: Dr. TB Darikwa

18 June 2024

Declaration

I, **Hlayisani Result Khoza**, hereby certify that this dissertation is submitted to the University of Limpopo in partial fulfillment of the requirements for the degree of Master of Science(E-Science) and is an original work of mine, that I have properly acknowledged all of the material in it, and that I have not submitted it for credit toward any degree at this or any other university.

Signature:.....*HR*.....Date:...05 June 2024.....

Khoza, H.R.

Abstract

The research inspects the application of machine learning approaches to forecasting customer attrition in telecommunication companies. Machine learning models such as extreme gradient boosting, random forest, k-nearest neighbour, adaptive boosting, support vector machine, and logistic regression were used to forecast and compared the best model and analysed churn behaviour. Cross-validation techniques were applied to enhance model performance, revealing critical predictors of churn such as contract length, customer tenure, and service usage patterns. The results emphasised the effectiveness of machine learning in accurately identifying potential churners. Furthermore, the study emphasises the importance of leveraging predictive analytics to proactively address customer attrition, enabling telecommunication companies to devise targeted retention strategies and enhance customer satisfaction and loyalty.

Dedication

Dedication:
I dedicate this master research to the unwavering support and boundless love of my family, whose encouragement has been the anchor of my academic journey. To my friends and mentors, your guidance and insights have illuminated my path, shaped the course of my research, and fostered intellectual growth.

This endeavor is a testament to the resilience and dedication of my professors and colleagues, whose collective efforts have enriched my academic experience. Above all, I dedicate this research to the pursuit of knowledge and the countless individuals around the world whose stories, struggles, and triumphs inspire me to contribute meaningfully to the field.

Acknowledgments

I extend my heartfelt gratitude to my supervisor, Dr. TB Darikwa, for guiding me through this research project. I am deeply thankful for his mentorship and support. May the Almighty God continue to bless him. I also wish to express my special thanks to the National e-Science Postgraduate Teaching and Training Platform (NEPTTP) for their sponsorship.

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

Customer attrition is the occurrence of clients discontinuing their services with a telecommunications company; it is a pervasive and pressing challenge in the telecommunications market (Tanui, 2008). In 2019, it was reported that the telecommunications sector experienced an approximate 35% annual churn rate globally (King and Rice, 2019). Furthermore, it was mentioned that recruiting new clients costs 5 to 10 times more than retaining the existing ones (King and Rice, 2019).

The complex and competitive nature of this sector makes retaining customers of paramount importance. Customer churn can have profound implications for telecommunications companies, leading to significant revenue losses and hampering long-term sustainability (Tanui, 2008).

Understanding the factors that influence churn, as well as developing effective strategies to predict and mitigate it, has become a central focus of research within the industry (Beeharry and Fokone, 2022). This research shed light on the intricate dynamics of customer churn within telecommunication companies, examining its root causes and offering insights into predictive models and strategies for proactively reducing churn rates.

Since technology evolves and competition intensifies, telecommunications companies face a dynamic environment where customer expectations constantly evolve. In this context, providing superior customer service and tailoring offerings to individual customer needs is critical for retaining a loyal customer base (Beeharry and Fokone, 2022). Consequently, this study seeks to explore the multifaceted aspects of client attrition, examining the role of customer satisfaction, service quality, pricing strategies, and other factors that influence customer decisions. By gaining a deeper understanding of these variables, telecommunication companies will be able to develop more effective strategies for client attrition and bolster their competitive edge in the marketplace (Fujo et al., 2022).

Furthermore, utilising advanced analytics and predictive models has emerged as a vital tool for telecommunications companies in identifying at-risk customers (Fujo et al., 2022). This research also delves into predictive analytics and machine learning (ML) approaches to forecast customer churn, providing companies with valuable insights to address issues before they lead to actual churn. By aligning their strategies with data-driven predictions, telecommunication companies can work toward reducing customer attrition and fostering long-lasting customer relationships. In summary, this work endeavors to outline a holistic view of client attrition within the telecommunications sector and equip companies with knowledge and tools to address this pervasive challenge effectively (Fujo et al., 2022).

1.2 Background

According to Edwine et al. (2022), the telecommunications industry experienced an average annual churn rate ranging from 15% to 22%. However, it was further stated that in the African market, approximately 80% of the revenue came from the top 20% of customers, making it crucial to reduce churn among high-value customers.

However, advancements in ML and artificial intelligence (AI) have recently showcased opportunities for effectively forecasting customer churn (Patil et al., 2022). This trend is driven by several factors, including the increasing competition among telecommunications companies, technological advancements, and changing consumer preferences (Patil et al., 2022). The ramifications of customer churn are profound, as it leads to substantial revenue losses, threatens market share, and disrupts the industry's dynamics. In response to these challenges, there is a need for further research to solve the underlying causes of customer churn and develop effective strategies for retention (Srinivasan et al., 2023).

A critical aspect of the customer churn puzzle lies in understanding the driving forces behind this behavior. Factors such as service quality, pricing, customer service, and network coverage are pivotal in influencing customer decisions to stay or switch. (Srinivasan et al., 2023). Moreover, as customer expectations evolve, telecommunications companies must adapt swiftly to meet these changing demands. In light of this, the industry is progressively employing ML and data analytics techniques to predict client attrition, allowing companies to preemptively address customer dissatisfaction and take corrective actions to mitigate churn (Srinivasan et al., 2023).

Telecommunication companies have recognised the urgency of proactively addressing customer churn by improving customer experiences, enhancing the

quality of their services, and personalising offerings to cater to individual customer needs (Liu et al., 2022). This background research underscores the significance of investigating customer churn within the telecommunications sector, not only to fathom its root causes but also to unlock potential solutions by predicting customers who are likely to churn and identifying factors that influence customers to churn. The findings revealed from this work serve as a foundation for the development of strategies and predictive models that empower telecommunication companies to minimise customer churn, foster lasting customer relationships, and thrive in an increasingly competitive landscape (Liu et al., 2022).

1.3 Problem statement

Customer churn, which refers to clients who are likely to voluntarily discontinue their services, is very high in the telecommunications industry, with a rate ranging from 20% to 40% per year in South Africa (Huang and Kechadi, 2013). A high customer churn rate is problematic as it results in huge associated revenue losses. Reducing high revenue losses will require companies to understand the customer characteristics that churn and why they churn.

Reasons for churning have been investigated and are varied. They include, among others, leaving a company due to: (1) customer service and support system dissatisfaction (Rajamohamed and Manokaran, 2018); (2) life events such as moving to a new place, changing their financial condition, or going through personal changes (Huigevoort and Dijkman, 2015); and (3) high prices or unanticipated price rises. (Keramati et al., 2016). The characteristics of customers that churn are: (1) new clients who lack familiarity with the company's goods and services (Xiahou and Harada, 2022); (2) low-engagement customers who seldom use the product or service (Saran Kumar and Chandrakala, 2016); (3) young adults; and (4) low-income earners (Umayaparvathi and Iyakutti, 2016).

Understanding the characteristics and reasons for customer churn can help in predicting customers who are likely to churn. Such predictions are useful to companies that want to develop strategies that retain customers and improve customer satisfaction. This study intends to use customer churn models to predict client attrition in the telecommunications sector.

1.4 Rationale

Customer attrition prediction is a crucial challenge for telecommunication companies. By accurately identifying clients who are likely to turnover, industries should take effective steps to retain them and prevent the loss of valuable customers to their competitors. (Jain et al., 2020). Predictive analytics techniques can be employed to forecast churn behavior based on historical customer data.

Researching to identify potential customers likely to churn is crucial because, according to (Wagh et al., 2024), customer churn leads to significant revenue loss for telecommunication companies, and acquiring new customers is often more costly than retaining existing ones. Wagh et al. (2024) further mentioned that the telecommunication industry is highly competitive, and reducing churn can provide a competitive edge and improve market share. Maduna et al. (2024) states that understanding why customers leave can help companies improve their services, leading to higher customer satisfaction and loyalty. Additionally, (Amin et al., 2023) discovered that predictive models can help tailor services and offers to individual customer needs, enhancing customer experience. Furthermore, predicting churn allows companies to take proactive measures to retain customers, such as targeted marketing campaigns, loyalty programs, and personalised customer service.

Different models can be applied for attrition prediction, including extreme gradient boosting (xgboost), adaptive boosting (adaboost), support vector machines

(SVM), logistic regression (LR), decision trees DT), neural networks, and random forests (RF). The model selection relies on the specific requirements and characteristics of the telecommunications company's data. The churn prediction model is studied using historical data where the churn outcome is known

The typical modeling approach splits the dataset into training and validation. Algorithm performance is determined by metrics that include recall, precision, accuracy, and the measure of test of accuracy in binary classification analysis (F-score) (Saran Kumar and Chandrakala, 2016). After training and validation, the model becomes capable of forecasting churn for both new and existing customers. By inputting relevant customer data into the model, telecommunication companies can obtain churn probability scores for each customer. These scores help prioritise retention efforts and interventions.

However, telecommunications industries should create tailored retention strategies that cater to the unique needs of clients at risk of leaving. This can include personalised offers, discounts, improved customer service, loyalty programs, or proactive outreach to resolve issues. (Saran Kumar and Chandrakala, 2016).

By effectively predicting customer churn, telecommunication companies can allocate resources more efficiently, concentrate on retaining high-value clients, and reduce customer acquisition costs (Saran Kumar and Chandrakala, 2016). This study will utilise and deploy ML approaches to customer attrition using a dataset in the telecommunications industry.

1.5 Aim and objectives

1.5.1 Aim

This study aims to utilise ML techniques such as xgboost, adaboost, random forest, logistic regression, support vector machine, and k-nearest neighbour to

evaluate customer churn in the telecommunications industry.

1.5.2 Objectives

The objectives of the study are to:

1. Apply six ML algorithms to model and forecast customer churn.
2. apply the k-fold cross-validation technique to assess the performance of each algorithm,
3. compare the best-performing model between support vector machines, random forests, xgboost, logistic regression, adaboost, and k-nearest neighbour regarding overall prediction accuracy,
4. identification of the factors affecting customer churn.

1.6 Significance of the study

Customer churn presents a significant challenge for telecommunication companies, often leading to substantial revenue losses. This research identifies key factors contributing to customer churn and develops and applies six predictive models to effectively anticipate potential churners. This approach is poised to assist telecommunications companies in proactively reducing churn rates and preserving valuable customer relationships.

1.7 Structure of the dissertation

The dissertation is composed of five chapters. Chapter 1 delineates study's foundation, articulating the problem statement, rationale, goals, objectives, and the study's significance. Chapter 2 delves into a comprehensive exploration of historical and contemporary methods utilised to address customer

churn in the telecommunications industry, drawing insights from various studies. Chapter 3 outlines the methodology employed in this research and presents the mathematical formulations of the methods utilised. In Chapter 4, the findings are presented, and the outcomes are discussed. Finally, Chapter 5 encapsulates the overarching conclusions drawn from the research study's findings.

Chapter 2

Literature review

2.1 Introduction

This section examines the implications of ML algorithms. The research utilised supervised ML algorithm to model customer attrition prediction since the target factor to be predicted is already known. In this case, whether the customer has turnover with an outcome of ”yes” or ”no”.

When creating prediction models, supervised ML matches the input and output pairs from the training dataset. The prediction model is then used in classification by using explanatory variables from a test dataset as input into the prediction model (classification model) to predict the unknown values as output. In this study, we reviewed a selected number of ML algorithms, namely, SVM, K-nearest neighbour (KNN), xgboost, RF, LR, and adaboost. This section provides a basis for selecting the algorithms to apply to the data.

2.1.1 K-nearest neighbour

The KNN is frequently used for pattern classification (Soares et al., 2020). The KNN algorithm works so that it has to be given an N number of training vectors. Then, the method identifies the KNN points of an unknown vector and aims at determining its class (Soares et al., 2020).

The KNN algorithm's strength lies in its straightforward yet efficient implementation; its calculation time is less; and it has relatively high accuracy. KNN can memorise training data and apply directly to classify new input. This approach is simple and effective (Soares et al., 2020).

A similar study conducted by Brito et al. (2022) indicated that implementing the KNN algorithm is straightforward, since it primarily depends on computing the distance (usually Euclidean) between the variables. Furthermore, the KNN algorithm demonstrated the capability of handling large and incremental multi-class data. However, it was also shown that KNN is expensive in determining K if the dataset is large and that it needs more memory storage than an effective classifier or supervised learning algorithm (Brito et al., 2022).

2.1.2 Support vector machine

SVM stands out among supervised learning algorithms, commonly applied in tasks such as regression, classification, and pattern identification. Initially designed as binary classes without a stochastic nature, SVMs have evolved to address multi-class problems effectively. According to Carvalho et al. (2019), SVM constructs an N-dimensional hyperplane to partition data into N groups or classes optimally.

The goal of this model is to employ a surface that optimises the margin between various classifiers in the training data, effectively separating them (Cervantes

et al., 2020). Furthermore, the margin of the classifier is determined by the distance between the decision surface (hyperplane) and the closest data point (Cervantes et al., 2020).
The study findings by Tao et al. (2019) indicate that SVM excels in handling higher-dimensional data, presenting a notable advantage. Its primary strength lies in avoiding overlearning and excessive dimensionality, factors that contribute to both computational complexity and local extremum.

Furthermore, study conducted by Ahmad et al. (2017) shows that SVM consistently achieves high accuracy, making it a well-established and widely utilised machine-learning technique for addressing both classification and regression problems. This classifier identifies optimal boundaries for distinguishing between positive and negative training examples.

SVM necessitates an extended training period, rendering it impractical for large datasets, and it also struggles with overlapping classes (Tao et al., 2019). The SVM algorithm is unsuited for large datasets, mainly when dealing with noisy data or overlapping target classes, leading to suboptimal performance (Tao et al., 2019). Furthermore, SVM becomes less effective when there are more characteristics per data point than training data samples (Tao et al., 2019).

2.1.3 Random forest classifier

Ahmad et al. (2017) state that RF comprise an ensemble of tree predictors, the value of each tree determined by a uniformly random vector and is randomly selected among the entire forest. Nazarenko et al. (2019) adds that the RF ML algorithm employs a dual approach. Firstly, it utilises the bootstrap aggregating method, constructing multiple decision trees and aggregating their predictions. Simultaneously, the algorithm incorporates the random subspace approach, selecting a subset of M randomly chosen attributes from the overall

pool of available attributes (Nazarenko et al., 2019).

In the same vein, the work by Nazarenko et al. (2019) indicates that each newly generated tree assigns the item to a particular class to ascertain the winner, with class receiving an inflating number of votes received from trees emerging as the final prediction. According to Breiman (2001), the principle of large numbers precludes the random forest to avoid overfitting, which makes the algorithm a powerful prediction tool, and it provides a rapid, efficient, and dependable solution for mining data with high dimensions.

Ziegler and Konig (2014) states that RF performs well even in scenarios with numerous features and limited observations. Additionally, RF offers methods for easy feature selection and measuring variable significance. Notably, the initial step involves employing a RF to identify the important variables. Ultimately, RF proves to be a highly effective classifier.

The algorithm speed is considerably slower than other classification algorithms, as it relies on multiple decision trees to generate predictions (Nazarenko et al., 2019). When a random forest classifier makes a prediction, each tree in the forest must individually predict the outcome for the same input and collectively vote on it. This process can be notably time-consuming (Farnaaz and Jabbar, 2016). Furthermore, due to their sluggish performance, random forest classifiers may not be suitable for real-time predictions (Farnaaz and Jabbar, 2016).

2.1.4 Logistic regression

LR extends its applicability by being employed in scenarios where the dependent variable, Y, is categorical (Fritz and Berger, 2015). Dalvi et al. (2016) state that LR was applied in order to scrutinise dataset, seeking to identify connections between several independent traits and the dependent variable. There are three available types of LR models: multinomial LR is used when

the dependent variable has more than two categories that are not ordered. This type of regression models the probabilities of the different possible outcomes, Ordinal LR is used when the dependent variable has more than two categories with natural order, but the distances between the categories are not assumed to be equal, lastly, Binary LR is used when the dependent variable is binary, meaning it has only two possible outcomes. (Dalvi et al., 2016). In this research, we used binomial LR because our target variable takes binary outcomes which can be a customer has churned or not churned.

Initially, LR stood out as the most widely adopted method for binary outcomes. A study by Park (2013) highlights its growing popularity, emphasising its suitability for modeling a binary dependent variable alongside multiple independent variables. The simplicity of logistic regression is underscored, as it does not require optimisation of hyperparameters (Park, 2013).

Despite having a tiny sample size, few events, and simple variables, logistic regression works well, as shown in the study by (Nusinovici et al., 2020). The research shows that logistic regression surpasses ML models in predicting the risk of major chronic illnesses. Another advantage lies in logistic regression's ability to swiftly identify suitable candidates from untested resources, thereby saving computation time and costs, as noted in Liu et al. (2021).

According to Park (2013), the logistic model believes in a linear association between output variables and input variables, which cannot always hold in real-world scenarios. As a linear model, any non-linear relationships must be explicitly defined by transforming the original data and introducing additional terms to the model. Sensitivity to outliers is a concern, impacting coefficients and predictions (Park, 2013). Overfitting may occur if the amount of independent features is excessively large relative to the sample size. (Park, 2013). Lastly, LR is specifically designed for analysing binary outcomes and may not

be suitable for other types of outcomes (Park, 2013).

2.1.5 Extreme gradient boosting

This algorithm is designed for machine boosting, specifically enhancing the effectiveness of the ML algorithms. It is crafted and optimised to be effective, versatile, and easily portable (Wade and Glynn, 2020).

The xgboost model outpaces other gradient-boosting implementations in terms of speed. Regarding model accuracy, it is currently recognised among the most precise ML algorithms (Zhao et al., 2020). Xgboost excels in training algorithm performance, demonstrating proficiency in classification and regression tasks, attributed to its underlying gradient-based approach that enhances adaptability to imbalanced datasets (Peng et al., 2019). Moreover, the scalability of xgboost is a key factor in its success, being scalable to accommodate billions of samples in distributed or memory-constrained environments, and operating at a speed greater than ten times quicker than conventional single-computer solutions (Zhao et al., 2020). xgboost is sensitive to outliers (Peng et al., 2019).

2.1.6 Adaptive boosting

Adaboost is an ML ensemble method which is used for grouping and linear regression tasks. To build a strong classifier, it combines several weak learners, usually decision trees. This type of learner prioritises sample distribution adjustments made for erroneously classified samples during training and repeats this process until the weak classifier has completed a predefined number of training cycles, denoting the end of the learning process. (Chen and Guestrin, 2016).

Adaboost offers the advantage of reclassifying previously misclassified datasets

of subset training runs by adjusting the error dataset's weight (Chen et al., 2019). Furthermore, adaboost exhibits a low intrinsic boosting overhead, with the training times of the core algorithms predominantly influencing the duration of constructing the final model (Mohammad et al., 2019). Research by Liu (2010) highlighted adaboost as a popular and effective boosting algorithm, especially when paired with decision trees as a strong or weak learner, showcasing its efficacy in tasks such as fingerprint classification. Adaboost is sensitive to outliers (Liu, 2010).

2.2 Review of the application of ML algorithms to customer churn and the associated risk factors

The study carried out in China by Edwine et al. (2022) investigated a proficient approach to assess the likelihood of customer churn in the telecommunications industry, employing sophisticated ML techniques that include RF, SVM, and KNN. The models' performance, was assessed using the following metrics: mean absolute error, F1-score, accuracy, recall, precision, and area under the curve (AUC) (Edwine et al., 2022). The results indicated that the top-performing models based on accuracy metrics were support vector machines, random forests, and KNN (Edwine et al., 2022).

Further investigation revealed that the range of services each customer subscribed to (phone, multiple lines, internet, account information tenure, contract type, payment method, preference for paperless billing, monthly charges, and total charges), as well as online security, device protection, online backup, streaming television and movies, and finally, tech support, were among the independent variables examined. The research also uncovered further infor-

mation about the customers' demographics. (Edwine et al., 2022).

Kiguchi et al. (2022) conducted research to determine customer churn rate in the telecommunications sector and develop an attrition prediction model by comparing LR, DT, adaboost, and RF models. The study indicates that AUC and, F1-score were applied to measure the best model accuracy. The findings indicated that LR performed best, followed by RF and adaboost based on AUC and F1-score. The research also revealed that commonly employed explanatory factors encompassed gender, tenure, contract type, payment method, monthly charges, and customers possessing multiple lines(Kiguchi et al., 2022).

Olufemi and Strydom (2018) researched to predict customer churn with customer demographic data, the range of services each customer has signed up for, and customer account information in the telecommunications industry. The study revealed that five algorithms were compared, which were adaboost, LR, xgboost, RF, and KNN. It was found that using F-score and mean squared error LR was the best-performing model, followed by xgboost, RF, KNN, and adaboost.

A research study was carried out in Ethiopia by Seid and Woldeyohannis (2022) to design and develop a ML model capable of accurately predicting churned customers from the total customer base of the Commercial Bank of Ethiopia (CBE). They applied and compared supervised algorithms for ML, such as LR, RF, SVM, and KNN, using customer demographic data, the range of services each customer has signed up for, and customer account information. The result revealed that the SVM was the best-performing model, followed by RF, LR, and KNN when accuracy and F-score were applied (Seid and Woldeyohannis, 2022).

Research was conducted by Wadikar (2020) to develop an ML model to predict customer churn for an Insurance company, using quantitative and deductive strategies, with methods including LR, RF, SVM, and Neural Network. The RF

model emerged as the best classifier with an accuracy of nearly 97%, precision of 91%, and recall of 98%.

Comparing the study conducted by Seid and Woldeyohannis (2022) we discovered that they applied LR, RF, SVM, and, KNN models to predict customer churn on the banking dataset. Their results revealed that RF was the best-performing model with an accuracy score of 93%, and an F-score of 88% while Wadikar (2020) have applied LR, RF, SVM, and Neural network models on the Insurance dataset. Their result indicates that RF was the highest performing model with an accuracy of nearly 97%, precision of 91%, and recall of 98%.

The research study in South Africa, conducted by Olufemi and Strydom (2018) employed a questionnaire-based approach to gather the dataset from customers of various South African telecommunications services. Their study showcased using Bayesian models for predicting customer churn, revealing favorable performance results. Additionally, the study indicated three factors that had a greater impact on client attrition in South Africa, as identified by the predictions derived from the model: Offers Promotions (OP), Customer Care Service (CCS), and Friends and Family Deals on Networks (FFD) (Olufemi and Strydom, 2018).

Studies have consistently shown that customer churn is a significant concern for South African telecommunications operators, with churn rates ranging from 10% to 30% (Glendah et al., 2019). A study by Abu Ghazaleh and Zabadi (2020) found that the main factors influencing customer churn in South Africa include poor network quality, high tariffs, and inadequate customer service. Similarly, a study by Fernandez Montoya et al. (2023) identified billing issues, lack of transparency, and limited value-added services as key drivers of churn.

Another study by Soltani-Nejad et al. (2021) compared churn rates across different mobile network operators in South Africa and found that smaller op-

erators experience higher churn rates due to their limited network coverage and infrastructure. The study also highlighted the importance of customer retention strategies, such as loyalty programs and personalised marketing, in reducing churn. Furthermore, research by Mashau (2020) explored the impact of social media on customer churn and found that social media engagement can significantly influence customer loyalty and retention.

”Customer Churn Prediction in the South African Telecom Industry: A Case Study of Vodacom and MTN” by Mashau (2020), the study used ML algorithms to predict customer churn for Vodacom and MTN customers, identifying key factors such as billing issues and network quality. Another study conducted by Molapo and Mukwada (2014) compared customer retention strategies between Cell C and Telkom, the results indicated the importance of personalised marketing and loyalty programs.

A review of the literature by Poudel et al. (2024) emphasised the need for South African telecommunications operators to adopt data-driven approaches to understand customer behavior and preferences. The study highlighted the importance of predictive analytics and ML algorithms in identifying high-risk customers and developing targeted retention strategies. Overall, the literature suggests that customer churn in the South African telecommunications industry is a complex issue influenced by a range of factors, and operators must adopt a multifaceted approach to reduce churn and improve customer retention.

Chong et al. (2023) conducted a study aimed to compare several supervised ML algorithms, including RF, adaboost, LR, xgboost, KNN, and SVM. They determined which model was the best by using criteria like precision, accuracy, and F1 score and, recall. It was found that the xgboost algorithm emerged as the most successful model for predicting customer churn, followed by LR, KNN,

RF, adaboost, and SVM, respectively (Chong et al., 2023). The results also indicated that gender, tenure, contract type, payment method, monthly charges, and customers with multiple lines were the most important explanatory variables (Chong et al., 2023).

The literature reviewed in this research applied similar ML models, including LR, RF, SVM, adaboost, xgboost, KNN, and neural network, to datasets from various sectors such as banking, insurance, and telecommunications. It was found that RF and xgboost consistently outperformed other models using metrics like F-score, accuracy, precision, and recall. Although the datasets differed, the results were comparable. In this study, we wanted to apply the same models LR, RF, SVM, KNN, xgboost, and adaboost to a telecommunications dataset to determine if the outcomes would be similar to those in the banking and insurance sectors.

Based on leading South African studies on customer churn in telecommunications, it was shown that the billing method, internet service type, and network quality had the most significant influence on customer churn, which aligned with the findings of this study. Additionally, personalised marketing and loyalty programs were identified as the most effective strategies for telecommunication companies to implement to reduce the customer churn rate.

2.3 Summary of the chapter

In this section, we have examined selected supervised ML models tailored for binary classification problems. The discussion encompasses the functionality, advantages, and disadvantages of each algorithm. Additionally, it explores the relevant work conducted by researchers to predict client attrition within the telecommunications industry and other industries. The top-performing models

for customer churn studies differ across various research endeavors. Nonetheless, some consistently studied models include LR, KNN, SVM, xgboost, adaboost, and RF. Based on the findings of the review study, we opted to employ these models in our dataset analysis. Furthermore, we identified the significant variables elucidating customer churn as the services each customer has subscribed to, customer account details, and demographic information. These were considered our independent variables in this study.

Chapter 3

Methodology

3.1 Introduction

This part primarily delves into a detailed explanation of the methodology and mathematical models, specifically descriptive statistics, that are applied to forecast customer churn.

3.2 Data source and study area

Datasets were extracted with 7043 rows from the Kaggle website (International Business Machines Corporation) from 2020 to 2022 (Kaggle, 2020). The Kaggle website is an open source for data scientists who want to participate in free data science projects. The dataset contains customer information that includes an independent variable named churn and explanatory variables that include, among others, demographic information such as customer gender and how long a customer has been in the company, call records, billing information, service

usage, complaints, and customer interactions.

3.3 Overall research approach and design

In this study, secondary data from Kaggle was used to perform descriptive data analysis, which was useful to discover the insights of the data and build six machine learning models to predict customer churn by comparing SVM, KNN, RF,xgboost, adaboost, and LR classifiers, respectively. The confusion matrix, roots mean squared, recall, precision, and ROC (Receiver Operating Characteristic) curve were utilised to identify the best effective algorithm since this is a classification problem.

The data used in this research was suitable for a machine learning approach since it included historical records of customer behavior, usage patterns, billing information, interaction history, internet usage, customer service interactions, payment history, and demographic information, all of which are relevant for understanding and predicting churn. The quality of the data was assessed, as machine learning requires a good dataset.

The accuracy of the data was validated to ensure it was free from errors and inconsistencies, including correct billing information, service usage records, and customer interaction logs. The completeness of the data was ensured by addressing any missing values. Consistency was also maintained by ensuring correct data types for all attributes, and the data was finally encoded into numeric values to enable machine learning models to interact with it effectively when making predictions.

3.4 Machine learning data analysis approach

In this study, quantitative data collected from Kaggle was coded and categorised using Python in Jupyter Notebook software. All the categorical explanatory variables were encoded to numerical values, as this is one of the requirements in machine learning analysis. Relationships were investigated to check which explanatory variables contribute to high customer churn; then, the RF, SVM, KNN, adaboost, xgboost, and LR models were fitted to predict customer churn using Python libraries.

3.4.1 Logistic regression classifier

LR falls under the umbrella of machine learning techniques tailored for binary classification tasks, such as forecasting the probability of an outcome, event, or observation. This study applied an LR algorithm to predict if a customer churns, with 1 or 0 as an outcome, respectively. To map predictions and their probability, logistic regression employs the sigmoid and logistic functions. This function converts any real value into a specific range spanning from 0 to 1, as seen by its S-shaped curve.

Moreover, once the sigmoid function's output (estimated probability) meets a pre-established cutoff on the graph, the model determines that the case belongs to that category. Conversely, if the estimated probability falls below the predefined cutoff, the model infers that the case does not pertain to the category. The sigmoid function represents the activation function for logistic regression and can be described as:

$$f(x) = \frac{1}{1 + e^{-x}}, \tag{3.1}$$

where e represents the base of natural logarithms, and x denotes the trans-

formation of the numerical value. The subsequent equation illustrates logistic regression:

$$y = \frac{e^{(b_0+b_1x)}}{1+e^{(b_0+b_1x)}}. \tag{3.2}$$

The above is a LR sigmoid function where: x is the input value, the explanatory variables that influence customer churn, y is the predicted output, the dependent variable that represents if a customer has churned or not, b_0 signifies the bias or intercept term, while b_1 indicates the coefficient linked to the input (x). This formula resembles linear regression's linear combination of input values with weights or coefficients to measure an output value. However, unlike linear regression, the predicted output value in this case is binary (0 or 1) rather than a continuous variable.

3.4.2 Random forest classifier

The RF algorithm for binary classification can be expressed using an ensemble of decision trees. Let's denote the output (churn or not churn) of the RF algorithm as $\widehat{Y}$, and the set of decision trees in the forest as $\{T_1, T_2, T_3, ..., T_n\}$. The final prediction of the random forest ensemble is typically decided by a majority vote or by averaging the outcomes generated by its component trees. For binary classification, this can be represented as:

$$\widehat{Y} = \frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_i, \tag{3.3}$$

where n is the sample size.

3.4.3 Support vector machine

In binary classification, SVM aims to determine a hyperplane that effectively divides data into two types while maximising the margin between them. (Perrotta et al., 2017). The decision function of a linear SVM for binary classification is represented by the following formula:

$$f(x) = sign((\sum_{i=1}^{N} \alpha_i y_i K(x, x_i)) + b), \tag{3.4}$$

where $f(x)$ denotes the decision function for predicting customer churn classes of the input (independent variable) x, $sign(.)$ function returns -1 for negative values, 0 for zero, and 1 for positive values. Here, N represents the count of support vectors, α_i represents the Lagrange multiplier linked with each support vector, y_i represents the class label of the i-th support vector, $K(x, x_i)$ represents the kernel function, which measures the similarity between the input x and the support vectors x_i, while b indicates a bias term. Throughout the SVM training process, the aim is to identify the optimal values of α_i and b that maximise the margin while satisfying the constraint:

$$y_i(\sum_{i=1}^{N} \alpha_i y_i K(x, x_i)) + b \geq 1, \text{for i = 1,2, ..., N}, \tag{3.5}$$

where x_i represents explanatory variables, y_i is the class label of the i-th support vector., and b denotes a bias term This constraint guarantees that data points are accurately classified and positioned on the correct side of the decision boundary, maintaining a margin of at least 1.

3.4.4 K-nearest neighbour

The KNN algorithm determines the K clusters that are a nearest neighbour of a given dataset point by employing a metric of distance that includes the distance in Euclidean space. The value or class of the dataset point is subse-

quently obtained by utilising the majority decision or the average of its K cluster neighbours. This method enables the algorithm to adapt to changing patterns and generate predictions according to the structure of the local dataset. K denote the set of nearest neighbours of dataset X as subset S_x. In this case, the subset is the set of independent variables that influence customer churn. Mathematically, S_x is denoted by $S_x \subseteq D$ such that $|S_x| = k$ for all

$$(x^i, y^i) \in D \; s_x, \tag{3.6}$$

$$dist(x, x^i) \subseteq MAX_(x^", y^") \in S_x dist(x, x^"), \tag{3.7}$$

for all the points in D that do not belong to S_x. The distance from X to any point in S_x is at least as great as the distance to the farthest point in S_x. We define the classifier $h(.)$ as a function that yields the most frequent label in S_x:

$$h(x) = mode(y^" : (x^", y^") \in S_x), \tag{3.8}$$

where $mode(.)$ means to select the label of the highest occurrence, in case either customer has churned or did not churn, denoted by 1 and 0, respectively.

3.4.5 Extreme gradient boosting classifier

The xgboost is among one of the most widely used and effective implementations of the gradient-boosted trees model (Chen and Guestrin, 2016). This supervised learning method emphasises function approximation by optimising particular loss functions, and incorporating various regularisation methods to enhance performance (Chen and Guestrin, 2016). At iteration t, we aim to minimise the following objective function, comprising both the loss function and regularisation components:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)}) + f_t(\mathbf{x}_i) + \Omega(f_t), \tag{3.9}$$

y_i is a real value (label) known from the training data set that can be seen as $f(x + \Delta x)$ where x =$\hat{y}_i^{(t-1)}$, $\mathcal{L}^{(t)}$ represents the objective function at iteration t. It's the function we seek to minimise during the model's training process. l stands for the loss function, quantifying the disparity between the true label y_i and the predicted value $\hat{y}_i^{(t-1)}$ from the previous iteration $(t-1)$, y_i denotes the true label or actual value for the i-th training instance, $\hat{y}_i^{(t-1)}$ represents the predicted value for the i-th training instance at iteration $(t-1)$ from the previous iteration $(t-1)$, f_t is the function (often a decision tree in the context of gradient boosting) being added at iteration t, x_i represents the feature vector of the i-th training instance.

This term represents the prediction for the i-th instance made by the new function added in the current iteration, $\Omega(f_t)$ this represents the regularisation term for the function. Regularisation is used to penalise model complexity to prevent overfitting. This term can include various factors such as the number of leaves in a tree, the depth of the tree, and other complexity measures. The objective function in xgboost is a composite function, where l depends on the output of the current and preceding classification and regression trees (CART) learners, essentially forming a cumulative sum of these additive trees. We use Taylor's theorem and gradient-boosted trees to minimise this function.

$$f(x) \approx f(a) + f'(a)(x - a) \tag{3.10}$$

$$\Delta x = f_t(\mathbf{x}_i)$$

$f(x)$ denotes the value of the function f at a given point x. $f(a)$ denotes the value of the function f at a specific point a, his denotes the derivative of the $f'(a)$ function f evaluated at point a. The derivative gives the rate at which f changes at a. Δx is defined as the output of the function f_t when applied

to the feature vector x_i of the i-th training instance. In the context of gradient boosting, $f_t(\mathbf{x}_i)$ represents the prediction refinement performed by the t-th model (typically a decision tree) for the i-th instance. In our context,$f(x)$ represents the loss function l, where a corresponds to the predicted value from the previous step $(t-1)$, and Δx indicates the new learner to be included at step t, with the aforementioned elements, in every iteration t, we can represent the objective (loss) function as a simple function of the newly introduced learner. This enables us to employ optimisation methods within the Euclidean space. As previously stated, a denotes the prediction at step $(t-1)$, and $(x-a)$ signifies the new learner that must be incorporated at step t to minimise the objective function greedily. Hence, opting for the second-order Taylor approximation yields:

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 \tag{3.11}$$

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t), \tag{3.12}$$

where g_i = $\partial \hat{y}^{(t-1)}(l(y_i, \hat{y}_i^{(t-1)})$ and h_i = $\partial^2 \hat{y}^{(t-1)}(l(y_i, \hat{y}_i^{(t-1)})$.

The above is the first and second order gradient statistics of the loss function, Finally, after removing the constant components, the objective to minimise at step t simplifies to the following:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t). \tag{3.13}$$

In binary classification utilising log loss optimisation, let's examine the scenario involving a log loss objective function:

$$(y)\ln(p) + (1-p)\ln(1-p), where, p = \frac{1}{1+e^{-}x}, \tag{3.14}$$

where y signifies the true label, which belongs to 0,1, and p indicates the probability score. It's crucial to emphasise that p (score or pseudo-probability) is calculated after applying the widely recognised sigmoid function to the output of the gradient boosting trees (GBT) model x. The model's output x is the sum of the predictions from the CART tree learners.

3.4.6 Adaptive boosting classifier

Adaboost, which stands for adaptive boosting. It stands out as one of the pioneering algorithms in the Boosting domain of machine learning, primarily tailored for binary classification tasks (Hastie et al., 2009). For a dataset containing N samples, we set the weight of each data point as $w_i = \frac{1}{N}$ during initialisation, For m ranging from 1 to M:

(a) Draw training samples x_i from the dataset using the weights $w_i^{(m)}$.

(b) Train a classifier K_m using all the training samples x_i.

(c) Calculate $\mathcal{E}$ using the formula $\mathcal{E} = \frac{\sum_{y_i \neq K_m(x_i)} w_i^{(m)}}{\sum_{y_i} w_i^{(m)}}$, where y_i represents the true value of the target variable, and w_i^m denotes the weight of sample i at iteration m.

(d) Calculate α_m using the formula $\alpha_m = \frac{1}{2} \ln \frac{1-\mathcal{E}}{\mathcal{E}}$.

(e) Adjust all the weights according to the formula $w_i^{(m+1)} = w_i^{(m)} e^{-\alpha_m y_i K_m(x)}$.

The new predictions are determined by $K(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m K_m(x)\right)$.

3.4.7 Chi-square test for feature selection

Within a dataset, categorical features are subjected to the chi-square test. We determine the optimal features with the highest chi-square scores by computing each feature's chi-square and target variable. Features demonstrating substantial interdependence with the target variable are crucial for prediction and

may warrant further scrutiny. Selecting the most pertinent attributes from the dataset and applying machine learning techniques to improve model performance is known as feature selection or attribute selection. An excess of superfluous characteristics increases training time and increases the overfitting probability.

Statistically, we initially established the two hypothesis statements to conduct a chi-square test. The null hypothesis (H0) posits that the two categorical variables exhibit no significant correlation. The alternative hypothesis (H1) contends that the two categorical variables display a significant correlation. We construct a contingency table illustrating the frequency distribution of the two categorical variables. Next, we determine the contingency table's predicted values by:

$$E_{ij} = \frac{R_i C_j}{N}, \tag{3.15}$$

where R_i denotes the sum of row i, C_j denotes the sum of column j, and N represents the total number of observations.
We calculate chi-square statistic using

$$\chi^2 = \frac{\sum (O_{ij} - E_{ij})^2}{E_{ij}}, \tag{3.16}$$

where O_{ij} represents the observed values and E_{ij} denotes expected values. Compare the computed chi-square statistic (χ^2) with the critical value derived from the chi-square distribution table corresponding to the selected significance level (e.g., 0.05). If the value of χ^2 exceeds the critical value, we reject the null hypothesis, suggesting a noteworthy association between the variables. If χ^2 is less than or equal to the critical value, we do not reject the null hypothesis, implying no significant association. Alternatively, when using α = 0.05 (selected significance level) and the p (p-value), if $p < 0.05$ we reject the null

hypothesis; otherwise, we do not reject the null hypothesis.

3.5 Evaluation techniques

3.5.1 k-fold cross-validation

Although cross-validation ensures that there is no overlap in the test, its mechanism resembles that of repeated random subsampling. In k-fold cross-validation, the initial learning set is divided into k distinct subsets; each subset is roughly the same size. Every fold denotes a resultant subgroup. The cases in the learning set are randomly sampled for this division without replacement. The training set comprises all the $k-1$ subsets used for model training. Afterward, the model proceeds to work with the remaining subset, referred to as the validation set, to evaluate its performance. This process repeats until each of the k subsets has been utilised as the validation set. The cross-validated performance represents the average performance obtained from the evaluations conducted across the k validation sets.

In a broader context, $\widehat{f_k}$ signifies the model trained on all subsets except the k-th subset of the training set. The value $\widehat{y_i} = \widehat{f_k}(x_i)$ denotes the predicted or estimated value for the actual class label, y_i, of case x_i, which is part of the k-th subset. The prediction error estimate from cross-validation, denoted as ϵ_{cv}, is then defined as:

$$\epsilon_{cv} = \frac{1}{n}\sum_{i=1}^{n} \iota(y_i, \widehat{f_k}(x_i)). \tag{3.17}$$

Cross-validation frequently employs a technique called stratified random sampling. This method ensures that the sampling reflects the class proportions present in the learning set within the individual subsets.

3.5.2 Root mean squared error

RMSE (root mean squared error) offers insight into the dispersion of outcome errors, which is especially valuable for assessing model performance relative to the scale of the target variable. Lower RMSE values signify superior model performance, indicating reduced prediction errors.

RMSE (root mean squared error), represented as E_i for a singular model, is computed using the subsequent formula:

$$E_i = \sqrt{\sum_{j=1}^{n}(p_{ij} - T_j)^2}, \tag{3.18}$$

where p_{ij} denotes the value predicted by the singular model i for entry j (among n entries), while T_j represents the target value for entry j. In an ideal prediction scenario, $p_{ij} = T_j$, resulting in $E_i = 0$. As a result, the index E_i ranges from 0 to infinity, with 0 indicating perfect prediction.

3.5.3 Confusion matrix for the binary classification

A confusion matrix provides a structured representation commonly utilised for assessing the efficacy of a binary classification algorithm. It presents a detailed breakdown of the model's predictions vis-à-vis the actual class labels. The key components of a confusion matrix in binary classification encompass: True Positive (TP): Occurs when the model accurately predicts the positive class, False Positive (FP): Occurs when the model erroneously predicts the positive class while the true class is negative (Type I error), True Negative (TN): Occurs when the model accurately predicts the negative class, False Negative (FN): Occurs when the model erroneously predicts the negative class while the true class is positive (Type II error).

From these values, various performance metrics can be calculated, including:

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Table 3.1: Confusion matrix

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

Precision (Positive predictive value) = $\frac{TP}{TP + FP}$

Recall (Sensitivity, true positive rate) = $\frac{TP}{TP + FN}$

F1 Score = $2 \times \frac{precision \times Recall}{Precision + Recall}$

Specificity (true negative rate) = $\frac{TN}{TN + FP}$

These metrics offer insights into various facets of the model's performance. Precision evaluates its capability to identify positive instances correctly; recall measures its ability to capture all positive instances; Accuracy provides an overall performance assessment; the F-score gauges the test's accuracy; and specificity indicates the test's ability to classify subjects who genuinely lack the outcome of interest.

3.5.4 Receiver operating characteristic area

The AUC-ROC curve functions as a performance measure for classification tasks across various threshold values. It illustrates a probability curve, with the AUC quantifying the degree of separability, which reflects the model's capacity to differentiate between classes. A high AUC suggests that the model is more effective at distinguishing between churning and non-churning clients.

3.6 Summary of the chapter

In this section, we have explored the mathematical formulations of our models and the evaluation methods used to assess their performance.

Chapter 4

Results and discussion

4.1 Introduction

This section discusses machine learning models and data analysis for predicting customer attrition in the telecommunications industry.

4.2 Exploratory data analysis

In this chapter, we analyzed our dataset and applied six machine learning models to predict customer churn. We also compared the performance of these models.

4.2.1 Descriptive statistics

We begin our analysis by looking at three numerical variables which represents the period the customer has been with the company in months; ”monthly total charges,” which display the current sum of the customer’s monthly charges

for all company services; and ”total charges,” which indicate the cumulative charges accrued by the customer until the end of the quarter.

Table 4.1: Descriptive statistics

	Count	Mean	Std	Min	25 perc	50 perc	75 perc	Max
Total charges	7043	2279.73	2266.80	0.00	389.55	1394.55	3786.60	8684.80
Tenure	7043	32.37	24.56	0.00	9.00	29.00	55.00	72.00
Monthly charges	7043	64.76	30.09	18.25	35.50	60.35	89.85	118.75

Table 4.1 shows that the company's longest-serving customers have remained loyal for 72 months, while the newest customers recently joined just last month, marking 0 months of tenure. On average, customers have been with the company for approximately 33 months. The highest monthly payment recorded was 118.75, whereas the lowest was 18.25. On average, customers have paid 64.45 per month. The highest total charges amount to 8684.80, with an average of 2279.73, while the lowest total charge is 0.

4.2.2 Bivariate and multivariate analysis

Customers who churn are influenced by different factors; we will investigate which factors are likely to drive customer churn.

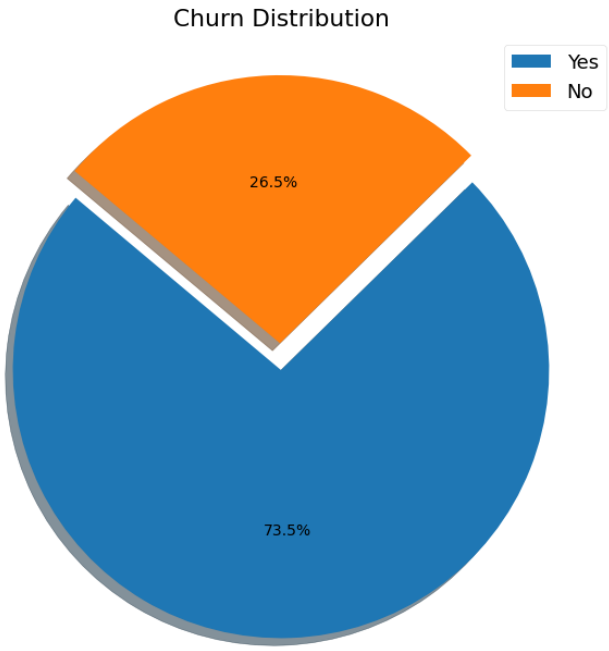

Figure 4.1: Customer churn distribution

Figure 4.1 shows that 26.5% of customers have churned and 73.5% of customers have not churned.

Table 4.2: Categorical feature distribution

	Churn (%)		Chi Square Test	
	YES	**NO**	**Alpha**	**P value**
Gender			0.05	0.490
Male	13.20	37.27		
Female	13.33	36.19		
Internet Service			0.05	0.009
Digital subscriber line (DSL)	6.52	7.86		
Fibre optic	18.42	25.54		
No	1.60	20.06		
Contract			0.05	0.005
Month-to-month	23.50	31.52		

Continue on the next page

Table 4.2 – *Continued from previous page*

	Churn (%)		Chi Square Test	
	YES	**NO**	**Alpha**	**P value**
One year	2.36	18.56		
Two year	0.68	23.384		
Payment Method			0.05	0.003
Bank transfer (automatic)	3.66	18.26		
Credit card (automatic)	3.29	18.32		
Electronic check	18.37	15.21		
Mailed check	18.51	4.37		
Partner			0.05	0.002
No	17.04	34.66		
Yes	9.50	38.80		
Dependents			0.05	0.004
No	21.13	48.13		
Yes	4.63	25.33		
Phone Service			0.05	0.34
No	7.27	2.41		
Yes	24.12	66.19		
Multiple Lines			0.05	0.003
No	20.74	28.92		
No phone service	1.60	20.06		
Yes	4.19	24.47		
Online Security			0.05	0.002
No	12.05	36.08		
No phone service	2.41	7.27		
Yes	12.07	30.12		
Online Backup			0.05	0.002
No	17.51	26.34		
No phone service	1.60	20.06		
Yes	7.43	27.06		
Device Protection			0.05	0.002
No	17.19	26.75		
No phone service	1.60	20.06		
Yes	7.74	26.65		
Tech Support			0.05	0.001
No	20.53	28.78		
No phone service	1.60	20.06		
Yes	4.40	24.62		
Streaming television			0.05	0.005

Continue on the next page

Table 4.2 – *Continued from previous page*

	Churn (%)		Chi Square Test	
	YES	**NO**	**Alpha**	**P value**
No	13.37	26.52		
No phone service	1.60	20.06		
Yes	11.56	26.88		
Streaming Movies			0.05	0.002
No	13.32	26.22		
No phone service	1.60	20.06		
Yes	11.61	27.18		
Paperless Billing			0.05	0.004
No	6.66	34.12		
Yes	19.88	39.34		

Table 4.2 shows Chi-square test for association between each categorical variable and customer churn demonstrating that only gender and phone service are not associated with customer churn based on their $p > \alpha = 0.05$ (chosen significant level). Consequently, these variables will be excluded from the model-building process. We discovered that the difference between customer churning based on gender is almost equal, with approximately 13.20% male churn rate and 13.33% female churn rate. Most clients who churn use fiber optics as their internet service, followed by DSL users with 18.42% and 6.52%, respectively. When it comes to contracts, we observed that those who opt for monthly contracts are likely to churn more than those who opt for yearly contracts, with 23.50%, 2.36%, and 0.68%, respectively.

Mailed and electronic checks are the leading payment methods to influence high customer churn rates with 18.51% and 18.37%, while bank transfer (automatic) and credit card (automatic) have approximately a 4% churn rate. Customers without partners and with no people depending on them seem to have a high churn rate, with 17.04% of customers without a partner and 21.13% of customers without people depending on them. This rationale is logical because

there are no individuals compelling them to maintain their subscription to the telecommunication service for an extended period. For instance, they lack dependents like children requiring internet access for school or siblings seeking employment opportunities and using smart televisions.

Additionally, customers without phone services exhibit elevated churn rates, influenced by factors such as streaming movies and movies, multiple lines, online security, online backup, device protection, and tech support. Moreover, individuals opting for paper billing are more prone to churn, with a rate of 19.88%, compared to those who prefer digital billing, with a churn rate of 6.66%.

4.2.3 Model performance comparison

Table 4.3: Model performance comparison

Metric	Xgboost	Adaboost	Logistic Regression	KNN	SVM	Random Forest
Accuracy	0.78	0.80	0.80	0.75	0.80	0.79
F-Score	0.56	0.59	0.59	0.52	0.55	0.55
Precision	0.61	0.66	0.65	0.53	0.66	0.64
Recall	0.52	0.54	0.55	0.50	0.48	0.50
ROC-AUC	0.84	0.86	0.86	0.79	0.82	0.84
RMSE	0.46	0.44	0.45	0.50	0.45	0.46

Table 4.3 shows that accuracy: adaboost, LR, and SVM have the highest accuracy (0.80), followed closely by RF (0.79). F-Score: adaboost and LR both have the highest F-Score (0.59). Precision: adaboost and SVM have the highest precision (0.66). Recall: LR has the highest recall (0.55). ROC-AUC: adaboost and LR have the highest ROC-AUC (0.86). RMSE: adaboost has the lowest RMSE (0.44), indicating the best performance in terms of error.
Based on the metrics, adaboost seems to be the best overall performer. It has the highest accuracy (0.80). It ties for the highest F-Score (0.59) and precision (0.66). It has a very competitive recall (0.54). It shares the highest ROC-AUC score (0.86). It has the lowest RMSE (0.44), indicating the least error.

4.2.4 Xgboost model performance

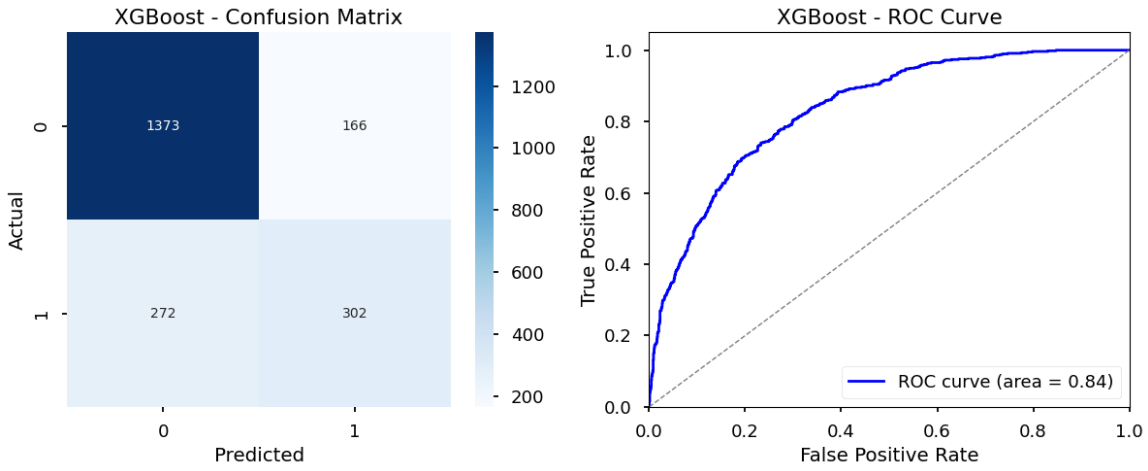

Figure 4.2: Confusion matrix and ROC curve for xgboost

The confusion matrix and ROC curve analysis of the xgboost model in figure 4.2 reveal the following:

The model accurately classified 302 instances as True Positives (TP) and 1373 instances as True Negatives (TN). Additionally, it misclassified 166 instances as False Positives (FP) and 272 instances as False Negatives (FN). The ROC AUC for this model is 0.84, signifying its robust capability to differentiate between the positive and negative classes.

The xgboost model demonstrates a commendable balance between sensitivity and specificity, as evidenced by its strong ROC AUC score, underscoring its overall effectiveness in class discrimination.

4.2.5 Adaboost model performance

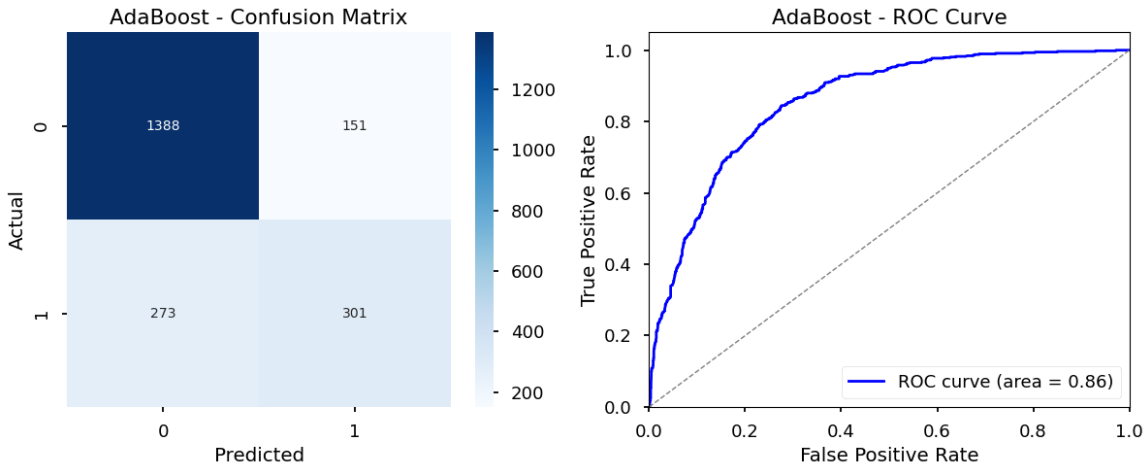

Figure 4.3: Confusion matrix and ROC curve for adaboost

The analysis of the confusion matrix and ROC curve for the adaboost model in figure 4.3 yields the following insights: The model accurately classified 301 instances as TP and 1388 instances as TN. It also misclassified 151 instances as FP and 273 instances as FN. With an ROC AUC score of 0.86, the model demonstrates a robust ability to distinguish between the positive and negative classes.

The adaboost model showcases high accuracy and a strong ROC AUC score, rendering it a dependable performer for this classification task. Its relatively low rates of false positives and false negatives underscore its proficiency in correctly identifying both the presence and absence of the target class. In summary, adaboost delivers reliable and efficient classification outcomes, striking a commendable balance between sensitivity and specificity.

4.2.6 Logistic regression model performance

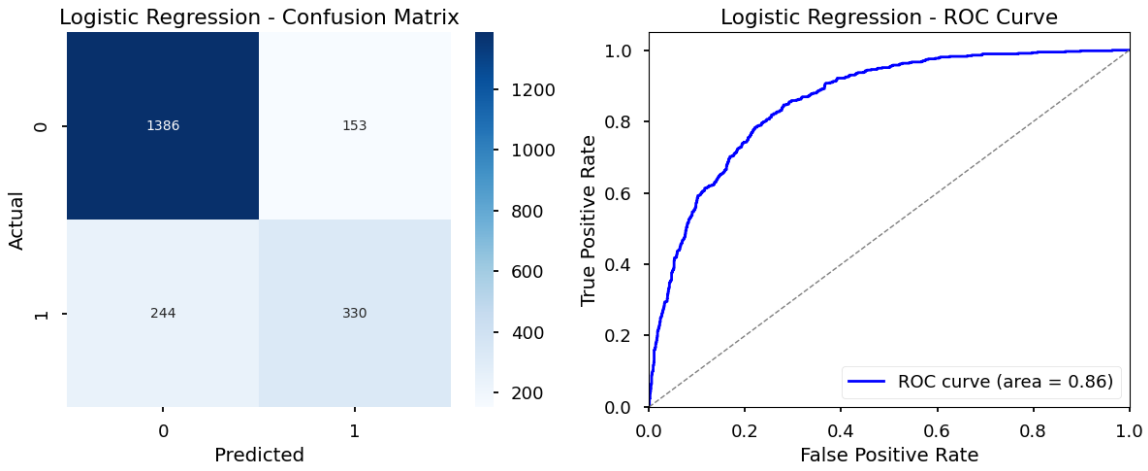

Figure 4.4: Confusion matrix and ROC curve for logistic regression

The analysis of the confusion matrix and ROC curve for the LR model in figure 4.4 reveals the following findings: The model accurately classified 330 instances as TP and 1386 instances as TN. Additionally, it misclassified 153 instances as FP and 244 instances as FN. With an ROC AUC score of 0.86, the model demonstrates a robust capability to differentiate between the positive and negative classes.

The LR model demonstrates high accuracy and a strong ROC AUC score, making it a robust performer for this classification task. The relatively low number of false positives and false negatives highlights the model's capability in correctly identifying both the presence and absence of the target class. Overall, Logistic Regression provides a reliable and efficient classification with a good balance between sensitivity and specificity.

4.2.7 Random forest model performance

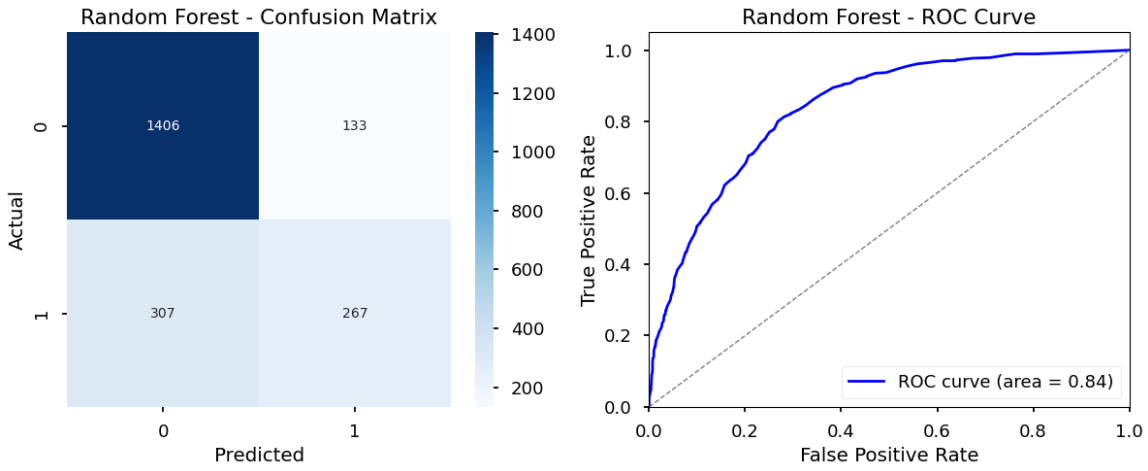

Figure 4.5: Confusion matrix and ROC curve for random forest

The evaluation of the confusion matrix and ROC curve for the RF model in figure 4.5 yields the following observations: The model accurately classified 267 instances as TP and 1406 instances as TN. Furthermore, it misclassified 133 instances as FP and 307 instances as FN. With a ROC AUC score of 0.84, the model exhibits a robust capability to differentiate between the positive and negative classes.

The RF model demonstrates high accuracy and a strong ROC AUC score, making it a robust performer for this classification task. The relatively low number of false positives and false negatives highlights the model's capability in correctly identifying both the presence and absence of the target class. Overall, random forest provides a reliable and efficient classification with a good balance between sensitivity and specificity.

4.2.8 KNN model performance

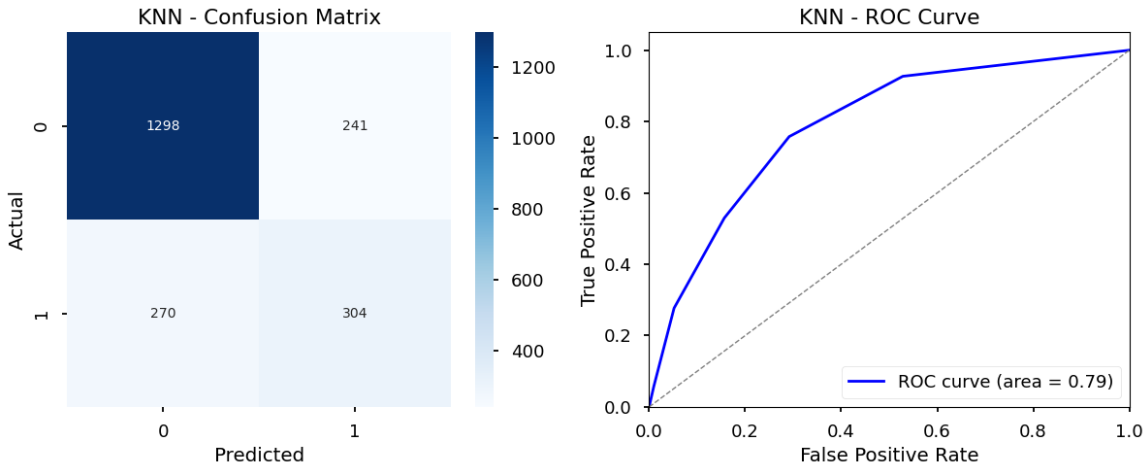

Figure 4.6: Confusion matrix and ROC curve for k-nearest neighbour

The analysis of the confusion matrix and ROC curve for the KNN model in figure 4.6 reveals the following results: The model accurately classified 304 instances as TP and 1298 instances as TN. Additionally, it misclassified 270 instances as FP and 241 instances as FN. With an ROC AUC score of 0.79, the model demonstrates a robust ability to distinguish between the positive and negative classes.

The KNN model exhibits a commendable balance between sensitivity and specificity, as indicated by its strong ROC AUC score, underscoring its overall effectiveness in-class differentiation.

4.2.9 SVM model performance

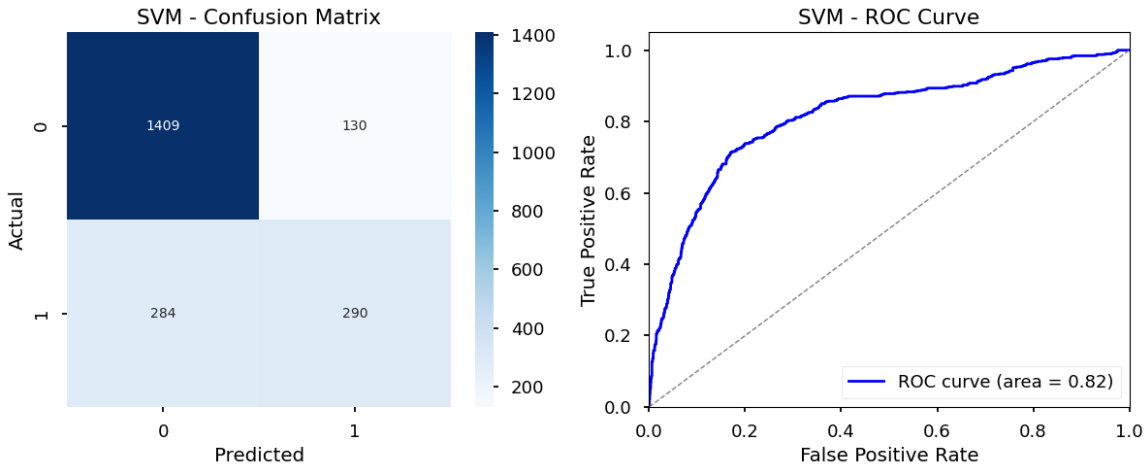

Figure 4.7: Confusion matrix and ROC curve for support vector machine

The examination of the confusion matrix and ROC curve for the SVM model in figure 4.7 reveals the following findings: The model accurately classified 290 instances as TP and 1409 instances as TN. Furthermore, it misclassified 284 instances as FP and 130 instances as FN. With an ROC AUC score of 0.82, the model demonstrates a robust capability to distinguish between the positive and negative classes.

The SVM model demonstrates a commendable balance between sensitivity and specificity, as evidenced by its strong ROC AUC score, indicating its overall effectiveness in class discrimination.

4.3 Summary of the chapter

This section explores the association between customer churn (”Churn”) and various independent variables, utilising metrics such as the matrix, accuracy score, and receiver operating characteristic (ROC) curve to evaluate four machine learning models: LR, RF, SVM,adaboost, xgboost, and KNN.

The primary findings reveal that the adaboost model displayed superior performance compared to other models. It achieved an AUC-ROC of 0.86, accuracy of 0.80, and macro-F1 score of 0.59. Following closely behind is logistic regression, which attained an AUC-ROC of 0.86 and an accuracy of 0.80. The random forest classifier exhibited an AUC-ROC of 0.84 and an accuracy of 0.79, while the support vector machine classifier yielded an AUC-ROC of 0.82 and an accuracy of 0.80. Xgboost classifier similarly presented an AUC-ROC of 0.84 and an accuracy of 0.78. Lastly, the k-nearest neighbors classifier recorded an AUC-ROC of 0.79 and an accuracy of 0.75.

Furthermore, insights gleaned from the literature review indicate that logistic regression, adaboost, and random forest models are frequently acknowledged as the top-performing models for predicting customer churn, as evidenced by their AUC-ROC and macro-F1 scores. Additionally, the results of the chi-square test suggest that gender and phone services do not exert a significant influence on customer churn.

Chapter 5

Conclusion

5.1 Conclusion

Within this study, six models that include LR, RF, SVM, KNN, adaboost, and xgboost were employed to anticipate and compare customer churn, supplemented by descriptive and multivariate analysis to explore factors influencing churn. The research concludes that payment method, senior citizen status, paperless billing status, internet service type, tenure, contract type, monthly charges, and dependents status significantly influence customer churn in the telecommunication sector. Adaboost surpassed support vector machine, logistic regression, random forest and k-nearest neighbors,xgboost models, achieving an AUC-ROC score of 0.86, accuracy of 0.80, and macro-F1 score of 0.59. Findings indicated that customers with month-to-month contracts exhibit higher churn rates, while those with internet service, especially fiber optic, are more prone to churn. Additionally, customers with longer tenures or higher total payments are less likely to churn.

This study made significant contributions to the telecommunication sector and the field of machine learning by employing six machine learning models to predict and compare customer churn, complemented by descriptive and multivariate analyses to identify key factors influencing churn. The research identified critical factors such as payment method, senior citizen status, paperless billing status, internet service type, tenure, contract type, monthly charges, and dependents status as significant influencers of customer churn.

In the context of machine learning, the study demonstrated the superior performance of the Adaboost model over other models, including SVM, LR, RF, KNN, and xgboost, with an AUC-ROC score of0.86%, accuracy of 0.80%, and a macro-F1 score of 0.59%. The findings highlighted that customers with month-to-month contracts and fiber optic internet service exhibit higher churn rates, while those with longer tenures or higher total payments are less likely to churn. These insights provide valuable guidance for telecommunication companies to develop targeted retention strategies and improve customer satisfaction.

5.2 Recommendations

The researcher makes the following insights:
Target younger, individuals because they are more loyal compared to senior citizens. Promote cost-effective services such as streaming Movies (like Netflix specials) and Music, offering longer contract duration, possibly spanning a year or two, since young people enjoy binge-watching series and listening to music.

Provide additional discounts for customers opting for one or two-year contracts to incentives more customers to commit to longer terms. Implement a loyalty rewards program where loyal customers receive complimentary services on select products for a month each year. Emphasise DSL over Fiber optic services

as DSL tends to be more affordable, and customers are less likely to churn than Fiber optic. Consider creating bundle family packages, encouraging younger customers to sign up with their older family members or friends, as younger customers without dependents or partners are more prone to churn.

5.3 Limitations of the study

The research constraint lies in the dataset's size, as machine learning techniques often benefit from extensive training data to enhance performance.

5.4 Future studies

Subsequent studies should focus on employing model hyperparameter tuning to enhance performance and should also explore the utilisation of larger datasets. Logistic regression and random forest are straightforward to implement and can enhance prediction performance. They are also capable of addressing overfitting issues through regularisation techniques. Therefore, I recommend using these methods in future studies.

References

ABU GHAZALEH, M. AND ZABADI, A. M. (2020). Promoting a revamped crm through the internet of things and big data: an ahp-based evaluation. *International journal of organizational analysis*, **28** (1), 66–91.

AHMAD, M., AFTAB, S., MUHAMMAD, S. S., AND WAHEED, U. (2017). Tools and techniques for lexicon driven sentiment analysis. *International Research Journal of Multidisciplinary Technovation*, **8** (1), 17–23.

AMIN, A., ADNAN, A., AND ANWAR, S. (2023). An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and naive bayes. *Applied Soft Computing*, **137**, 110103.

BEEHARRY, R., YOGESH AND FOKONE, T. (2022). Hybrid approach using machine learning algorithms for customers' churn prediction in the telecommunications industry. *Concurrency and Computation: Practice and Experience*, **34** (4), 6627.

BREIMAN, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.

BRITO, J. B., BUCCO, G. B., HELDT, R., BECKER, J. L., SILVEIRA, C. S., LUCE, F. B., AND ANZANELLO, M. J. (2022). A framework to improve churn prediction performance in retail banking. *Financial Innovation*, **10** (1), 1–29.

CARVALHO, T. P., SOARES, F. A., VITA, R., FRANCISCO, R. D. P., BASTO, J. P., AND ALCAL, S. G. (2019). A systematic literature review of machine learn-

ing methods applied to predictive maintenance. *Computers and Industrial Engineering*, **137**, 106–177.

CERVANTES, J., GARCIA-LAMONT, F., RODRIGUEZ-MAZAHUA, L., AND LOPEZ, A. (2020). A comprehensive survey on support vector machine classification: applications, challenges, and trends. *Neurocomputing*, **408**, 189–215.

CHEN, M., LIU, Q., CHEN, S., LIU, Y., ZHANG, C. H., AND LIU, R. (2019). Xgboost-based algorithm interpretation and application on post-fault transient stability status prediction of power system. *Institute of Electrical and Electronics Engineers Access*, **7**, 13149–13158.

CHEN, T. AND GUESTRIN, C. (2016). Xgboost: a scalable tree boosting system. *In Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*. Springer, pp. 785–794.

CHONG, A. Y. W., KHAW, K. W., YEONG, W. C., AND CHUAH, W. X. (2023). Customer churn prediction of telecom company using machine learning algorithms. *Journal of Soft Computing and Data Mining*, **4** (2), 1–22.

DALVI, P. K., KHANDGE, S. K., DEOMORE, A., BANKAR, A., AND KANADE, V. (2016). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. *In 2016 Symposium on Colossal Data Analysis and Networking*. Institute of Electrical and Electronics Engineers, pp. 1–4.

DUFFETT, R. G. AND THOMAS, S. (2024). Health non profit organizations use of social media communication and marketing during covid-19: A qualitative technology acceptance model viewpoint. *Social Sciences & Humanities Open*, **10**, 101042.

EDWINE, N., WANG, W., SONG, W., AND SSEBUGGWAWO, D. (2022). Detect-

ing the risk of customer churn in telecom sector. *Mathematical Problems in Engineering*, **2**, 200–253.

FARENIUK, Y., ZATONATSKA, T., DLUHOPOLSKYI, O., AND KOVALENKO, O. (2022). Customer churn prediction model: a case of the telecommunication market. *Economics*, **10** (2), 109–130.

FARNAAZ, N. AND JABBAR, M. (2016). Random forest modeling for network intrusion detection system. *Procedia Computer Science*, **89**, 213–217.

FERNANDEZ MONTOYA, L., ALAFO, C., MARISOLER, H., MSQUINA, M., MALHEIA, A., SACOOR, C., ABILIO, A. P., MARRENJO, D., CUAMBA, N., GALATAS, B., ET AL. (2023). An evaluation of llin ownership, access, and use during the magnitude project in southern mozambique. *PloS one*, **18** (3), 282209.

FRITZ, M. AND BERGER, P. D. (2015). *Improving the user experience through practical data analytics: gain meaningful insight and increase your bottom line*. Morgan Kaufmann.

FUJO, S. W., SUBRAMANIAN, S., KHDER, M. A., ET AL. (2022). Customer churn prediction in telecommunication industry using deep learning. *Information Sciences Letters*, **11** (1), 24.

GLENDAH, S. T., ALALA, O. B., AND DISHON, M. W. (2019). Social media marketing and brand loyalty at safaricom company, kenya. *European Journal of Business and Strategic Management*, **4** (5), 49–64.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. H., AND FRIEDMAN, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

HUANG, Y. AND KECHADI, T. (2013). An effective hybrid learning system

for telecommunication churn prediction. *Expert Systems with Applications*, **40** (14), 5635–5647.

HUIGEVOORT, C. AND DIJKMAN, R. (2015). Customer churn prediction for an insurance company. *Eindhoven University of Technology*, **11** (3), 45–58.

JAIN, H., KHUNTETA, A., AND SRIVASTAVA, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost. *Procedia Computer Science*, **167**, 101–112.

KAGGLE (2020). Machine learning and data science community. Accessed: January 23, 2023.
URL: *https://www.kaggle.com/*

KERAMATI, A., GHANEEI, H., AND MIRMOHAMMADI, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, **2** (1), 1–13.

KIGUCHI, M., SAEED, W., AND MEDI, I. (2022). Churn prediction in digital game-based learning using data mining techniques: logistic regression, decision tree, and random forest. *Applied Soft Computing*, **118**, 108–491.

KING, B. E. AND RICE, J. (2019). Analysis of churn in mobile telecommunications: Predicting the timing of customer churn. *AIMS International Journal of Management*, **13** (2), 4627.

LAFFONT, J. J. AND TIROLE, J. (2001). *Competition in telecommunications*. Massachusetts Institute of Technology press.

LALWANI, P., MISHRA, M. K., CHADHA, J. S., AND SETHI, P. (2022). Customer churn prediction system: a machine learning approach. *Computing*, 1–24.

LIU, M. (2010). Fingerprint classification based on adaboost learning from singularity features. *Pattern Recognition*, **43** (3), 1062–1070.

LIU, R., ALI, S., BILAL, S. F., SAKHAWAT, Z., IMRAN, A., ALMUHAIMEED, A., ALZAHRANI, A., AND SUN, G. (2022). An intelligent hybrid scheme for customer churn prediction integrating clustering and classification algorithms. *Applied Sciences*, **12** (18), 9355.

LIU, Y., ESAN, O. C., PAN, Z., AND AN, L. (2021). Machine learning for advanced energy materials. *Energy and AI*, **3**, 10–49.

MADUNA, M., TELUKDARIE, A., MUNIEN, I., ONKONKWO, U., AND VERMEULEN, A. (2024). Smart customer churn management system using machine learning. *Procedia Computer Science*, **237**, 552–558.

MASHAU, M. L. (2020). Customer loyalty programmes in the south african banking sector. *Customer churn in telecommunication industry*, **48** (3), 1386.

MOHAMMAD, N. I., ISMAIL, S. A., KAMA, M. N., YUSOP, O. M., AND AZMI, A. (2019). Customer churn prediction in telecommunication industry using machine learning classifiers. *In Proceedings of the 3rd International Conference on Vision, Image and Signal Processing*. Springer, pp. 1–7.

MOLAPO, M. E. AND MUKWADA, G. (2014). Does customer retention strategies matter in the south african cellphone industry. *Mediterranean Journal of Social Sciences*, **5** (23), 144–151.

NAZARENKO, E., VARKENTIN, V., AND POLYAKOVA, T. (2019). Features of application of machine learning methods for classification of network traffic (features, advantages, disadvantages). *In 2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon)*. Institute of Electrical and Electronics Engineers, pp. 1–5.

NUSINOVICI, S., THAM, Y. C., YAN, M. Y. C., TING, D. S. W., LI, J., SABANAYAGAM, C., WONG, T. Y., AND CHENG, C.-Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, **122**, 56–69.

OLUFEMI, O. G. AND STRYDOM, T. (2018). Churn forecasting model for south african pre-paid service providers. *In 7th International RAIS Conference on Social Sciences*. Springer, pp. 110–126.

PARK, H. A. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing*, **43** (2), 154–164.

PATIL, K., PATIL, S., DANVE, R., AND PATIL, R. (2022). Machine learning and neural network models for customer churn prediction in banking and telecom sectors. *In Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems*. Springer, pp. 241–253.

PENG, Z., HUANG, Q., AND HAN, Y. (2019). Model research on the forecast of second-hand house price in chengdu based on xgboost algorithm. *In 2019 Institute of Electrical and Electronics Engineers 11th International Conference on Advanced Information Technology*. Institute of Electrical and Electronics Engineers, pp. 168–172.

PERROTTA, F., PARRY, T., AND NEVES, L. C. (2017). Application of machine learning for fuel consumption modelling of trucks. *In 2017 Institute of Electrical and Electronics Engineers International Conference on Big Data (Big Data)*. Institute of Electrical and Electronics Engineers, pp. 3810–3815.

POUDEL, S. S., POKHAREL, S., AND TIMILSINA, M. (2024). Explaining customer churn prediction in telecom industry using tabular machine learning models. *Machine Learning with Applications*, **17**, 100567.

RAJAMOHAMED, R. AND MANOKARAN, J. (2018). Improved credit card churn prediction based on rough clustering and supervised learning techniques. *Cluster Computing*, **21** (1), 65–77.

SARAN KUMAR, A. AND CHANDRAKALA, D. (2016). A survey on customer

churn prediction using machine learning techniques. *International Journal of Computer Applications*, **975**, 88–87.

SEID, M. H. AND WOLDEYOHANNIS, M. M. (2022). Customer churn prediction using machine learning: commercial bank of ethiopia. *In 2022 International Conference on Information and Communication Technology for Development for Africa*, volume 18. Institute of Electrical and Electronics Engineers, pp. 1–6.

SOARES, P. M., JOHANNSEN, F., LIMA, D. C., LEMOS, G., BENTO, V. A., AND BUSHENKOVA, A. (2020). High-resolution downscaling of cmip6 earth system and global climate models using deep learning for iberia. *Geoscientific Model Development*, **17** (1), 229–259.

SOLTANI-NEJAD, N., VAKILIMOFRAD, H., FAZLI, F., SABERI, M. K., DOULANI, A., AND MAZLOUM, J. (2021). Developing a model to identify the factors contributing to user loyalty of university libraries. *The Journal of Academic Librarianship*, **47** (5), 102386.

SRINIVASAN, R., RAJESWARI, D., AND ELANGOVAN, G. (2023). Customer churn prediction using machine learning approaches. *In 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering*, volume 12. Institute of Electrical and Electronics Engineers, pp. 1–6.

SRITRUSTA, R., UDIN HARUN AL RASYID, M., AND SUKARIDHOTO, A. K. (2019). Lovhealth: loviotech healthcare iot-cloud platform for patient care based on diagnosis system with fuzzy logic and machine learning approach. *International Journal of Computing and Digital Systems*, **15** (1), 1–14.

TANUI, J. K. (2008). *Strategic Responses to Increasing Competitive Challenges in the Telecommunications industry in Kenya-A case of Telkom Kenya Limited*. Ph.D. thesis, University of Nairobi.

TAO, Z., HUILING, L., WENWEN, W., AND XIA, Y. (2019). Ga-svm based feature selection and parameter optimization in hospitalization expense modeling. *Applied Soft Computing*, **75**, 323–332.

UMAYAPARVATHI, V. AND IYAKUTTI, K. (2016). A survey on customer churn prediction in telecom industry: datasets, methods and metrics. *International Research Journal of Engineering and Technology*, **3** (04), 162–170.

WADE, C. AND GLYNN, K. (2020). *Hands-on gradient boosting with xgboost and scikit-learn: perform accessible machine learning and extreme gradient boosting with python*, volume 6. Packt Publishing Ltd.

WADIKAR, D. (2020). Customer churn prediction. *Medicine Journal of Social Sciences*, **3** (27), 144–151.

WAGH, S. K., ANDHALE, A. A., WAGH, K. S., PANSARE, J. R., AMBADEKAR, S. P., AND GAWANDE, S. (2024). Customer churn prediction in telecom sector using machine learning techniques. *Results in Control and Optimization*, **14**, 100342.

XIAHOU, X. AND HARADA, Y. (2022). B2c e-commerce customer churn prediction based on k-means and svm. *Journal of Theoretical and Applied Electronic Commerce Research*, **17** (2), 458–475.

ZHAO, W., LI, J., ZHAO, J., ZHAO, D., LU, J., AND WANG, X. (2020). Xgb model: research on evaporation duct height prediction based on xgboost algorithm. *Radioengineering*, **29** (1), 81–93.

ZIEGLER, A. AND KONIG, I. R. (2014). Mining data with random forests: current options for real-world applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **4** (1), 55–63.

Appendix

Some selected Python codes

churndata = pd.readcsv("WA$_Fn - UseC_Telco - Customer - Churn.csv$")

churndata.info()

churndata.head()

Python codes for fitting logistic regression

lr = LogisticRegression()

lr.fit(Xtrain,ytrain)

pred = lr.predict(Xtest)

Python codes for fitting Random Forest model

rf = RandomForest()

```
rf.fit(Xtrain,ytrain)


pred = rf.predict(Xtest)
```