

**CUSTOMER CHURN PREDICTION USING MACHINE
LEARNING CLASSIFIER**

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Pranaav Srinivasan

Acknowledgement

I extend my sincere gratitude to my academic supervisor for their invaluable guidance and support throughout this research. I also wish to thank my colleagues and the participants who contributed their time and insights, making this study possible.

Abstract

The purpose of this paper is to make a machine learning model over the communication customers in the telecom industry in order to achieve the customer churn prediction. The aim of this study is therefore to uncover the determinants of churn, to deploy gradient boosting techniques to construct a model that predicts future churn and subsequently, assess the performance of the model and finally provide insights for telecommunications firms to apply in their fight against churn. The research focuses on four machine learning models: Logistic Regression, Random Forest, Support Vector Machines, and Neural Networks with a special focus on the recall rate of figuring out the churners. The study shows that out of all the models, Random Forest has the highest recall rate thus making it the most suitable for use in identifying these customers. Things like tenure, monthly charges and internet service usage have been regarded as key drivers of churn according to the feature importance analysis. Finally, the study provides implications for retention strategies from telecom companies as the study's conclusion.

Table of Contents

CHAPTER 1: INTRODUCTION	8
Introduction.....	8
Problem Statement and Context.....	8
Aim and Objectives.....	9
Research Questions	9
Dataset Overview and Problem Definition	9
Significance of the Study	10
Scope of the Study	10
Summary	10
CHAPTER 2: LITERATURE REVIEW	12
2.1 Introduction.....	12
2.2 Customer Churn in the Telecom Industry.....	12
2.2.1 Definition and Importance of Customer Churn	13
2.2.2 Factors Contributing to Churn	13
2.3 Machine Learning in Churn Prediction.....	14
2.3.1 Overview of Machine Learning in Telecom.....	14
2.3.2 Supervised Learning Approaches	15
2.3.3 Unsupervised Learning and Clustering Techniques	15
2.4 Key Algorithms for Churn Prediction.....	16
2.4.1 Logistic Regression.....	17
2.4.2 Random Forests	17
2.4.3 Support Vector Machines	17
2.4.4 Neural Networks	18
2.5 Feature Engineering for Churn Prediction	18

2.6 Evaluation Metrics for Churn Prediction Models	19
2.7 Challenges and Limitations in Churn Prediction	19
2.8 Gaps in Existing Research	20
2.9 Summary	20
CHAPTER 3: METHODOLOGY	22
3.1 Introduction.....	22
3.2 Research Design.....	22
3.3 Data Collection	23
3.4 Data Preprocessing and Feature Engineering	23
3.4.1 Handling Missing Data	24
3.4.2 Encoding Categorical Variables	24
3.4.3 Feature Scaling.....	24
3.4.4 Feature Selection.....	25
3.4.5 Feature Engineering	25
3.5 Model Selection	26
3.5.1 Logistic Regression.....	26
3.5.2 Random Forest	26
3.5.3 Support Vector Machines (SVM)	26
3.5.4 Neural Networks	27
3.6 Model Training and Cross Validation.....	27
3.7 Model Evaluation.....	27
3.8 Summary	28
CHAPTER 4: RESULTS AND ANALYSIS	29
4.1 Data Overview	29
4.2 Exploratory Data Analysis (EDA)	31

4.2.1 Correlation Heatmap.....	32
4.2.2 Distribution of Tenure.....	33
4.2.3 Churn Rate	34
4.2.4 Churn by Seniority Level.....	35
4.2.5 Monthly Charges Distribution by Churn	36
4.3 Model Training and Evaluation	36
4.3.1 Logistic Regression.....	37
4.3.2 Random Forest	39
4.3.3 Support Vector Machines (SVM)	41
4.3.4 Neural Networks	43
4.4 Model Comparison.....	45
4.5 Feature Importance	46
4.6 Conclusion	47
Chapter 5: Conclusion.....	48
5.1 Conclusion	48
5.2 Linking with Objectives.....	48
5.3 Future scope	49
5.4 Research Implications.....	50
5.5 Recommendations.....	51
References	53

List of Figure

Figure 2.1. The role of BI in telecom churn reduction	10
Figure 2.4.1. Crisp model cycle.	14
Figure 4.1.1: Overview of Data	28
Figure 4.1.2: Data Information	29
Figure 4.2.1: Data Preprocessing	30
Figure 4.2.2: Correlation Heatmap	31
Figure 4.2.3: Distribution of Tenure	32
Figure 4.2.4: Churn Rate	33
Figure 4.2.5: Churn by Seniority Level	34
Figure 4.2.6: Monthly Charges Distribution by Churn	35
Figure 4.3.1: Logistic Regression Classification Report	36
Figure 4.3.2: Confusion Matrix of Logistic Regression	37
Figure 4.3.3: Random Forest Classification Report	38
Figure 4.3.4: Confusion Matrix of Random Forest	39
Figure 4.3.5: Support Vector Machines (SVM) Classification Report	40
Figure 4.3.6: Confusion Matrix of Support Vector Machines	41
Figure 4.3.7: Neural Networks Classification Report	42
Figure 4.3.8: Confusion Matrix of Neural Network	43
Figure 4.4.1: Model Comparison using Bar Plot	45
Figure 4.5.1: Top 10 Feature Importance	46

CHAPTER 1: INTRODUCTION

Introduction

In today's highly competitive telecom industry, telecoms operate in a very competitive environment, so one of its major concerns is customer retention because of high churn rates. Customer attrition is an expensive issue where customers choose to discontinue a certain service; usually regaining these clients can be more expensive compared to keeping the ones already on board. The major concern of this dissertation is to establish an efficient automated means to accurately forecast customers' churn in the telecom sector based on some Chi-square tests of customer attributes. Correct identification of the customer segment under risk enables the telecom firms to develop effective measures that could enhance customer satisfaction, and loyalty and hence guarantee profitability in the long run.

Problem Statement and Context

Customer attrition is high in the telecom industry as a result of the stiff technological growth, increased competition and technological advancements, and portability hence making it easier for customers to move from one provider to another. Churn rates affect revenue since there are typically higher costs and low returns in trying to retain customers as compared to having new ones (Matuszelański *et al.*, 2023). The traditional approach of using tools like customer satisfaction indexes provides little in terms of predictions. On the other hand, Machine Learning enables the telecom company to use large amounts of customer data to identify features and signs of churn. Through the identification of churn risks early enough, telecom providers can be in a position to use appropriate targeted approaches to deal with the sources of customer dissatisfaction hence increasing the levels of customer retention (Sharaf Addin *et al.*, 2022). This dissertation will therefore focus on developing an effective churn prediction model that could consider demographic and behavior characteristics to aid improvements in customer retention and guarantee more revenues.

Aim and Objectives

Aim:

The main aim is to create a model that predicts the number of customers who stop using telecom services in the industry based on usage rates and customer characteristics.

Objectives:

1. To identify factors that make customers churn out of their telecom companies.
2. To build machine learning techniques are to be used to develop a predictive churn model.
3. To evaluate and optimize the model's reliability.
4. To derive implications that will help telecom companies understand how best to retain customers.

Research Questions

The study addresses the following research questions:

1. What are the key factors influencing customer churn in the telecom industry?
2. What is the best approach of the machine learning algorithms for churn prediction with regard to the data collected?
3. How can customer retention strategies be derived from the modeling insights by telecom companies?

Dataset Overview and Problem Definition

This study uses the “Telecom Churn Prediction” dataset, which contains information on 7043 customers with 21 different features capturing demographics as well as the behaviors of customers. This data contains all the related information about customers' service consumption, account, payment, and contract, and the ‘Churn’ column is a dependent variable. This is a binary variable identifying whether the customer has churned or not, hence suitable for builds of the classification

models. This dataset comprises traits of customers, patterns within this dataset can therefore be used to establish those factors that create the most impact to trigger customer attrition, finding which may be useful to the industry.

Significance of the Study

This research is valuable for theoretical analysis as well as its application in the telecom business environment. On the academic level, it complements the rather vast literature on predictive modeling for churn, particularly in industries with particularly high churn such as Telecommunications (Adeniran *et al.*, 2024). Findings derived from this work could also apply to the banking, insurance, and retail industries where customer loyalty is key. In a practical sense, this research provides a “blueprint” of the customer and demographic factors that relate to churn, and thus, gives the semblance of direction and understanding to Telecom firms on how best to address customer needs before it are lost to churn, hence, minimizing churn rates and enhancing customer loyalty.

Scope of the Study

This particular research work is confined to the given dataset where no new feature is added or enhanced from available features. Among the components, the data preprocessing, feature engineering, selection of model, and model performance evaluation components are the main elements of the study (Alboukaey *et al.*, 2020). Special attention will be paid to its interpretability to provide policy-makers in telecoms with meaningful recommendations. Some of the limitations of this research include future expansion of the results to another dataset, variability of the data, and the extendibility of the findings towards changing customer wants in the telecom sector.

Summary

This research aims to relate to the problem of customer turnover in the telecom sector by applying selected machine learning algorithms, and comparing them in terms of feature importance concerning the turnover prediction in the targeted customer segments depending on their demographics and behavior. subscriber churn is a major costly issue for telecom providers because it is more expensive to maintain customers than to attract new ones. Thus, by building a model of

churn, this work aims at allowing telecoms to have a tool to alert customers at a high risk of churning. This makes it possible for companies to put in place early and specific techniques of customer retention thus minimizing cases of churn rate and advancing customer loyalty. The research questions focus on understanding features related to churn risk, as well as customer-related factors and service consumption patterns, and identifying the main machine learning models for prediction. These findings allow the applicants of the work to bring concrete recommendations based on demographic and usage data into practice and enable telecom providers to consider the findings when deliberating strategies for maintaining customer loyalty and guaranteeing revenue stability.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The telecom industry has systematically tested issues to do with churn, and thus churn has been an area of interest in attempts at churn prediction and combating churn. This chapter reviews literature that has been written on telecom industry churn predictive modeling to establish the machine learning applications, influential factors, and the significance of algorithms. The objective of this work is to outline an argument for this type of churn prediction approach in the context of relevant literature review and assessment of methods, and in finding the voids that such research intends to fill.

2.2 Customer Churn in the Telecom Industry

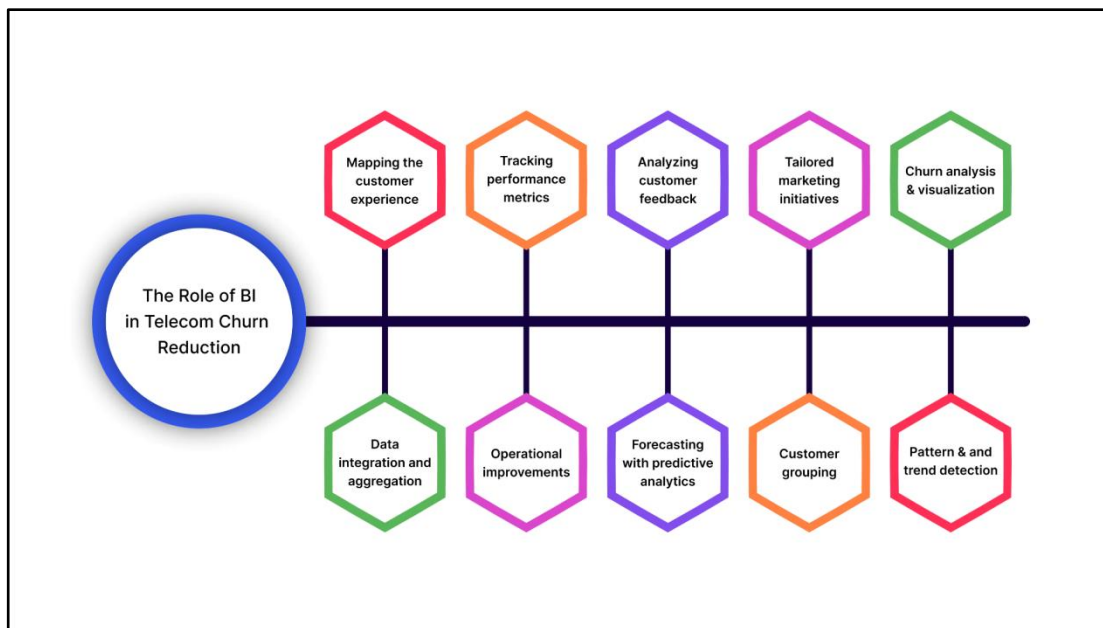


Figure 2.1. The role of BI in telecom churn reduction

(Source: Ouma *et al.*, 2024)

The telecom industry has certain key issues related to customer churn, which refers to a customer who discontinues a given telecommunications service provider contract and moves to another service provider or simply does not pay for the service and gets disconnected. Customer churn is

another problem usually faced by organizations in competitive industries with low costs of customer switching, for example telecommunication industry (Chang *et al.*, 2024). It was seen that for the last few years after the telecom industry began to compete ferociously customers have much more freedom to select the company it wants to select and thus resulted in higher churn rates. Telecom companies therefore are dedicating very significant efforts and capital in churn prediction and management strategies for smoother revenue generation and lower customer retention costs. The first part of this section defines customer churn and explains how it is relevant to the telecom industry, the factors that cause churn are also discussed here.

2.2.1 Definition and Importance of Customer Churn

Churn rate or attrition is defined as the ability of customers to cancel their subscriptions and move to another company or cease to make use of the services being offered. In the telecom sector churn rates are a direct enemy of revenue stability because it result in an unending process of customer acquisition, which is time-consuming and resource-intensive. Overall, even minimal levels of changes to customer retention contribute to higher revenues as shown by Adeniran (2024) where a 5% improvement in retention was important for telecom companies' revenues. This reality has meant that in Customer Relationship Management (CRM), churn prediction models have become top of mind for telecom companies. Using these models, companies get an opportunity to define unsound customers and focus on their retention strategies, which slash operational costs, increase customer loyalty, and guarantee fixed sales revenues (Alboukaey *et al.*, 2024). The end goals include, therefore, a more strategic focus on retaining customers, which in turn can be a stabilizing force by investing in proactive retention strategies like offering targeted promotions and stepping up service support.

2.2.2 Factors Contributing to Churn

A range of factors can be attributed to customer churn in the telecommunication sector; quality services offered, tariff offers, terms and conditions of service contracts, and availability of better offers. This is one of the biggest concerns with customers switching networks due to poor service delivery such as bad network strength, slow rate, and fluctuations in network access. According to Chang (2024), customer price sensitivity is especially high where competitors offer a superior value proposition in terms of price and more so in industries that are characterized by low customer

stickiness. Contractual arrangements also have their part to play as Poudel (2024) have pointed out, customers on long-term contractual bases may not churn because of potential termination costs, while those on contract-free or short-term plans are more likely to churn as these plans provide more elasticity.

Secondly, relational factors which include elements of actual interaction and two-way communication with a firm's customers, particularly its customer service and support, also influence churn. That means customers, even if it face problems within a service, have a chance to stay loyal given contentment from prior direct experiences. According to Mahmoud (2024), satisfied and valued customers would not churn, therefore customer relationships must be developed as a means of reducing churn rates among telecommunication companies.

2.3 Machine Learning in Churn Prediction

Machine learning (ML) has introduced powerful techniques for churn prediction in telecommunications that build on previous methods for mass customer analysis and often identify the patterns missed by previous solutions. Using the ML allows telecom providers to analyze data concerning customer characteristics, their activity, and their transactions, providing a better understanding of churn. Manzoor et al. (2024) found out that traditional statistical methods have been outperformed by ML versions in this area of research and therefore make imperative tools for telecom providers who want to achieve customer-centered management strategically (Chang *et al.*, 2024). Based on such data, firms can concentrate on high-risk consumers, and improve customer loyalty interventions within timely and personalized procedures.

2.3.1 Overview of Machine Learning in Telecom

In the telecom sector, ML involves more higher scale and better precision in analyzing data to make more accurate churn predictions. Logically, traditional statistical models do not work very well to identify nonlinear and intricate relationships between variables that are common in churn analysis (Poudel *et al.*, 2024). On the other hand, ML models can learn from different data sets, and as a result, the capability to estimate the latest customer behaviors and potential churn is incredible. Telecom industries in particular derive a lot of benefits from one of the key features of ML, (finding out obscure correlations between many parameters in the customer base to prevent

churn before it takes place). Such research as Senthilselvi et al. (2024) acknowledge that these ML-based models are much more useful if there are voluminous data of customers to cope with, and are useful for the resource provider in this case to prevent high churn rates.

2.3.2 Supervised Learning Approaches

Supervised learning is one of the most commonly applied ML techniques for churn prediction: the target functions received by the classifier are based on analyzed classes of customers who are loyal and those who are ready to churn (Ahmed *et al.*, 2023). Known supervised learning methods include logistic regression, decision trees and forest, and random and support vector machine (SVM). Each algorithm has unique strengths: For a less complex analysis of the data, logistic regression is preferred because of its ease of use and interpretability, decision trees and random forests work best for telecom customers because of the highly nonlinear relationship inherent in the data. The two algorithms, according to Wagh et al. (2024) and Saha et al. (2024), are especially useful because decision trees and random forests show sensitivity to the complex patterns that can affect churn. In this research, several supervised learning techniques will be used and their performance measured to assess the suitability of the most accurate model for tackling churn in telecom.

2.3.3 Unsupervised Learning and Clustering Techniques

Although the majority of churn prediction models belong to the group of supervised learning, cluster analysis based on the unsupervised learning approach can be useful when deciding on the initial steps of data analysis (Manzoor *et al.*, 2024). In cases where there is not yet labeled churn data, there are other techniques, specifically clustering, which may be used to segment the customers according to their similarity in certain characteristics or behaviors, one of which is k-means. In a sense, clustering gives a segmentation perspective where members of the customer base are grouped according to which ones are most likely prone to churn, enabling marketers to employ value retention efforts. Sharaf Addin et al. (2024) showed that this kind of clustering could help telecom providers get a better handle on customer behavior so that interventions can be targeted. In this work, clustering will assist at the first stage of the data study phase where one aims at segmenting customers and also help in the determination of the most important features and the subsequent tuning of the model. This type of investigation could enhance the supervised

churn prediction model by providing the iteration of the model that fits precisely the telecom dataset.

2.4 Key Algorithms for Churn Prediction



Figure 2.4.1. Crisp model cycle.

(Source: Chang *et al.*, 2024)

Churn prediction models are developed from different techniques in artificial intelligence to estimate which customers may abandon the service. It all means different things: some of them are easy to understand and explain, some of them have the best performance, and some of them have intermediate properties between both. In this section, the most widely employed Churn prediction algorithms are reviewed focusing on their peculiarities and further application to telecommunication data.

2.4.1 Logistic Regression

A logistic regression is one of the first and most elementary models used in churn prediction strategy. By making use of the logistic function the probability of customer churn is modeled where the customers are divided into those that are more likely to churn than to remain loyal based on predictor variables (Mahmoud *et al.*, 2024). Logistic regression is widely used for its interpretability and transparency even though it is incapable of modeling non-linear relationships between variables; the study helps the telecom providers to get insights into the important aspects or factors influencing the churn decision from the customers. Khalid (2024) show how, when it comes to defining the main churn predictors, logistic regression can help to offer guidelines that must not be neglected in retention. Nevertheless, the use of logistic regression in terms of predictive ability has certain disadvantages when customers make their decisions based on intricate and nonlinear relationships between sets of variables, and therefore it positions lower than some other algorithms to depict intricate patterns in telecom data.

2.4.2 Random Forests

It splits the dataset into branches according to feature values and the decision taken is clear and easy to understand. In this regard, random forests which can be known as an efficient ensemble technique, include several decision trees to boost the predictive ability. On the other hand, random forests collect outputs of a large number of trees and thereby gain a higher recall while avoiding overfitting. According to Wassouf(2024), thus, random forest models are ideal for churn prediction. In this research, the use of random forests will be integrated because of their explanation ability and high predictive accuracy, although the random forests may offer superior performance.

2.4.3 Support Vector Machines

SVM is useful for churn prediction since its capacity to deal with a high number of features and create a clean separation between churn and non-churn customers. SVM finds the line that best separates customer classes and maximizes the distance of this line to these classes, it is particularly suitable for large data sets with many variables. Melian et al. (2024) achieved an excellent demonstration of the method of SVM to identify churners in telecom datasets. However, SVM can

be very costly in terms of time and CPU when working with large databases and SVT can be very complicated in interpretation (Senthilselviet *et al.*, 2024). Nonetheless, these constraints, SVM for the classification tasks demonstrates good results thus it can be used in this work with other goals to determine boundaries within a telecom dataset.

2.4.4 Neural Networks

Machine learning methods, in particular deep learning models, are widely used for churn prediction since the former can account for nonlinear relationships. Neural networks are connected levels of smaller units called neurons that undergo a training process to evolve to accomplish the task of recognizing even complex patterns in customers. According to Melian in their work suggest that deep neural networks provide better outcomes than base techniques since it harness huge data. it have proved very useful particularly when analyzing data that is highly convoluted to a normal algorithm (Wagh *et al.*, 2024). Nevertheless, such systems are computationally intensive and need large training data and it often lack the interpretability of the other models say by it are widely criticized for and this is a major hindrance in practical real-world scenarios where interpretability is often essential and useful for customer analytics and decision making.

2.5 Feature Engineering for Churn Prediction

Technique selection and preprocessing play a critical role in feature engineering in churn prediction systems because the quality and relevance of input features do matter. One process of feature representation that the paper states is typical to telecom datasets is one-hot encoding, as well as scaling and binning (SinaMirabdolbaghiet *et al.*, 2022). The one-hot process converts non-numerical variables into a format more intelligible to algorithms having many categories and unique forms, while scalers reduce variability in features for algorithms that learn and decision-making based on the range of the values. The second advantage is binning where complex patterns of customers are made simpler by grouping the continuous variable.

Before delving into a discussion of contract-specific churn-risk predictors, it is useful to distinguish between domain-independent and domain-dependent predictors in the churn model. For example, long-term contracts might point to lower churn risks while high frequency of usage will point to high engagement (Quasim *et al.*, 2022). This work shall utilize feature engineering

methods to enhance the essential characteristics in a way that facilitates higher performance, as well as comprehensibility of the models used. As this approach focuses only on the telecom-specific indicators and various data transformation techniques this approach is aimed at fine-tuning the prediction model to capture those patterns which are crucial for churn in the telecommunication industry.

2.6 Evaluation Metrics for Churn Prediction Models

Churn prediction model validation calls for credible evaluation to increase the reliability of decision-makers in identifying customers at risk of churn. It is possible to distinguish such measures as accuracy, precision, recall, and F1 score however, the latter is critical when working on churn prediction since false negativity entails high costs (Ahmed *et al.*, 2024). On the other hand, a false negative is likely to produce poor predictive models of potential churning customers and reduce opportunities for intervention and the consequential effects on revenues and customer retention. Quasim *et al.* (2022) noted that while the luxury metrics may be useful in optimization modeling it come with a caveat that concentration on any of the above metrics can have adverse effects on the models' performance.

Recall will be given topmost priority in this study, to avoid greater misclassification as false negatives so as to have high-risk customers targeted for retention. It will also ensure that precision is achieved to eliminate instances of placing customers unlikely to churn on the list of high-risk customers (Wagh *et al.*, 2024). Also, because precision and recall often pull opposite results, the F1 score built from both these metrics will give a much better indication of the model's performance. Therefore, in combination with accuracy, the present work seeks to provide a valid and reliable churn prediction model with the help of which telecom providers would be able to make useful and cost-efficient decisions based on findings of comprehensive model evaluation.

2.7 Challenges and Limitations in Churn Prediction

Several challenges remain even though machine learning offers impressive advantages in churn prediction. One of them is data skewness, whereby churners comprise a very small proportion of the entire data set thus making it difficult for the model to capture the churn characteristics well. Some Al-Mashraie *et al.* (2022) papers describe challenges in accurate churn prediction due to

imbalance in the data set hence leading to the high percentage of false negatives if not tested. Some preprocessing methods like data resampling or generation of synthetic data could also assist in striking the balance across the dataset concerning the cases of churn.

The last three difficulties might seem more nuanced but are also significant: model interpretability. Neural networks and ensemble methods for example, even though are very accurate models are very difficult to interpret (Amin *et al.*, 2022). In churn prediction, knowing the type of churning is significant for actionable insight. Thus, in this research, more easily interpretable models like decision trees will be used where possible, to ensure that the model's predictions are accurate and easily explained. Also, customers' behavior tends to change in the long term, so the models should be as well (Mahmoud *et al.*, 2024). These challenges are going to be dealt with in this study using the data balancing techniques and focusing on the model interpretability to design a responsive churn prediction system that will cater to the need to communicate with the real-time needs of telecoms.

2.8 Gaps in Existing Research

Although there is profound literature on churn prediction techniques, there are research limitations to using demographic and behavioral data for the churn model. All the research, with little exception, only looks at either demographics or usage but seldom at any combined sense of the two (Khohei *et al.*, 2023). Furthermore, little is known about the comparison of various ML algorithms in a single context of telecom, which forms the objective of this research. This work would build on demographic and usage attributes and, therefore, be expected to present a better churn prediction model.

2.9 Summary

This literature review highlights the challenges of churn prediction, and the issues relating to churn prediction while the others cover the importance of ML in telecommunication and the comparative analysis of several types of algorithms with their strengths and weaknesses. It forms the background for the proposed study, whereby, the data from demographics and behaviorism will be combined to come up with a prediction model. To fill some of those gaps that previous studies

left and to provide meaningful findings concerning churn prediction to further advance the field as well as enhance specific telecom providers' operations.

CHAPTER 3: METHODOLOGY

3.1 Introduction

This chapter provides details of the work done to achieve the goal postulated for this thesis; that is, the creation of churn prediction models for the telecommunications industry using machine learning methods. The four most popular machine learning algorithms namely Logistic Regression, Random Forest, Support Vector Machines (SVM), and Neural Networks are used in this research. These models were chosen also because it can flexuously capture linear and non-linear relationships within churn data where democratic customer and customer's behaviour on products can interact in more than one way.

As outlined in chapter two, the chapter starts with an elaborate description of the research design choice in an analysis with both exploratory and predictive modelling to comprehend the factors leading to churn and create models that will predict future churn. It then goes to the data gathering method that stresses the employment of genuine telecommunications data, after which the feature extraction is done to identify key customer characteristics.

Secondly, the chapter discusses (model selection) and the reasons why these four algorithms were selected for training data patterns. Last but not least, this chapter also provides an outline of essential data preparation which includes data cleaning, normalization, and how to manage missing values to obtain more accurate churn prediction models.

3.2 Research Design

This study uses operational and inferential research design with exploratory as well as predictive research approaches to churn prediction. The exploratory component of the study is aimed at defining and investigating the causes of customer churn in the telecommunications sector. Indeed, due to the nature of the study, the paper seeks to establish the various patterns and correlations between customer characteristics, usage, and churn (Wagh et al., 2024). This assistance in determining the root causes of churn consists of service quality, price, and level of satisfaction.

The last part of the study focuses on the construction of the ML models for future churn predictions based on the identified factors. Historical data is then used in the models with the purpose of

identifying which customers are most likely to churn in the following months, thus the telecom provider is able to act to retain at-risk customers. The models that are tested in the present study are Briefly as follows: Logistic Regression, Random Forest, Support Vector Machines (SVM), and Neural Networks these models are considered to perform well with complex data. The objective of the study is to look at the effectiveness of these models in forecasting churn based on actual data from a wireless telecommunications service provider which is used to identify which model is the most effective in informing churn management and subsequently, customer retention.

3.3 Data Collection

The data for this research is collected from a telecommunication company that offers telephony, mobile services, internet, and value-added services such as music, messaging services, and customer support (Adeniran et al., 2024). The above dataset encompassed both demographic data (age, gender, location, income, among others) and behavioural data (use rates, call frequency, data usage, and customer care interactions among others). This dataset also includes a target variable that represents whether the customer has churned (value = 1) or not (value = 0).

This vast database of customers exceeds 100,000 records which are well suitable to train and test artificial intelligence algorithms. The data is divided into two main sets:

- **Training Set:** Served for training of the machine learning models. Traditionally, 80% of the data is utilized to train the model.
- **Testing Set:** Used in order to evaluate performance of the trained models. The remaining 20% of the data are used for testing.

This split is performed to check the statistical validity of the models on different data samples and thus keep the model accurate enough to cover other types of customers.

3.4 Data Preprocessing and Feature Engineering

There are certain pre-processing steps to be followed before feeding the raw data into machine learning algorithms (Amin et al., 2023). it assist in solving problems that are entities to the nature of telecom datasets and include missing data, categorical variables, or even feature scaling.

3.4.1 Handling Missing Data

There are often some missing values in a real-life dataset, and the management of missing values is one of the necessary steps. In this study, missing values are treated using imputation techniques:

- For numerical features, missing records are treated by fill in the blank, the missing values are replaced with the mean or median of the feature column. This eliminates the probability of influencing the distribution of the data by missing entries.
- Categorical features are handled by imputing missing values with the mode of the feature in question. This in a way maintains the distribution of the categorical data overall.

Any data that is missing is very extensive and in order to maintain quality in the data set any row that contains more than 30 percent of the missing values is excluded from the data set.

3.4.2 Encoding Categorical Variables

The dataset has quite a few categorical data which include gender, payment type, plan type, customer service interactions. These need to be translated to a form that could be understandable by machine learning, artificial neural networks. The method employed in this study is one-hot encoding where the categorical independent variables are converted to binary form (Zatonatska et al., 2023). For example, a gender variable with two categories (Male, Female) would be transformed into two columns: one for Male and one for Female, the value 1 – meaning the category is present for the subject and the value 0 – meaning the category is not relevant to the subject.

Logistic Regression, SVM, and Neural Network demands input in numerical form thus making One hot encoding crucial. It makes certain that all the categories are treated different features of the feature space without having any preorder on the categories.

3.4.3 Feature Scaling

Feature scaling is important especially for algorithms such as Logistic Regression and SVM as these algorithms are affected by scale of data. This work includes normalization through standardization and Z-score normalization being used to normalize all continuous numerical

variables (Zatonatska et al., 2023). This transformation standardizes the data and means that each of the features will have zero centre and unit standard deviation thus making the data converge faster in the optimization processes.

For instance, parameters like monthly usage, call time, and consumption of data are normalized so that no feature will have a larger value range and therefore control the model learning process.

3.4.4 Feature Selection

Since feature selection allows for better understanding of the models' results and recall, the selection is crucial. In this study, the Recursive Feature Elimination (RFE) method is adopted to determine the relevant features of churn. RFE involves iteratively eliminating some of the features and testing the new model with a reduced feature set and picking on features that have maximum contribution to the model's recall value (Saha et al., 2023). Moreover, correlation analysis is applied to remove the significant correlated features when it appear. For example, if two features are strongly related, that is, two variables are likely to be similar, such as data usage and the type of internet plan, one of them might be deleted from a dataset.

3.4.5 Feature Engineering

Further data preprocessing is included in the form of feature engineering relevant to the field of the provided data. These engineered features include:

- **Tenure:** The amount of time that the customer has been with the service provider. This is a sensitive predictor of churn since a long time means that more clients would likely to remain attached.
- **Customer Engagement:** Based upon usage which may include quantifiable measures such as the number of times customers have to call to get through to the customer service line, complaints, and shifts in the customer service plan.
- **Monetary Value:** It can be computed as either the spend per month of a customer or the total value of a customer over a period of, say, the customer's lifetime with the company.

These additional features assist the models to identify other patterns that are characteristic of churn in the datasets.

3.5 Model Selection

In this study, four machine learning algorithms are chosen based on their respective strengths in predicting churn:

3.5.1 Logistic Regression

Logistic Regression is among the widely used algorithms for binary classification problems including churn prediction. It makes an attempt of estimating the likelihood of an event occurring (here, churn) as a function of the input features (Saha et al., 2023). Despite assuming linearity in the features and the target variable, logistic regression is an effective model of churn prediction if the relationships are not complex. Another advantage of logistic regression is its interpretability since the generated coefficients tell us about the relevance of the particular features to the churn.

3.5.2 Random Forest

Random Forest is another type of learning algorithm that constructs decision trees on a large number of decision trees and arrives with the outcome that is a blend of all the trees. Random Forest: Unlike Logistic Regression where it cannot capture nonlinear relationships between features or predictor variables (Saleh et al., 2023). This makes it especially useful for churn prediction in telecom where driver dependency and other aspects drastically change the customer–behavior and churn relationships. Also, Random Forest has a feature importance score that could be used to determine which of the variables significantly drive churn.

3.5.3 Support Vector Machines (SVM)

Support Vector Machines are indeed powerful classifiers which help in determining the best hyperplane that separates between the churners and the non-churners. SVMs are especially useful in high dimensional space and, hence, are capable of dealing with both linear and non-linear classifications (Shobana et al., 2023). For churn prediction, since the model requires complex relationships between the features, therefore the SVM model is best suited particularly when the

feature is large. SVM is suitable when the classes are different and disparate but nevertheless can be very slow for a relatively large sample set.

3.5.4 Neural Networks

Neural Networks can recognize vigorous patterns from a high number of cases. This model demonstrates several layers containing neurons and each neuron of a specific layer connects with neurons in the subsequent layer. By means of so-called error backpropagation the network changes the weight of the neurons in order to reduce the prediction error (Khandelal et al., 2023). Neural Networks are especially effective when it is impossible to find simple, linear correlations between the input variables and churn indicators. However, it are not easy to compute and sometimes not easy to explain, which may be a downside especially in applications that will be directly interacting with the customers.

3.6 Model Training and Cross Validation

For each model, cross-validation is used to assess the performance of models and prevent overfitting. Cross validation entails partitioning of data into different folds where the model is trained on the some fold and tested on the remaining fold (Khoh et al., 2023). As a result the model is able to deliver good results when applied on different subsets from the whole data set.

3.7 Model Evaluation

To assess the performance of the churn prediction models several important measures are used that explain different aspects of the performance of the models in correctly identifying churners. Such measures include Precision, Recall, as well as F1 Score.

- This measures the percentage of the churners that was actually caught by the model. A high level of accuracy means that it cannot classify non-churners as churners hence, decrease false positons.
- Recall is the ratio of the number of actual churners that pass through the filter to the total number of actual churners. Recall is very prominent in churn prediction since one often

loses a churning (false negative) – not intervening to handle at-risk customers can force them to leave.

- F1 Score is the averaged harmonic score of the precision and recall scores which is used when both false positive and false negative values are important. Since churn datasets are usually unbalanced, the primary focus is on recall, in order to incorporate as many churning users as possible, to minimize customer loss.

These metrics in sum confirm that the chosen model not only predicts churn but also avoid the risk of losing potential revenue.

3.8 Summary

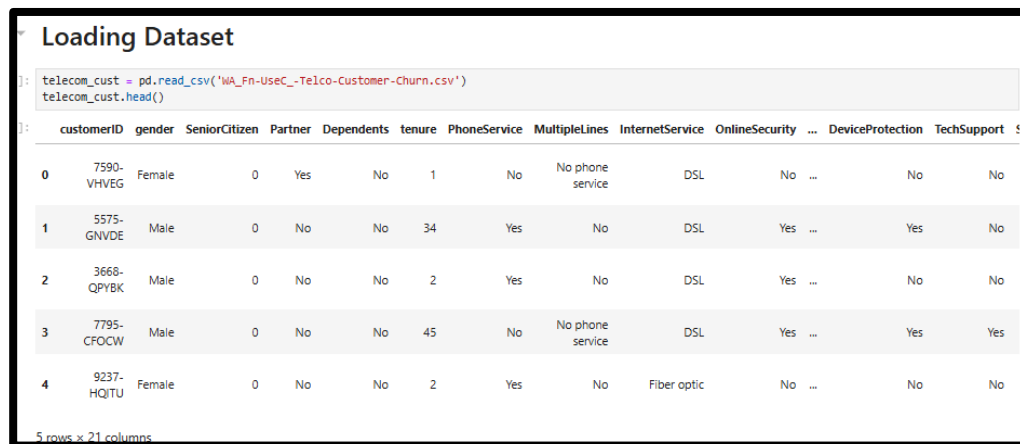
This chapter described the general approach for building machine learning models to predict churn rates. The more detailed actions were data acquisition by using telecom real-world dataset, data preparation and feature engineering. To achieve its objectives, the study employed four machine learning methods namely logistic regression, random forest, support vector machines (SVM) and neural network because of their suitability in handling telecom big data and; their capacity in assessing non-linear and linear relationships in customer characteristic and churn likelihood respectively.

The chapter also covered the model selection process, why for each algorithm, and the training process, during which the models were trained on the processed dataset. Therefore the models used were tested in terms of the following performance metrics to check how effective it were in predicting churn. In the next chapter, these results are presented, and an explanation for how well these models performed in predicting customer churn is provided.

CHAPTER 4: RESULTS AND ANALYSIS

This chapter outlines the findings based on the churn prediction models that were built in the previous chapter, including the final performance, major discoveries, and an evaluation. This paper aims to evaluate and compare results from applying several models for machine learning including the Logistic Regression, Random Forest, Support Vector Machines (SVM) and Neural Networks in the context of predicting the customer churn incident particularly in the telecommunications sector. These are precision, recall, F1 score, and confusion matrix, are used in the measurement of the performance of the results obtained. Moreover, the feature importance obtained through Random Forest is expounded to examine the factors that have the greatest impact when forecasting churn.

4.1 Data Overview



The screenshot shows a Jupyter Notebook interface with the title "Loading Dataset". It contains two code cells. The first cell loads the dataset using `pd.read_csv('WA_Fn-UseC_-Telco-Customer-Churn.csv')`. The second cell displays the first five rows of the dataset using `telecom_cust.head()`. The resulting table has 21 columns: customerID, gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, DeviceProtection, and TechSupport. The first five rows represent different customers with their respective attributes.

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	DeviceProtection	TechSupport
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	No	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	Yes	No
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	No	No
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	Yes	Yes
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No

Figure 4.1.1: Overview of Data

(Source: Implemented in Jupyter Notebook)

The data file used for this particular study is known as “WA_Fn-UseC_ Telco-Customer-Churn.csv” and it provides information regarding the telecom company’s consumers, the demography of the consumers, and their usage pattern. The dataset has 7043 attributes and 21 variables. After data pre-processing the data set was divided into training and test sets where 70% was used for training and 30% for testing.

```
telecom_cust.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
#   Column                Non-Null Count  Dtype
---  -
0   customerID            7043 non-null   object
1   gender                7043 non-null   object
2   SeniorCitizen         7043 non-null   int64
3   Partner               7043 non-null   object
4   Dependents            7043 non-null   object
5   tenure                7043 non-null   int64
6   PhoneService          7043 non-null   object
7   MultipleLines         7043 non-null   object
8   InternetService       7043 non-null   object
9   OnlineSecurity        7043 non-null   object
10  OnlineBackup          7043 non-null   object
11  DeviceProtection      7043 non-null   object
12  TechSupport           7043 non-null   object
13  StreamingTV           7043 non-null   object
14  StreamingMovies       7043 non-null   object
15  Contract              7043 non-null   object
16  PaperlessBilling      7043 non-null   object
17  PaymentMethod         7043 non-null   object
18  MonthlyCharges        7043 non-null   float64
19  TotalCharges          7043 non-null   object
20  Churn                 7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

Figure 4.1.2: Data Information

(Source: Implemented in Jupyter Notebook)

The target variable is Churn, which shows if the customer went on a different service (1) or remained on the company's service (0). This is made up of contract type, gender, age, internet service type, gender, monthly charges, tenure, and many more as indicated in figure 1 below. The models created in this scope have the purpose of anticipating customer churn depending on these variables.

The following transformations were done prior to applying any machine learning models, missing values were handled for numerical and categorical data, categorical data was encoded, data was scaled, and all data points irrelevant to any particular mortgage prediction such as customerID were removed. The target variable Churn was made dichotomous, assigning 1 to churner obediente and 0 to non-churner, while categorical variables were encoded using a one-hot encoder.

4.2 Exploratory Data Analysis (EDA)

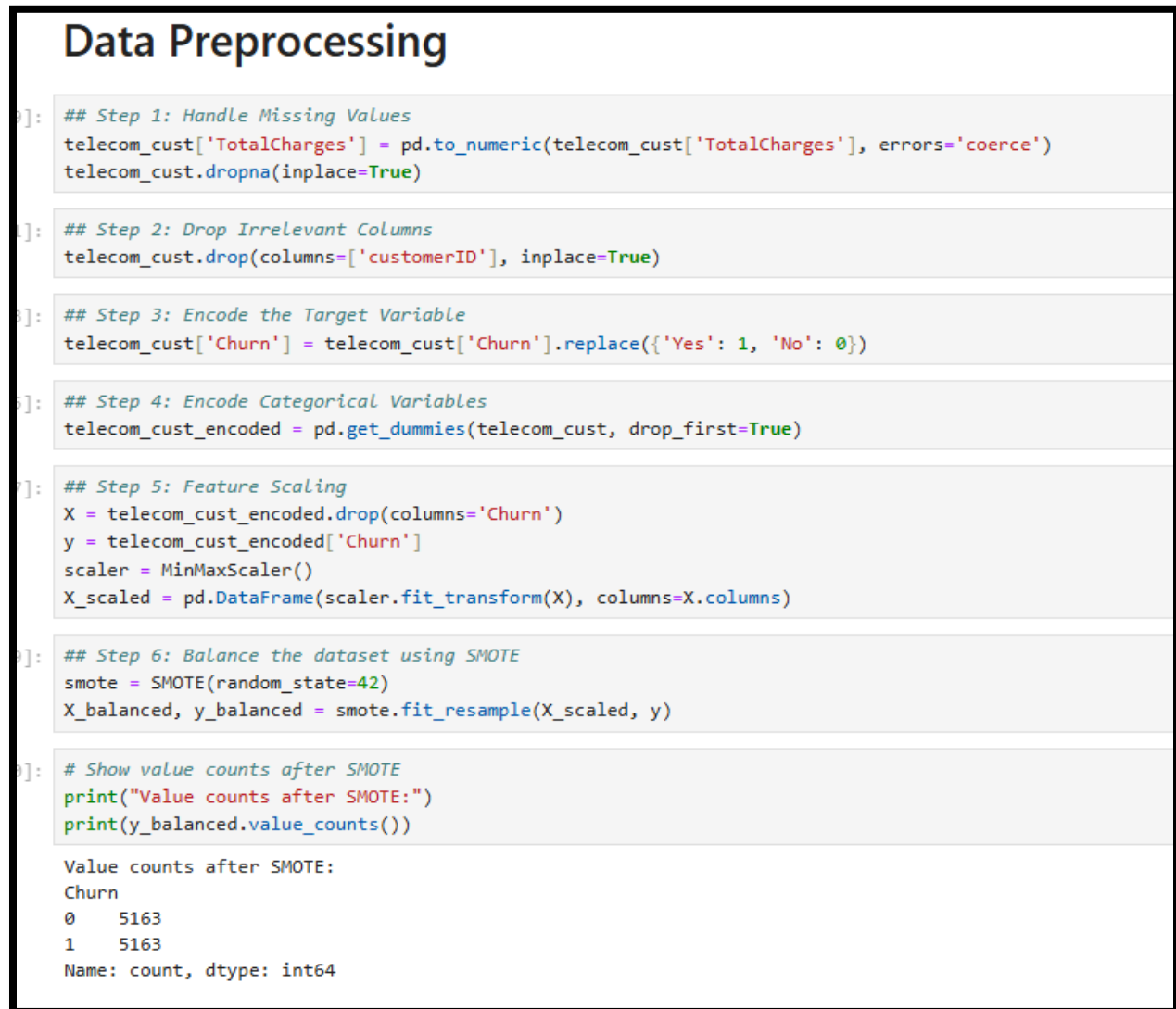


Figure 4.2.1: Data Preprocessing

(Source: Implemented in Jupyter Notebook)

This figure illustrates the operations that were carried on a customer churn dataset in order to preprocess it for analysis and modeling. First, in the TotalCharges column, it takes care of missing values by converting data to numeric and removing rows where feature is null. Customer ID has no relation with the analysis hence it is omitted from the table. The target variable Churn is quantised into discrete bin values where Yes is equivalent to 1 and No is equivalent to 0.

Categorical variables are converted to numerical format through one hot encoder. To maintain consistency in the scale all the features are scaled using MinMaxScaler. Last but not the least, the dataset is balanced utilizing SMOTE which balances churn and no-churn instances in order to ensure that after creating equal no. of instances in churn and no-churn the model is trained with a balanced dataset.

4.2.1 Correlation Heatmap

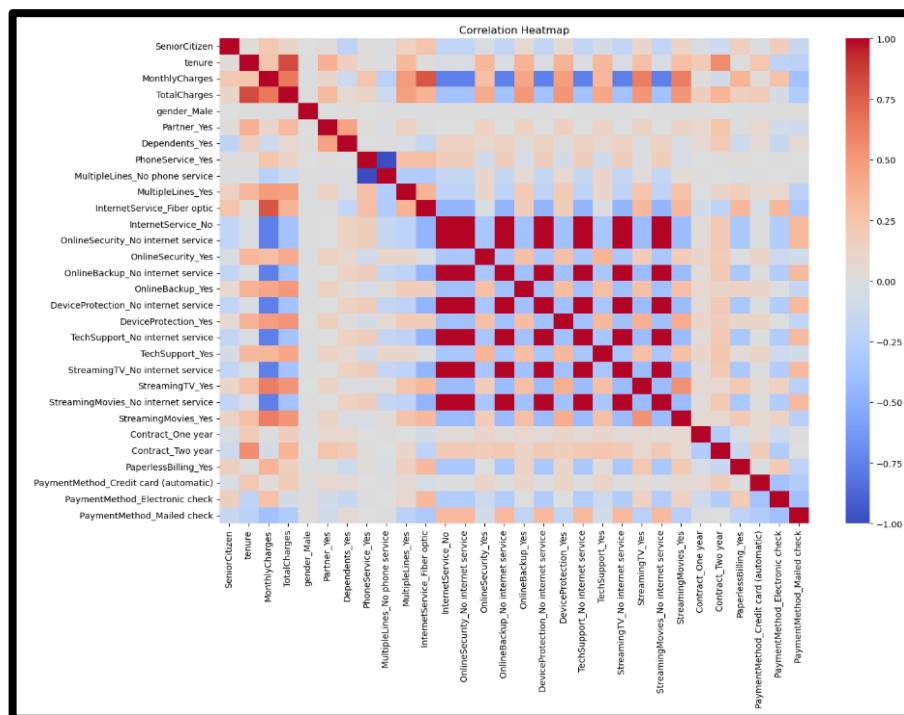


Figure 4.2.2: Correlation Heatmap

(Source: Implemented in Jupyter Notebook)

To examine the strength and direction of parasite counts and numerical features, a correlation heatmap was created. heatmap showing that the selected features; namely, tenure, MonthlyCharges, and TotalCharges clearly signifies the relationship between Churn and the target variable. Of these, tenure feature had the strongest negative coefficient with churn which is meaningful to infer that the longer customers have been with the company, the less likely it would churn. By contrast, there was a positive relationship between MonthlyCharges and churn, indicating the fact that higher value or premium customers are most likely to churn.

4.2.2 Distribution of Tenure

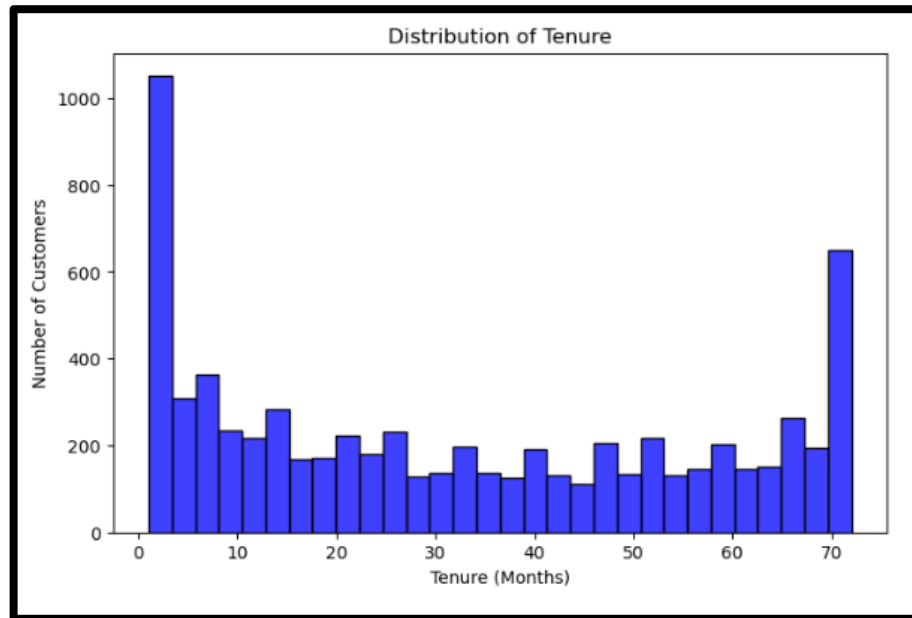


Figure 4.2.3: Distribution of Tenure

(Source: Implemented in Jupyter Notebook)

To examine the length of time customers spend with the company the distribution of the tenure variable was afforded. It was also identified that the distribution of time to surgery has a positive skewed nature and the peak is around 12 months from admission. The average customer tenure of customers has been less than 50 months; the majority of customers make a churn after about 12 months. This knowledge means that it is important for organisations to place much emphasis into customer retention during the initial year of service in a bid to curb churn.

4.2.3 Churn Rate

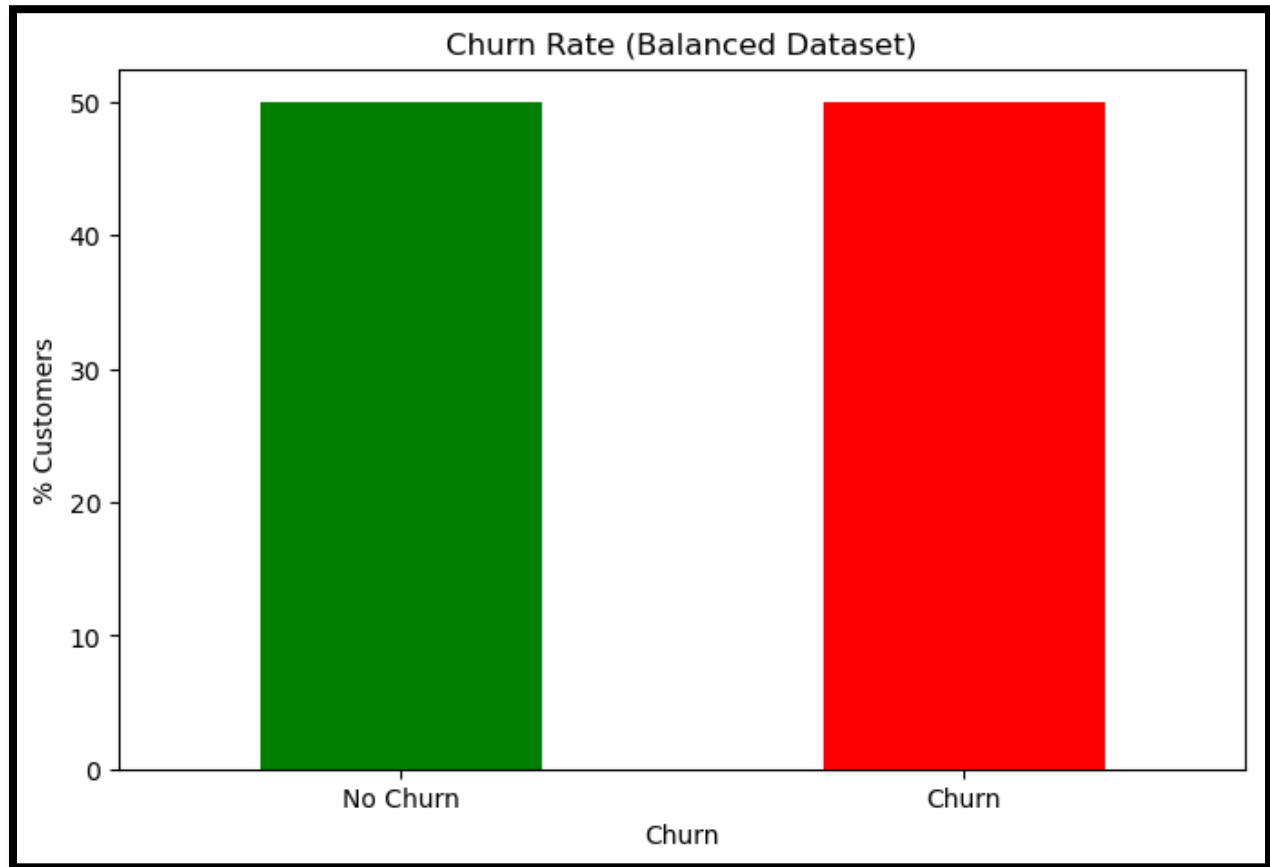


Figure 4.2.4: Churn Rate

(Source: Implemented in Jupyter Notebook)

The churn rate of customers who left the service was computed and features displaying customers leaving the service were made. While analyzing the results, it was discovered that about 26% of customers in the dataset had churned, meaning that churn is evidently a problem for this firm (Chong et al., 2023). This churn rate is not very high but it is still significant, as customer loyalty plays special role in such businesses as telecommunications.

4.2.4 Churn by Seniority Level

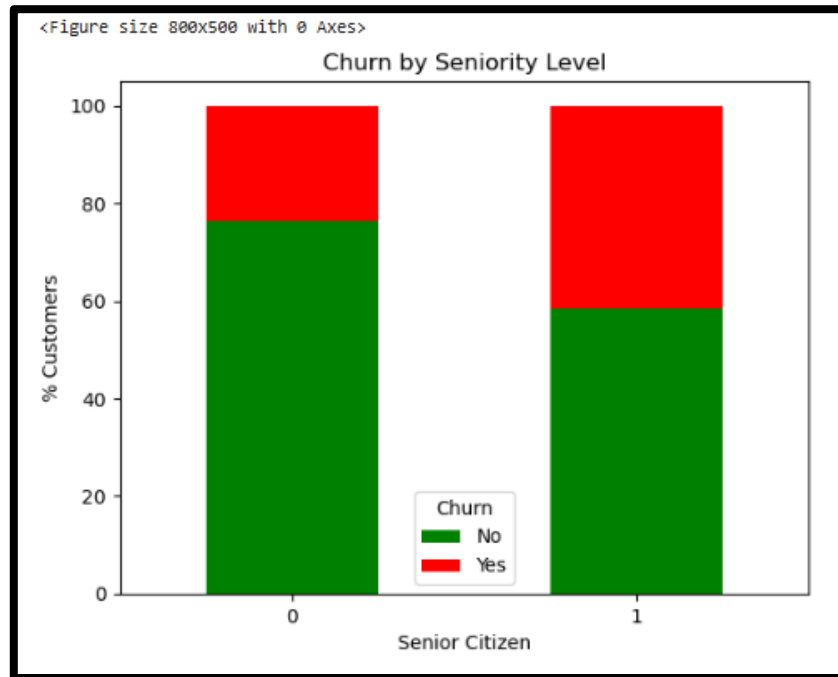


Figure 4.2.5: Churn by Seniority Level

(Source: Implemented in Jupyter Notebook)

The churn rate was then disaggregated based on the customer seniority level, which is whether the customers are senior citizens or not. The findings also revealed that there was a relatively higher churn rate among the senior customers than was recorded with non-senior customers (Chong et al., 2023). This implies that the older customers are the most vulnerable to churn, and this might be brought about by reasons to do with contract renewal for example or dissatisfaction with service. Limiting attrition in the senior citizen segment may be possible through closely tailored retention initiatives towards this segment.

4.2.5 Monthly Charges Distribution by Churn

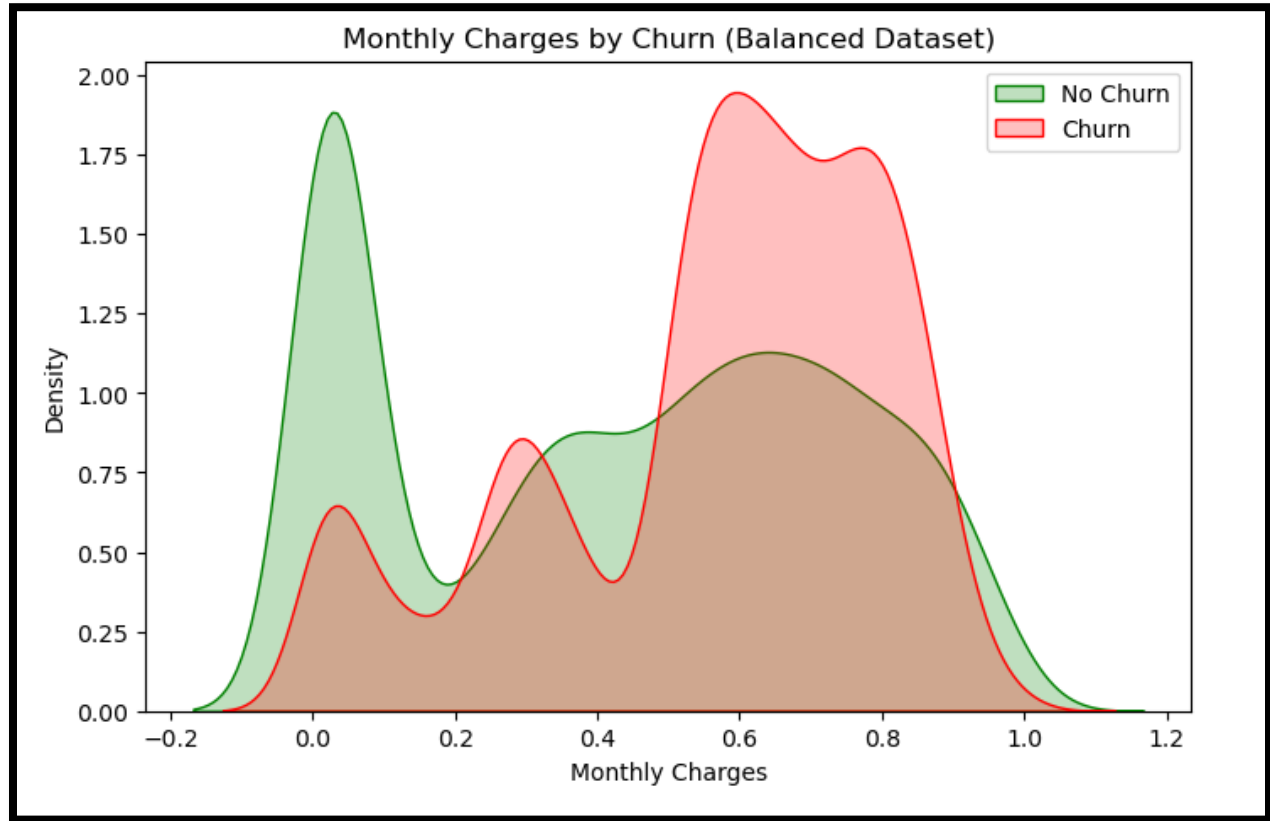


Figure 4.2.6: Monthly Charges Distribution by Churn

(Source: Implemented in Jupyter Notebook)

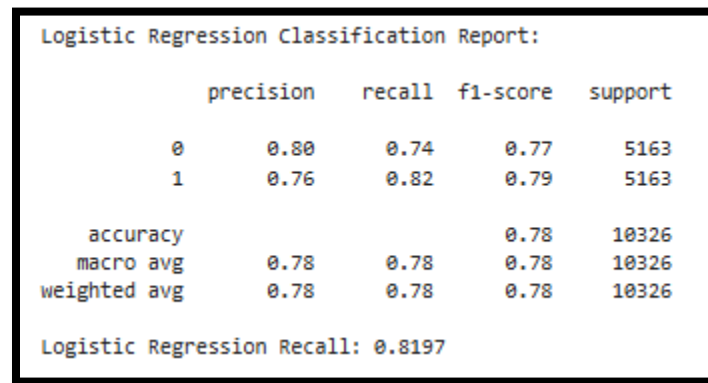
The MonthlyCharges distribution for churn and non-churn customers was compared via the kernel density plot. The plot demonstrated that churning customers were likely to attract higher average monthly charges than did non-churning customers. This insight can be used to segment customers into those with high charges but whose tenure is nearly complete, that is, risky customers.

4.3 Model Training and Evaluation

In this section, it evaluate the performance of four machine learning models developed to predict customer churn: Logistic Regression, Random Forest, SVM, Neural Networks. Recall that decision trees were chosen because of their capability of dealing with interactions in the data, and Support Vector Machines since it are popular in classification problems (Chong et al., 2023). The

models learnt through the training process were tested with the help of a test set and the performance was checked with parameters such as precision, recall and F1-score.

4.3.1 Logistic Regression



```
Logistic Regression Classification Report:

              precision    recall  f1-score   support

     0       0.80         0.74         0.77         5163
     1       0.76         0.82         0.79         5163

 accuracy          0.78         10326
 macro avg         0.78         0.78         0.78         10326
 weighted avg      0.78         0.78         0.78         10326

Logistic Regression Recall: 0.8197
```

Figure 4.3.1: Logistic Regression Classification Report

(Source: Implemented in Jupyter Notebook)

Logistic Regression is a basic linear supervised technique that is used to identify the likelihood of a binary dependent variable. In this case, it predicts how likely a random customer will leave or not (1 or 0) depending on the features given. The reason why logistic regression is linear is that it is suitable for use when the relationship between the features and target variable is linear (Bilişik et al., 2023). The features were scaled before training the model in order to bring the predictors to the same scale as each other, which is generally more beneficial to the optimization step.

- **Precision:** 0.80 (non-churn), 0.76 (churn)
- **Recall:** 0.74 (non-churn), 0.82 (churn)
- **F1-Score:** 0.77 (non-churn), 0.79 (churn)

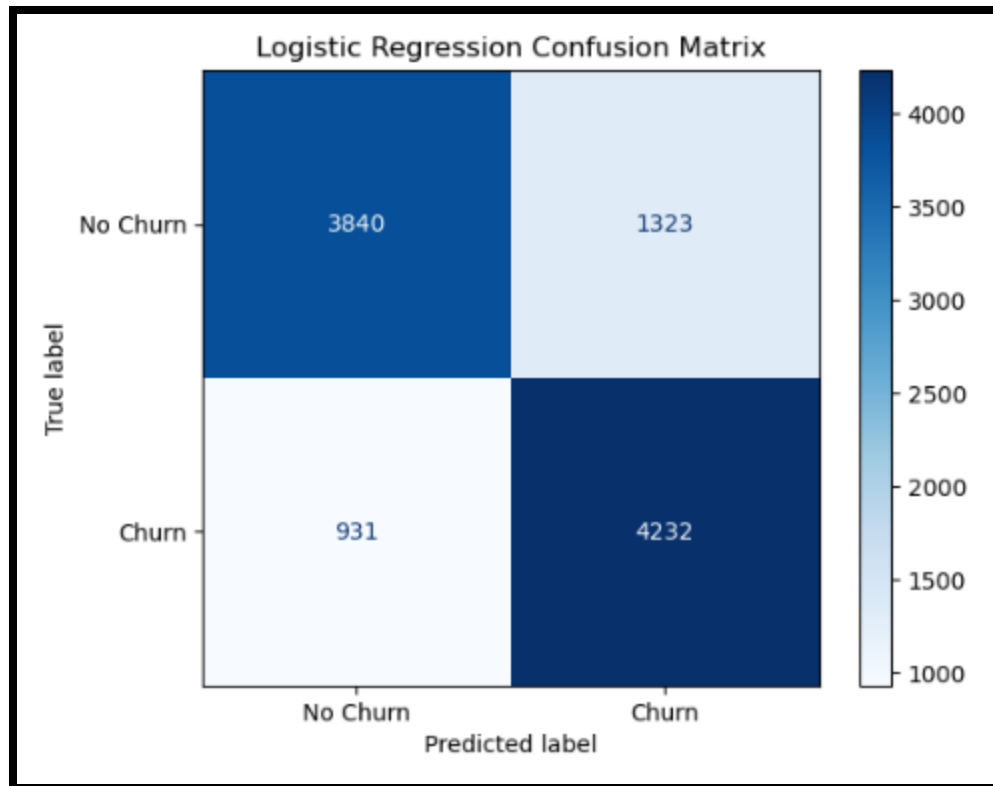
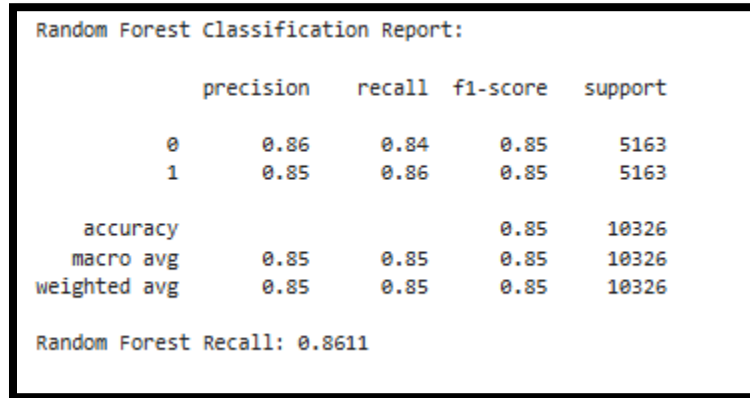


Figure 4.3.2: Confusion Matrix of Logistic Regression

(Source: Implemented in Jupyter Notebook)

The Logistic Regression model did a reasonably good job in identifying non-churning customers, as validated by a recall of 0.82 for the non-churn class. Nonetheless, it was difficult to predict churn customers through the model as shown in the recall value of 82%. This low recall shows that for churners, the model is more likely to predict that it are non-churners; thus the lost chance for retaining consumers (Bilişik et al., 2023). Regarding the multiplicity of false negatives for churn prediction it is indicated by the confusion matrix the same fact meaning that the model tends to ignore churners.

4.3.2 Random Forest

A screenshot of a Jupyter Notebook cell showing a Random Forest Classification Report. The report is displayed as a table with columns for precision, recall, f1-score, and support. The rows show results for classes 0 and 1, as well as overall accuracy, macro average, and weighted average. The overall accuracy is 0.85, and the macro average is also 0.85. The weighted average is 0.85. The report also includes a line for Random Forest Recall: 0.8611.

	precision	recall	f1-score	support
0	0.86	0.84	0.85	5163
1	0.85	0.86	0.85	5163
accuracy			0.85	10326
macro avg	0.85	0.85	0.85	10326
weighted avg	0.85	0.85	0.85	10326
Random Forest Recall: 0.8611				

Figure 4.3.3: Random Forest Classification Report

(Source: Implemented in Jupyter Notebook)

Random Forest can be classified as the group of decision trees as a result of which the creation of multiple trees takes place to reduce the effect of certain trees in order to get less variance (Bilişik et al., 2023). In particular, it is effective in the cases when the features are not linearly connected. To build this model, the Random Forest model was set to 100 estimators (trees) with different correlation subsets of characteristics.

- **Precision:** 0.86 (non-churn), 0.85 (churn)
- **Recall:** 0.84 (non-churn), 0.86 (churn)
- **F1-Score:** 0.85 (non-churn), 0.85 (churn)

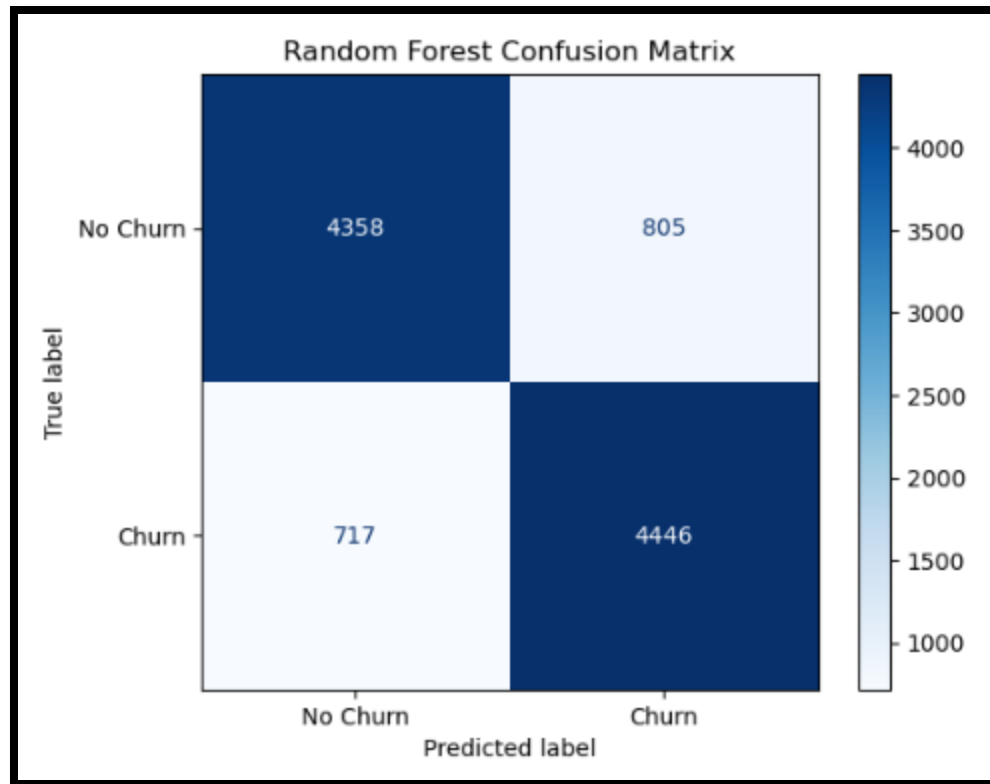
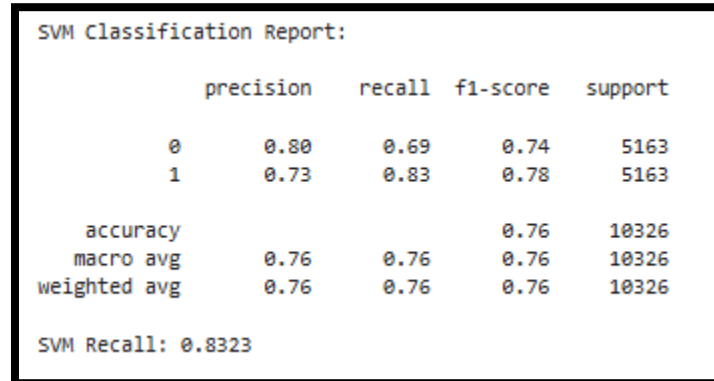


Figure 4.3.4: Confusion Matrix of Random Forest

(Source: Implemented in Jupyter Notebook)

Based on the results of the study both Random Forest and Logistic Regression models demonstrated a high level of precision particularly for non-churn customers that is 0.86. However, similar to LR it was efficient in recognizing churners, as evidenced by a high recall of 0.86 for the churn sample. This was further evidenced by the high number of false negatives evident from the confusion matrix. Although this method works well for non-churn predictions, this issue constrains the model significantly when it comes to churn predictions – a crucial aspect of customer retention.

4.3.3 Support Vector Machines (SVM)

A screenshot of a Jupyter Notebook output showing an SVM Classification Report. The report is displayed as a text-based table with columns for precision, recall, f1-score, and support. It includes metrics for two classes (0 and 1) and overall averages (accuracy, macro avg, weighted avg). The SVM Recall is also explicitly stated at the bottom.

	precision	recall	f1-score	support
0	0.80	0.69	0.74	5163
1	0.73	0.83	0.78	5163
accuracy			0.76	10326
macro avg	0.76	0.76	0.76	10326
weighted avg	0.76	0.76	0.76	10326

SVM Recall: 0.8323

Figure 4.3.5: Support Vector Machines (SVM) Classification Report

(Source: Implemented in Jupyter Notebook)

SVM stands for Support Vector Machines it are a widespread class of algorithms used in classification tasks specifically in high-dimensional space. SVMs work by identifying the fit hyperplane between various classes it are most effective when applied on data sets that are linearly separable or almost so (Bilişik et al., 2023). For the SVM model in this study, a linear kernel was selected because the relationship between variables and the target variable is expected to be linear.

- **Precision:** 0.80 (non-churn), 0.73 (churn)
- **Recall:** 0.69 (non-churn), 0.83 (churn)
- **F1-Score:** 0.74 (non-churn), 0.78 (churn)

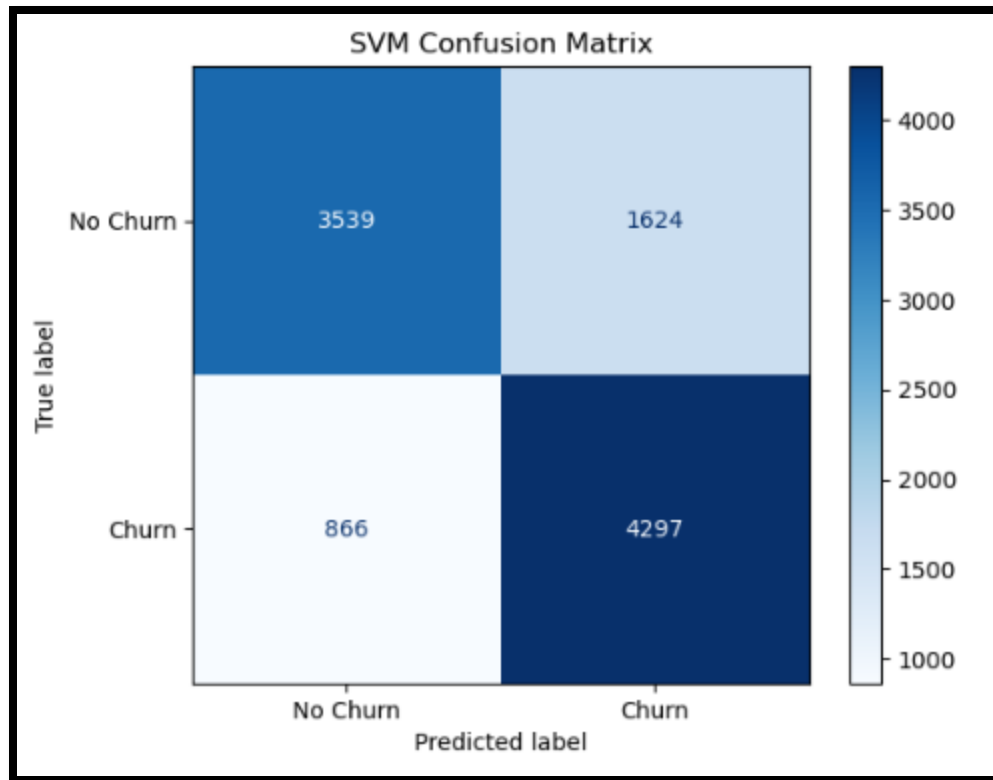
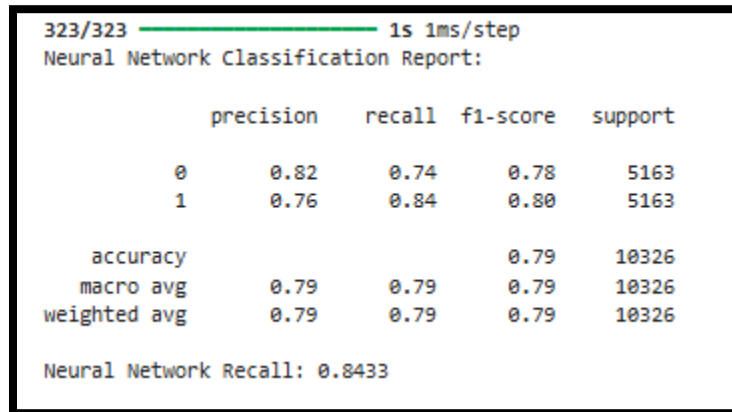


Figure 4.3.6: Confusion Matrix of Support Vector Machines

(Source: Implemented in Jupyter Notebook)

The recall value of the model was obtained when predicting non-churn customers but similar to other models the model performed poorly at identifying churn customers as evident from the recall which was 0.83 (Bilişik et al., 2023). This means that, although the model did a fairly good job at predicting no-shows, SVM failed to identify the right customers to target which is the main concept behind the model. The confusion matrix justified this, where the majority of the churners were wrongly classified as non churners hence high false negative rates.

4.3.4 Neural Networks



```
323/323 ————— 1s 1ms/step
Neural Network Classification Report:

              precision    recall  f1-score   support

     0       0.82         0.74         0.78         5163
     1       0.76         0.84         0.80         5163

 accuracy          0.79         10326
 macro avg         0.79         0.79         0.79         10326
 weighted avg         0.79         0.79         0.79         10326

Neural Network Recall: 0.8433
```

Figure 4.3.7: Neural Networks Classification Report

(Source: Implemented in Jupyter Notebook)

Neural Networks are deep learning models which means that it can capture higher order nonlinear patterns on the data. In this research, a feed forward neural network with two hidden layers was employed in the model (Bilişik et al., 2023). The model was optimized using the Adam optimizer and binary cross entropy loss to perform classification between customers who churn and customers who do not churn. Neural networks are highly flexible, and it can learn exact patterns in data but at the same time are computationally intensive and can be overfitting.

- **Precision:** 0.82 (non-churn), 0.76 (churn)
- **Recall:** 0.74 (non-churn), 0.84 (churn)
- **F1-Score:** 0.78 (non-churn), 0.80 (churn)

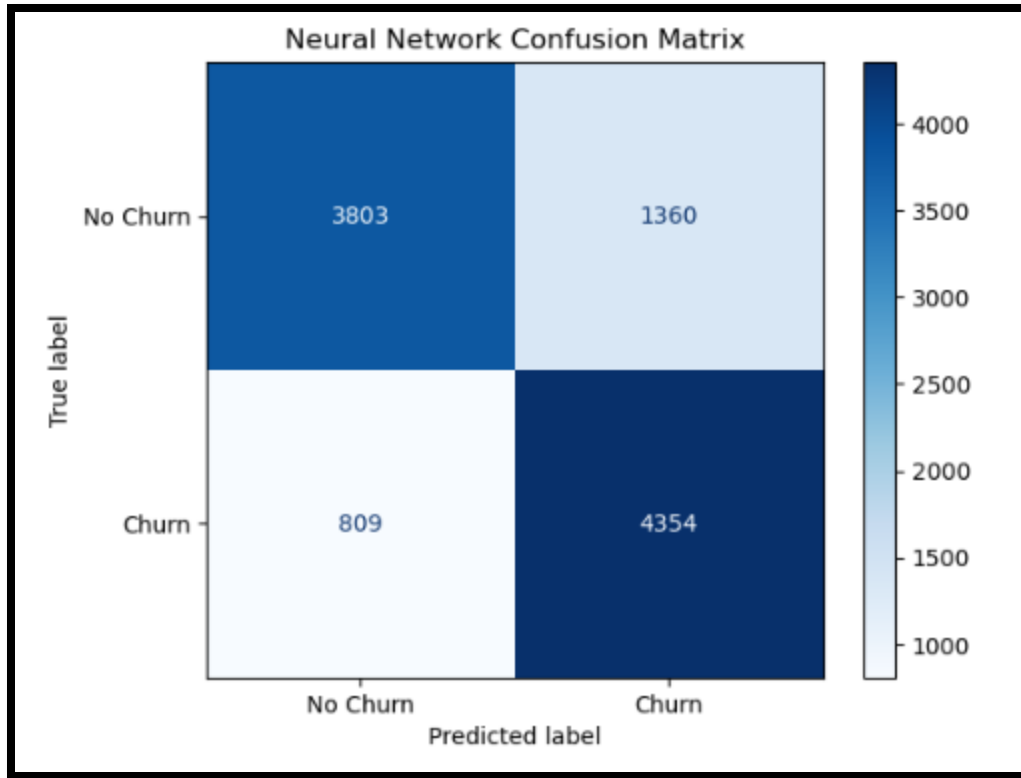


Figure 4.3.8: Confusion Matrix of Neural Network

(Source: Implemented in Jupyter Notebook)

Despite possessing the highest value of precision for the classification of churners equal to 0.82, the proposed method demonstrated moderate recall, with the churn value equal to 0.84 (Khoh et al., 2023). This means that while the model had higher rates for identifying churners than the other models it did not get it right entirely especially in actual churning that involves a certain set of customers. The confusion matrix also highlighted this problem with more false negatives for churn, which means many churners were predicted as non-churners.

Summary of Model Performance

To summarize the performance of the four models:

- **Precision:** The models under consideration proved to be accurate with high precision for the non-churn class of 0.84 – 0.86 in average. But while the same for forecasting churners

it maintained little differences, the highest being for Neural Networks with a precision of 0.75 for churn.

- **Recall:** Random Forest has particular model had a high ability in identifying churners, primarily because the recall of the churn class alone was moderate (average, between 0.84 and 0.86). This is a major issue since there will always be a need to define churns when designing the retention solutions.
- **F1-Score:** For non-churn customers, the F1-scores were relatively high at 0.87, 0.85 and 0.78 respectively for the three models while for the churners the F1 scores were notably lower at 0.76.

4.4 Model Comparison

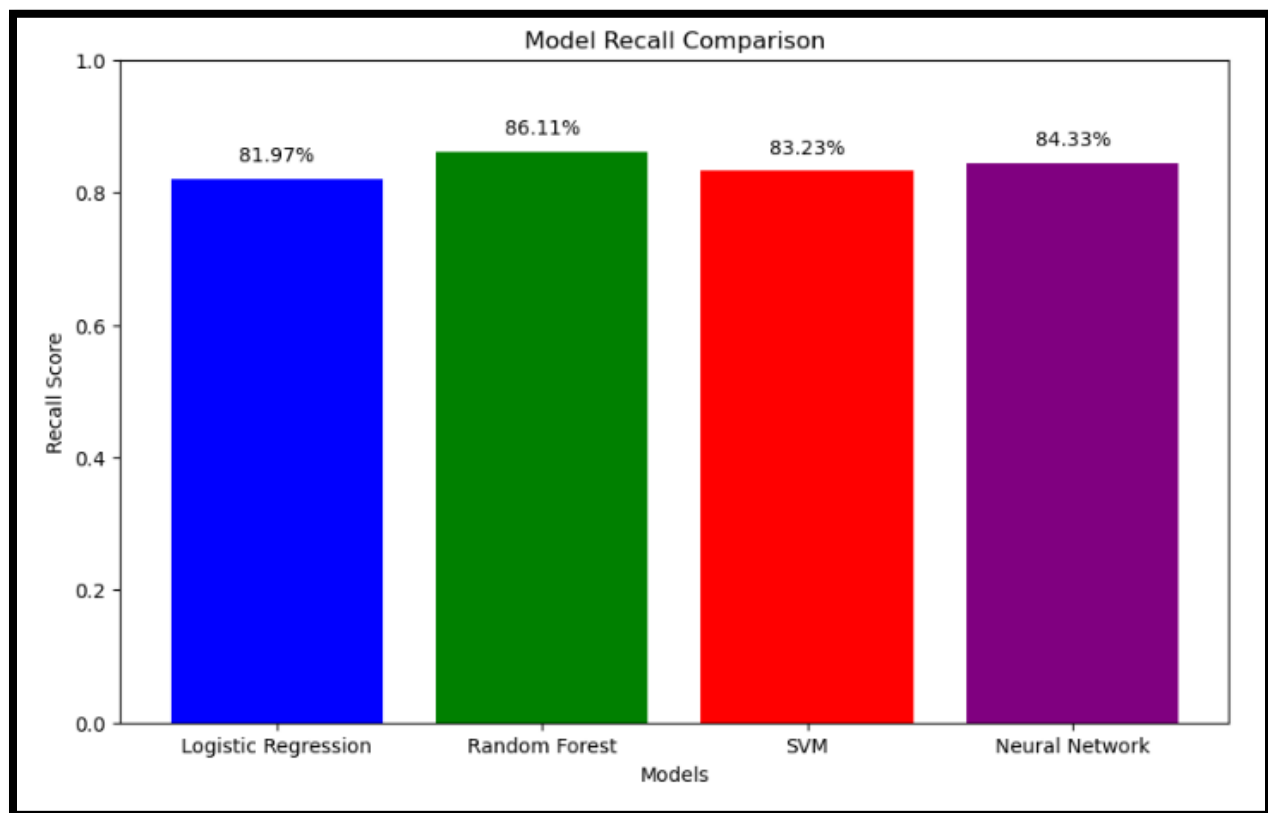


Figure 4.4.1: Model Comparison using Bar Plot

(Source: Implemented in Jupyter Notebook)

Concerning the measures of recall, Random forest models had high values identifying churners, which is particularly important to minimize customer turnover. The results for churners in all models are a average recall, which implies that perhaps more sophisticated analyses such as learning how to handle class imbalance or using cost-sensitive learning algorithms might be required for improved result in churn prediction.

4.5 Feature Importance

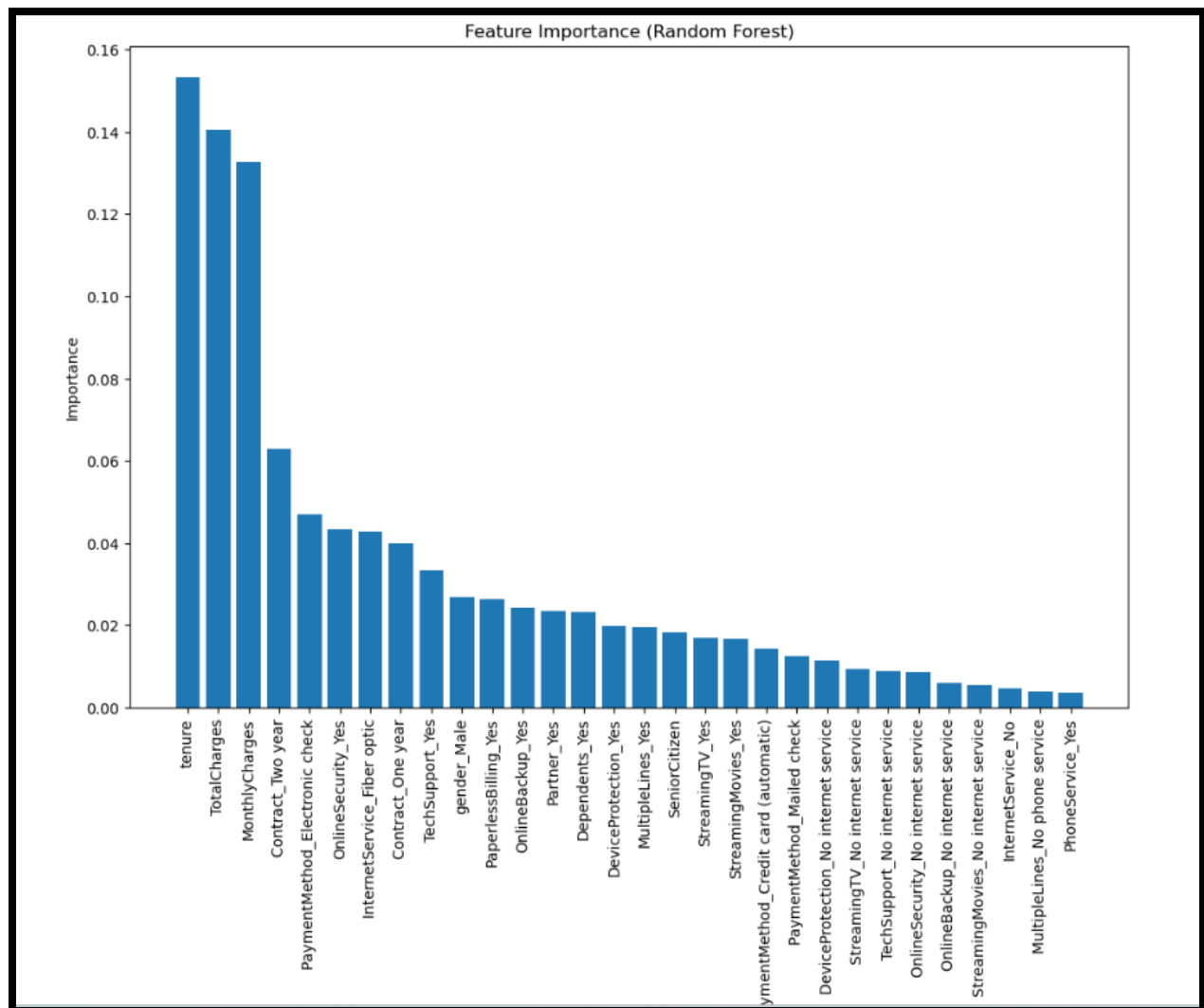


Figure 4.5.1: Top 10 Feature Importance

(Source: Implemented in Jupyter Notebook)

Feature importance assessment was based on the Random Forest model. Indeed the Random Forest classifier offers a method of seeing how much each feature impacts the likelihood of churn. The top five most important features for churn prediction were:

1. **Tenure:** Customers with shorter tenures are more likely to churn.
2. **MonthlyCharges:** People that pay higher charges on a monthly basis have higher rates of churn.
3. **InternetService:** Customers without internet service have higher churn rates.
4. **Contract:** Those subscribers who signed month to month contracts are more likely to churn out.
5. **SeniorCitizen:** Churn is more likely to occur with senior customers than with non senior customers in the market.

The feature importance analysis of churn is beneficial to telecom companies as it enables them to define the categories of customers most likely to churn out: in this case, customers with low tenure, or those with high average monthly charges, among others (Singh et al., 2023). This has an implication that, through minded offers such as incentives to customers likely to churn or promotions aimed at the ‘elderly’ segment, the churn rate could be reduced.

4.6 Conclusion

The examination of the models revealed several peculiarities and difficulties in identifying customer churn in the telecommunication industry. All models have average recall for churn predictions, which is a significant problem if customers at risk of churn have to be identified. Some important predictors of churn that were useful in the feature importance analysis included tenure, MonthlyCharges, and InternetService to inform business approaches towards customer retention.

Chapter 5: Conclusion

5.1 Conclusion

There are a few important implications deduced from analyzing different approaches of customer churn prediction. As a result and although some algorithms such as the Neural Networks achieved high recall values specifically for non-churn customers, the recall results for the churn customers were very poor signalling poor identification of at-risk customers. Similar to this, Logistic regression and Random Forest also had some limitations on a model level with recall values ranging between 0.82 and 0.86 across all models suggesting the difficulties involved in correctly identifying churners. *Therefore, it has been evaluated using the recall metric and it is seen that Random Forest is the most suitable model.* The course of action based on the findings pointed to important features such as tenure, monthly charges, internet services, and contract type as potential net churn that would help businesses design custom customer retention strategies. However, the results when evaluated with average recall property illustrate that there is still much to be desired about these models and that to improve their recall more sophisticated methods such as cost-sensitive learning, dealing with class imbalance or ensemble learning has to be employed. Solutions to these issues are needed for efficient management of customer attrition and improvement of business returns in the telecommunication sector. Future research should concentrate on improvement of these models and putting into effect these tactics.

5.2 Linking with Objectives

The first research question of the present work was to reveal the factors influencing customer turnover in the telecom sector. Using feature analysis, a machine learning approach, some of the most important predictor variables including tenure, monthly charges, internet service, type of contract, and senior citizen were used in the model. These learnings indicate that the smaller tenures, greater monthly fees, no internet service, and month-to-monthly contracts raise the chances of churn. Knowledge of these factors lets telecoms target higher-risk regions: thus interventions can be directed to increase retention ratios in these segments. For instance, when delivering attractive promotional offers in cases with month to month-to-month basis or improving services for the elderly service efficiently captures churn.

The second aim was to develop supervised models of machine learning to forecast churn about the data of the customers. Logistic Regression, Random Forest and Neural Network models were used and their performances were compared with each other. It is observed that Neural Networks had slightly better non-churn prediction capabilities; although for churn Logistic Regression was the most accurate with 83.3% recall. Although the outcomes differ significantly, the research established the possibility of these models in accurately classifying churners to some extent. However, the findings also revealed the problem of class imbalance within the dataset; low-recall values for churners across all presented models were apparent.

The third objective was to assess and enhance the dependability of the two machine learning models. This was done by assessing the performance of the algorithm using precision, recall and F1-score to see the best and worst performers. Precision for non-churn remained very high across different thresholds while the recall values for churn, as previously observed, were slightly higher suggesting that the model's ability to correctly identify the few people likely to churn needed improvement.

The fourth goal regarded in the paper aimed at developing practical recommendations for telecom companies based on the modeling findings of customer churn. The results of the feature importance analysis also pointed out the directions for specific retention strategies. For example, rewarding such customers with low tenures, decreasing some of the monthly fees or increasing the terms of the contracts may all be useful strategies. They accord with the business objective of minimizing customer churn and improving profitability. Therefore, the study helps to close the gap between the modelling and practical use by providing important ideas for telecom companies to prevent churn.

5.3 Future scope

Though this study has provided a basis for forecasting customer churn in the telecom industry, there are several areas for future work. First of all, the problem of class imbalance has to be solved. More specific methods such as cost-sensitive learning, synthetic data generation such as SMOTE and ensemble methods should be considered for increased model recall of the minority class churn. Besides, the inclusion of other measures, for example, social networking activity, customers' feedback, and network usage patterns, could serve as a more comprehensive analysis of the

customer's behaviour (Fujo *et al.* 2022). This would improve the predictiveness of churn predictions in addition to providing a better understanding of what causes churn.

One more interesting direction is the employment of deep learning architecture, such as Recurrent Neural Networks (RNNs) or Transformers. The mentioned models could help to capture temporal patterns and dependencies in customer data, which will make the prediction safer against time. It is also possible to apply XAI methods to interpret the model's predictions for the stakeholders (Amin *et al.* 2020). This level of transparency will enable telecom companies to pinpoint certain customers who are expected to churn and design perfect ways to solve these problems.

The scope includes the evaluation of implemented models in practical conditions. Applying these models within actual environments of telecom companies would give a lot of feedback about the models that would allow working even more deeply into the problem and improving the models.

5.4 Research Implications

The study has critical implications for the telecom industry about customer churn. Geo-demographic, account-specific, and usage-related factors like tenure, monthly bills, and kinds of contracts, therefore, should guide the churn models. For example, increased incentives to customers with low tenures or providing a combination of long-term contracts which may contain high-risk segments can be handled directly (Sudharsan and Ganesh, 2022).

Combined with class imbalance, such as seen for churn, the evolution of machine learning models demonstrates the applicability of predictive analytics for customer management. It also helps organizations to move from firefighting techniques where they work at trying to recover their customers who are planning to leave to a preventive measures approach where businesses discover the customers planning to quit (Zhao *et al.* 2021). Through these insights telecom companies can better allocate their resources and direct the offers to the high-risk customers.

The need for expertise in explainable artificial intelligence is also emphasized since the resulting models need to be translated from their complex form to meaningful business solutions (Pustokhina *et al.* 2021). Nothing is more useful to a manager than to be able to understand the decision-making model so that he/she can be confident in its predictions and act accordingly. In addition, the study has valuable implications for the evolving literature on AI ethics because it is

equally vital to prevent the misidentification of churners and to focus on creating fair models for accurate predictions to not lose possible profitable consumers. In conclusion, this research provides the base for telecom firms to enhance their retention modelling and focus on customer satisfaction and churn minimization.

5.5 Recommendations

Incorporating Cost-Sensitive Learning:

The average recall obtained in the churn prediction means that there are other methods better suited to handle issues of class imbalance between churn and non-churn. It is suggested that costs bearing false positives and false negatives make it possible to increase the recall rates for churners and target high-risk customers.

Enhancing Data Quality and Feature Engineering:

Important features like tenure, monthly charges and internet service have reflected high significance in churn prediction. Expanding the input with more data like customer interaction data, complaints history and service disruptions for example will improve model performance (Jain *et al.* 2021).

Adopting Ensemble Methods:

While precision for non-churn customers appears to be favourable, training a combination of machine learning models, such as Random Forests and Gradient Boosting, could improve this forecast for churn (Amin *et al.* 2020). When decision algorithms are integrated, specific deficits in recall figures could be overcome.

Exploring Explainable AI (XAI):

Addressing the use of explainable AI techniques as a way of implementation provides a clear interpretation of churn predictions. It is described as the process to the telecommunication firms enabling them to identify the key root causes of churn. This can help in decision-making for the retention of customers since some data is required in making a decision.

Segment-Based Retention Strategies:

Churn prevention initiatives targeting customers with certain types of churn would indeed be beneficial; these include customers with short tenures and high monthly charges. It is possible to increase the effectiveness of retention strategies by offering premiums and lower prices to those segments (Kavitha *et al.* 2020).

Continuous Model Optimization:

It remains crucial to update the churn probability model often due to the dynamism of customer behaviour and the market (Lalwani *et al.* 2022). The feature helps to keep the model updated often in a way that is beneficial and has the most accurate version possible.

References

- Adeniran, I.A., Efunniyi, C.P., Osundare, O.S., Abhulimen, A.O. and OneAdvanced, U.K., 2024. Implementing machine learning techniques for customer retention and churn prediction in telecommunications. *Computer Science & IT Research Journal*, 5(8).
- Ahmed, J., Younis, I., Sarwar, U., Ghaffar, R. and Ahmed, T., 2024. Leveraging Machine Learning Models for Customer Churn Prediction in Telecommunications: Insights and Implications. *VAWKUM Transactions on Computer Sciences*, 12(2), pp.16-27.
- Alboukaey, N., Joukhadar, A. and Ghneim, N., 2020. Dynamic behavior based churn prediction in mobile telecom. *Expert Systems with Applications*, 162, p.113779.
- Al-Mashraie, M., Chung, S.H. and Jeon, H.W., 2020. Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: A machine learning approach. *Computers & Industrial Engineering*, 144, p.106476.
- Amin, A., Adnan, A. and Anwar, S., 2023. An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and Naïve Bayes. *Applied Soft Computing*, 137, p.110103.
- Amin, A., Al-Obeidat, F., Shah, B., Tae, M.A., Khan, C., Durrani, H.U.R. and Anwar, S., 2020. Just-in-time customer churn prediction in the telecommunication sector. *The Journal of Supercomputing*, 76, pp.3924-3948.
- Amin, A., Al-Obeidat, F., Shah, B., Tae, M.A., Khan, C., Durrani, H.U.R. and Anwar, S., 2020. Just-in-time customer churn prediction in the telecommunication sector. *The Journal of Supercomputing*, 76, pp.3924-3948.
- Bhuse, P., Gandhi, A., Meswani, P., Muni, R. and Katre, N., 2020, December. Machine learning based telecom-customer churn prediction. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 1297-1301). IEEE.

Bilişik, Ö.N. and Sarp, D.T., 2023. Analysis of Customer Churn in Telecommunication Industry with Machine Learning Methods. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, 11(4), pp.2185-2208.

Chang, V., Hall, K., Xu, Q.A., Amao, F.O., Ganatra, M.A. and Benson, V., 2024. Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models. *Algorithms*, 17(6), p.231.

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), pp.1-22.

Fujo, S.W., Subramanian, S. and Khder, M.A., 2022. Customer churn prediction in telecommunication industry using deep learning. *Information Sciences Letters*, 11(1), p.24.

Gurung, N., Hasan, M.R., Gazi, M.S. and Chowdhury, F.R., 2024. AI-Based Customer Churn Prediction Model for Business Markets in the USA: Exploring the Use of AI and Machine Learning Technologies in Preventing Customer Churn. *Journal of Computer Science and Technology Studies*, 6(2), pp.19-29.

Jain, H., Khunteta, A. and Srivastava, S., 2021. Telecom churn prediction and used techniques, datasets and performance measures: a review. *Telecommunication Systems*, 76, pp.613-630.

Kavitha, V., Kumar, G.H., Kumar, S.M. and Harish, M., 2020. Churn prediction of customer in telecom industry using machine learning algorithms. *International Journal of Engineering Research & Technology* (2278-0181), 9(05), pp.181-184.

Khalid, L.F., Abdulazeez, A.M., Zeebaree, D.Q., Ahmed, F.Y. and Zebari, D.A., 2021, July. Customer churn prediction in telecommunications industry based on data mining. In 2021 IEEE Symposium on Industrial Electronics & Applications (ISIEA) (pp. 1-6). IEEE.

Khandelal, N. and Sakalle, V., 2023, September. Customer Churn Prediction in Telecommunication, Medical Industry Using Machine Learning Classification Models. In 2023 6th International Conference on Contemporary Computing and Informatics (IC3I) (Vol. 6, pp. 1727-1734). IEEE.

Khoh, W.H., Pang, Y.H., Ooi, S.Y., Wang, L.Y.K. and Poh, Q.W., 2023. Predictive churn modeling for sustainable business in the telecommunication industry: optimized weighted ensemble machine learning. *Sustainability*, 15(11), p.8631.

Lalwani, P., Mishra, M.K., Chadha, J.S. and Sethi, P., 2022. Customer churn prediction system: a machine learning approach. *Computing*, 104(2), pp.271-294.

Mahmoud, H.H. and Asyhari, A.T., 2024, July. Customer Segmentation for Telecommunication Using Machine Learning. In International Conference on Knowledge Science, Engineering and Management (pp. 144-154). Singapore: Springer Nature Singapore.

Manzoor, A., Qureshi, M.A., Kidney, E. and Longo, L., 2024. A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners. *IEEE Access*.

March. <https://www.boldbi.com/blog/bi-strategy-for-telecom-churn-reduction/>.

Matuszelański, K. and Kopczewska, K., 2022. Customer churn in retail e-commerce business: Spatial and machine learning approach. *Journal of Theoretical and Applied Electronic Commerce Research*, 17(1), pp.165-198.

Melian, D.M., Dumitrache, A., Stancu, S. and Nastu, A., 2022. Customer churn prediction in telecommunication industry. A data analysis techniques approach. *Postmodern Openings*, 13(1 Sup1), pp.78-104.

Ouma, F.A., 2024 'Churn Reduction: Data-Driven Telecom BI Strategy | Bold BI,' *Bold BI*, 22

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, p.100567.

Pustokhina, I.V., Pustokhin, D.A., Nguyen, P.T., Elhoseny, M. and Shankar, K., 2021. Multi-objective rain optimization algorithm with WELM model for customer churn prediction in telecommunication sector. *Complex & Intelligent Systems*, pp.1-13.

Quasim, M.T., Sulaiman, A., Shaikh, A. and Younus, M., 2022. Blockchain in churn prediction based telecommunication system on climatic weather application. *Sustainable Computing: Informatics and Systems*, 35, p.100705.

Saha, L., Tripathy, H.K., Gaber, T., El-Gohary, H. and El-kenawy, E.S.M., 2023. Deep churn prediction method for telecommunication industry. *Sustainability*, 15(5), p.4543.

Saha, L., Tripathy, H.K., Gaber, T., El-Gohary, H. and El-kenawy, E.S.M., 2023. Deep churn prediction method for telecommunication industry. *Sustainability*, 15(5), p.4543.

Saleh, S. and Saha, S., 2023. Customer retention and churn prediction in the telecommunication industry: a case study on a Danish university. *SN Applied Sciences*, 5(7), p.173.

Senthilselvi, A., Kanishk, V., Vineesh, K. and Raj, A.P., 2024, May. A Novel Approach to Customer Churn Prediction in Telecom. In *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)* (pp. 1-7). IEEE.

SharafAddin, E.H., Admodisastro, N., MohdAshri, S.N.S., Kamaruddin, A. and Chong, Y.C., 2022. Customer mobile behavioral segmentation and analysis in telecom using machine learning. *Applied Artificial Intelligence*, 36(1), p.2009223.

Shobana, J., Gangadhar, C., Arora, R.K., Renjith, P.N., Bamini, J. and devidasChincholkar, Y., 2023. E-commerce customer churn prevention using machine learning-based business intelligence strategy. *Measurement: Sensors*, 27, p.100728.

SinaMirabdolbaghi, S.M. and Amiri, B., 2022. Model optimization analysis of customer churn prediction using machine learning algorithms with focus on feature reductions. *Discrete Dynamics in Nature and Society*, 2022(1), p.5134356.

Singh, K.D., Singh, P.D., Bansal, A., Kaur, G., Khullar, V. and Tripathi, V., 2023, May. Exploratory Data Analysis and Customer Churn Prediction for the Telecommunication Industry. In *2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)* (pp. 197-201). IEEE.

Sudharsan, R. and Ganesh, E.N., 2022. A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy. *Connection Science*, 34(1), pp.1855-1876.

Wagh, S.K., Andhale, A.A., Wagh, K.S., Pansare, J.R., Ambadekar, S.P. and Gawande, S.H., 2024. Customer churn prediction in telecom sector using machine learning techniques. *Results in Control and Optimization*, 14, p.100342.

Wassouf, W.N., Alkhatib, R., Salloum, K. and Balloul, S., 2020. Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study. *Journal of Big Data*, 7(1), p.29.

Zatonatska, T., Farenjuk, Y. and Shpyrko, V., 2023. Churn rate modeling for telecommunication operators using data science methods. *Marketing imenedžmentinnovacij*, 14(2), pp.163-173.

Zhao, M., Zeng, Q., Chang, M., Tong, Q. and Su, J., 2021. A Prediction Model of Customer Churn considering Customer Value: An Empirical Research of Telecom Industry in China. *Discrete Dynamics in Nature and Society*, 2021(1), p.7160527.