İSTİNYE ÜNİVERSİTESİ

FACULTY: ECONOMICS AND ADMINISTRATIVE SOCIAL SCIENCES

DEPARTMENT: MANAGEMENT INFORMATION SYSTEMS

MIS407 - BA Degree Project 2

How can telecommunications companies leverage machine learning algorithms to accurately predict high-value customer churn and implement proactive retention strategies to minimize customer loss and optimize revenue?

PREPARED BY

NURETTİN KARTAL

200527122

ADVISOR: Head of Department Prof. Şebnem Özdemir.

ISTANBUL, JUNE 2024

Chapter 1.

Introduction 1.

1.1. Background

1.1.1 The Telecommunications Industry

The telecommunications industry is one of the most important sectors among all the industries in this age. It progressed very fast recently because of the fast technological world with it's constant inventions. But first what is telecommunications? The telecommunications industry contains different types of companies that specialize in the field of communication. This communication is just to make people to link up from different areas around the world, the field of telecommunications include technologies such as telephone networks, mobile phones, data service providers, and others. The telecom sector is known for companies that provide the following services: such as internet service providers, mobile manufacturers, mobile operators, network operators, and more.

Telecom companies in the industry are now focused solely on providing valuable services to mobile device or smart phone users, and that is due to increased connectivity, and the abundance of mobile devices, (Beers Brian, Dr. Brown, JeFreda R. Brown., 2023). There are a lot of financial difficulties that companies in the industry face, and one of them that stand out is customer churn. Customer churn is a crucial problem that needs fast solution, because of its effect on the companies' financial standing in the telecom sector.

1.1.2 What is Customer Churn?

"Customer churn" is an important issue in the telecommunications sector, Customer churn means when a customer terminates his relationship with a provider provide for a period of time. Churn has negative consequences on the reputation and the financial status of the. company, churn affects the service provider's status in the industry and reduces its income streams. The impact is of customer churn can also be seen in the real world as bad reputation between customers.

First, customer churn causes a loss in revenue which is an issue to be worried about, because it is one of the great problems that have an immediate and direct impact on the financial stability of the company. Customer churn leads to an obvious decline in the number of customers and leads to the shrinking of the revenue stream, therefore it leads to low competitiveness and financial problems. When a customer leaves a company, he or she originally may have contributed on a monthly or weekly basis to the company's revenue, so when the customer leaves the company, it significantly shrinks the company's revenue.

Second, the consequence of this issue is not only related to the financial aspect, but churn also has an impact on the company's status and reputation in the industry. Customer churn can create a negative impression of the company to potential customers. Reichheld et al.(1990) had stated in his well cited paper that attaining new customers is usually more expensive than keeping existing ones, to apply this concept in the telecommunications industry, companies need

to handle the existing customer at risk of churn early, instead of trying to obtain new customers. And in modern days, for companies in the industry to properly handle this difficulty, they must use machine learning algorithms, artificial intelligence, prediction models, and more, to accurately predict customers that might end their relationship with the company.

Customer churn in a telecom company typically means that there is low loyalty among customers, companies in the sector desperately need to counteract disloyalty, as it is one of the main drivers of churn, so there must be an understanding of what the term “Patronage” means, Patronage is defined as a gift, or a constant stream of payments of purchases made by customers to their service provider, it is a unique way for businesses to have beneficial spurts of revenue, customers establish patronage when they feel loyal to their service provider and trust in them. Companies need to implement patronage in their relationship with customers, and this will happen when they must first understand the importance of customer retention, and the strategies that need to be implemented to prevent churn and promote loyalty.

Telecom companies nowadays are forced to change their products according to the always changing nature of customers, Companies find themselves constantly having to change their way of conducting business to deal with the high rates of customer churn, because it acts as a threat to their revenue and profit. But unfortunately, this continuous change slows them down and limits the company’s ability to invest in the improvement of its products and services.

That’s why Customer lifetime value (CLV) needs to be discussed, CLV is a method that is always implemented in companies battling with churn. It is known the telecom companies receive their profit from subscription fees paid by clients. Thus, it is very significant to understand the concept of Customer Lifetime Value (CLV), CLV is the most commonly used method for calculating the company's long-term customer value (Kumar et al.2008), the reason that companies use CLV is to measure their customer’s value to deal with each segment accordingly, so when companies have a calculated value of the customer, they then process it by using critical strategic decision-making processes regarding customer acquisition, retention and marketing campaigns; then they are able to allocate their resources efficiently, and give attention to the high-value customers that contribute the most, then finally they implement the retention strategies needed to focus on keeping the customers that contribute the most.

It is already well established that high-value customers are identified by their ability to bring significant revenue to the company, normally telecom companies find themselves in a place where they have to focus more on the high value segment and offer specialized services and more individual attention to keep their relationship with valuable clients stable. Take this example for a more practical application, A provider applied the CLV method on their customer data, the provider ended up finding a customer who is subscribing to the international calling and data roaming package which is the most expensive one, normally this client is classified as a “high value customer” with a value incrementing over time, which is more than the sum of a couple of customers who subscribe to the basic plans, so it is a must that the provider ends up worrying about keeping the high value customer to maximize profit, by implementing retention strategies.

The awareness of CLV among telecom companies will definitely increase the companies' chance of launching retention campaigns. Retention strategies are tailored to attract high-value customers by personalized discounts, loyalty programs, or customer service improvements (Banasiewicz, 2004), these strategies essentially are based on the application of machine learning algorithms that detect the high value customers to therefore target in personalized retention techniques.

1.2 Problem Statement

The main focus in this research is to reduce high value customers at risk of churn in telecom companies using prediction models. The segment of customer at risk of churn will be predicted by applying multiple machine learning algorithms such as logistic regression, random forest, kernel support vector machine, and gradient boosting machines (XGboost). The Telco customer churn dataset will be analyzed by these algorithms to extract patterns and indicators of churn, and point out accurately the segment churning, after predicting the churn segment, the implementation of retention strategies will be recommended for telecom companies.

This main task to be executed in this research will be a comparative analysis of the algorithms, by applying different models to the data, there will be a comparison in accuracy of each model, this will help telecom companies gain an understanding of these models to compete against their competitors.

1.3 Research Question

How can telecommunications companies leverage machine learning algorithms to accurately predict high-value customer churn and implement proactive retention strategies to minimize customer loss and optimize revenue?

1.4. Purpose and Scope

1.4.1. Purpose

The purpose of this thesis is to analyze the churn in customers, particularly in the telecom sector. There will be a brief descriptive analysis that will help in understanding the causes of customer loss and predicting churn. The research is to be based on the analysis of the Telco Customer Churn dataset, which consists of information about customers in a fictional telecom company.

The main goal to be achieved by techniques like logistic regression and random forest is to predict the segment of high value customers who have a high probability to churn. These

models are essentially implemented to be utilized in identify churn customers, this will help lay the foundation of personalized retention plans for high-value customers.

1.4.2. Scope

This project's essence is in the comparison of the four machine learning models mentioned. The first step of conducting the project will involve the preprocessing of the dataset to standardize the dataset, remove all the errors, remove missing values, and deal with outliers. Next, the following prediction models will be applied to the Telco Customer Churn dataset: logistic regression, kernel support vector machine, random forest, gradient boosting machines (xgboost in specific. The main objective is to determine which of these models is the most reliable one in terms prediction accuracy.

By that, not only the predictive modeling will be covered but also the retention strategies will be recommended in order to counteract churn among high-value consumers. Telecom company will be able to retain its high-value customers through personalized approaches. One of the core components of this project will be to use the findings from the analysis to create a customized retention strategy that will help telecom companies to achieve the aims of keeping the customers loyal.

The dataset used contains data of 7043 rows and 21 tabular columns. Because this is a Graduation Project thesis in the area of Management Information Systems, which will be an all-encompassing thesis aimed at predicting the customer segment churning within the telecommunication sector.

1.5 Objectives of the Study

In this research, the objectives are multi-dimensional. Firstly, this research is going to assess and compare the four predictive modeling methods/techniques, the logistic regression, kernel support vectore machine, random forest, and gradient boosting machine, in basically predicting the customer churn precisely.

The objective of this research will be to determine the most efficient model for detecting customer churn behavior among the telecommunications customers. The research also aims to identify high-value consumers in the telecom dataset in order to preserve their significant financial contribution, when this segment of clients if identified, it is then very important to keep them because they contribute so much to the business's profitability, which shows the value of implementing efficient retention plans to keep them loyal.

1.6 Significance of the Study

The significance of this paper is in the following notion, this research essentially targets these customers unlike the common churn prediction research, by putting the high-value customers in a segment isolated from all types of customers, this is to show that valuable customers shouldn't be treated like all customers in the business field, because they are valuable and not redundant. High-value customers' presence in the company ensures its survival in the market and increases its financial status (Abbasimehr, 2013). Therefore, increasing the priority of these faithful consumers should be the number one objective for the industry, to be profitable and keep the sustainability for the long term.

To further elaborate on that, the business model that the telecommunication sector applies just highlights even more the necessity of this research. Because most telecommunications businesses take the subscription-based models in generation of their revenues. Customer churn then poses a danger to this sector due to its way of conducting business, that is the core reason why the whole industry needs to get out of this problem for staying competitive and grow. This research will use an approach of segmenting customer groups and creating customized retention approaches for the most valuable customers to preserve the valuable income, this research opens a way to telecom organizations which can secure their financial status and stop customers disloyalty.

1.7 Research Methodology

1.7.1 Introduction to Research Methodology:

The research methodology for this study is designed to provide insights and solutions to customers' churn in telecommunication industry. First, the telco customer churn dataset has been obtained from Kaggle, a popular online repository for dataset. Then comes the preparation of the data, at first it involves a lot of care whereby data missing values and outlier data points are removed in order to facilitate data integrity.

After that, the feature engineering methods will be applied to the data, and the attributes related to high-valued clients should be given the special consideration in order to engineer a new featuer that acts as a criterea. One way to calculate the value of customers is CLV (Customer Lifetime Value) and this is an important step, because it helps to detect customers generating a high margin of income based on the attributes present in the dataset.

Segmentation tools are applied in order to separate customers into distinguished categories by using the data set attributes. These specific groups serve as inputs for further

predictive modeling, Then there is the application of the four machine learning algorithms that identify high-value customers with risk of churn.

Evaluation metrics are then being used to rate the models and identify the best model. As a result, the most effective model is recommended to implement into the real world. Likewise, result-oriented responses based on the findings of the analysis are recommended by comprising of strategies that protect high-value customers and reduce rate of churn.

1.7.2 Telco Customer Churn Dataset:

The data set that has been chosen for this research comes from the Kaggle platform, that hosts a large repository of datasets from many diverse fields. Kaggle has shown itself as a data provider with high quality data, it offers users an easy way to get data required for analysis and modeling, and it allows for people involved in the data science field to share their expertise and experiments on dataset in a collaborative manner. The data set is originally created by IBM, the data set has the terms of license stated as “Data Files Original Authors” on the kaggle website. Even though the dataset is associated with the IBM company, it is now widely known in the data community simply as the “Telco Customer Churn” dataset. And that is due to its robustness, depth and importance in the field of data science.

1.7.3 The Rationale for Dataset Selection:

The rationale for the choice of the Telco Customer Churn dataset has multiple reasons, it is perfectly applicable for versatile analysis. First of all, the dataset originally was developed by IBM company and thereafter was updated in later versions, which indicates its validity & most up to date trends. The dataset is very much available on Kaggle, algonside extensive and diverse datasets being uploaded on the platform, Kaggle is a very popular platform for sharing datasets about various fields among data enthusiasts. The telco dataset reached the mark of 1.98 million views and 238.000 downloads on Kaggle, it certainly is known and widely used by various data scientists, including the applications of it for the machine learning methods, like logistic regression or deep learning and many more.

On the other hand, the dataset provides insights and information that make it very convenient for the analysis. These 7,043 rows and 21 variables provide a great data mass that includes a big amount of data related to customer history to analyze deeply. The data attributes present in the data are very clear and full of insights about customers. The "churn" variable is not only the target variable in the model but also the main focus of the analysis because it is considered as a binary outcome in this study.

Another reason why the telco Customer Churn dataset is a good choice for the research, is that it works well with the telecom industry's vision on the customer-driven strategies. Which ultimately highlights that we need to fight customer churn with prediction models. This dataset has full customer profiles which shows us various aspects such as age, service statistics and payment patterns. These are variables that are very much related and are relevant to drivers of churn. Also, this dataset has been particularly recommended by plenty of data scientists and researchers for its use in customer churn research and analysis, this confirms that the selected dataset is very appropriate for analysis which makes it the best choice for comparative analysis.

1.7.4 Data Cleaning:

Data cleaning is also a vital part of preprocessing the data; The predictive models' typical nature picks up on NA values and outliers, these models are just known for their sensitivity, so it's essential to deal with such data anomalies so that there are no biases generated by the models. To solve the problem of the missing data, one needs to implement methods such as mean imputation, median imputation or predictive imputation, the type of imputation depends on the type of dataset or the context, but the mean imputation method will be carried out in this research, because there isn't a significant amount of missing data, as the NA values were only found in one variable in the data, that has only 11 missing values.

Data cleaning tasks will be implemented via methods such as z-scores, and the interquartile range (IQR). Corresponding to that, the outlier detection and data validation are largely the target of these methods. The outcome will be easily understood and obtained because these processes are incorporated in the R program, so it is much easier to catch data outliers or anomalies or thanks to these methods. Basically, what the z-scores method does, how far is every data point from the mean of all values, this way any value that goes far beyond the mean value will be detected. And the I.Q.R. is also very useful in the context of picking out outliers, this measurement is done by measuring the distance of all data points to the quartiles. These methods will surely by relied upon in this project for better interpretations and clarity

1.7.5 Data Preprocessing:

After data cleaning, the dataset goes into the data preparation stage that means standardizing of and preparing the data for analysis. Here, normalization or scaling are performed of the extreme values of the variables so that the influence of the high value would not have effect on the analysis. Eliminating diversity is also an essential process to be conducted between numerical attributes, techniques like Min-Max scaling or Z-score normalization could be applied by adjusting the range and type of data elements based on characteristics of the data.

In addition to this the cleaning process, there will be an implementation of the standardization and data normalization processes, Standardization is the process of transforming data into a standardized format so that users can handle and analyze it, this process will facilitate

feature transformation and data reshaping to retrieve relevant information about consumer value. The structure and format of the dataset will be suitably organized for additional predictive modelling because these preprocessing tasks will certainly improve the analytical process that will come later.

1.7.6 The Feature Engineering:

Feature engineering could be the most important step to take in this research. feature engineering as a technique is essentially made to ensure the simplicity of the dataset and enhancing its interpretability, it's basically all about recognizing particular variables in the data and converting them into new variables. For example, "tenure," "monthly charges," and "total charges" are examples of variables, A new variable may be developed By feature engineering based on these current ones. This happens by doing calculations similar to the Customer Lifetime Value (CLV). so, these variable; the "tenure" and "monthly charges" could be used to calculate the CLV, because the Total charges of the customer since the start of their relationship with the business could determine the customer's worth.

1.7.7 Segmentation:

Segmentation is the most imperative technique used in this research; because without it, there wouldn’t be any focus on high-value customers, segmentation essentially helps in identifying some groups of customers based on their Total charges, tenure, and many more purchasing preferences. The analysis process will go smoothly once there is a targeted segment of customers engineered. Segmentation is very useful in this thesis because the main focus is to know how to retain the high value clients, in order to keep their revenue stream.

Segmentation will help very much in highlighting the highest priority customer to deal with, for example, segmentation will help in making an isolated group of customers with a long tenure and a high monthly charge who are liable to churn, with this group segmented and highlighted, it is then easy to start implementing the retention measures. When these segments of high-value customers are addressed, the predicting process can start, this way the retention of high-value customers will be accomplished.

1.7.8 Handling Categorical Variables:

Managing categorical variables in the Telco Customer Churn Data is also very important. The variables gender and payment method play a very crucial role in in understanding customer behavior, these variables along with the numerical ones are what could very much drive customer churn. So, without handling categorical variables, there will not be any successful prediction of churn. The methods of one-hot encoding or label encoding can be applied to categorical variables to solve the issue of character format, when one hot or binary

encoding methods are applied, that will result in them having a numerical format suitable for the analysis, this is important because machine learning algorithms only understand numerical variables.

1.7.9 Rationale for Choosing R as the Coding Language:

The main reason that R was chosen for coding in this research is for several motives. First, R is famous for having plenty of libraries and many packages that help in machine learning and statistical analyses. R is indeed the perfect tool, for manipulating, analyzing, and visualizing data. Its richness includes a collection of packages designed to facilitate machine learning operations, for example the package “caret” deals with workflows (Kuhn, 2008), “tidyverse” tackles data processing and visualization (Wickham et al., 2019), and “randomForest” performs background for random forest algorithms which directly aims at the study's main task (Liaw & Wiener, 2002).

Additionally, R’s open-source framework brings a lot of people together, firstly, through the community of users as well as the developers, which basically provides plenty of support, resources, and collaboration opportunities. R also provides flexibility in that it can be integrated smoothly with other tools/platforms, popular for data analysis, thus the interoperability, common workflow design findings and ultimately better efficiency comes about (Grolemund & Wickham, 2016).

1.7.10 Application of Machine Learning Algorithms:

The most important stage of the research is indeed the machine learning algorithms application, where the prediction of the churn behavior is especially significant among high-value customers. The preprocessed telco customer churn dataset is then put to test. What this research aims to achieve is to use many machine learning algorithms in order to test out the best model for such robust dataset like “Telco Customer Churn”, This is to overcome the challenge of customer churn in the telecommunications industry. The methods chosen for this comparative analysis are Logistic Regression, Kernel Support Vector Machine, The Random Forrest model, alongside the gradient boosting machine method specifically (XgBoost).

In this research, logistic regression was chosen for it is a simple and interpretable model for binary class classification, it is very basic, and is known to be the most famous model to predict customer churn, In a paper by Nie et al. (2011) when it came to analyzing correlation and economic data, it turned out that logistic regression was actually better at predicting credit card churn compared to decision trees. Logistic regression also has been used in many other studies including churn prediction in the literature and it has been tested thoroughly, here is a key takeaway insight from Surya, P., & Anitha, K. (2022), When they applied the Logistic Regression model on a customer data, it outperformed Random Forest, SVM, and KNN, in

forecasting client loyalty in the telecom industry, the resutlts were 80% accuracy and 95% precision, which is great reason to choose logistic regression in such setting.

Then the SVM will be applied, but first It is vital to understand what SVM does, The Cortes and Vapnik (1995) papers defined SVMs as linear decision surfaces which are fitted within an n-dimensional feature space with applications, that is for the binary classification problems when given data coded as 0 and 1. This support vector networks model, illustrates great generalization capacity and it constantly outdid other machine learning algorithms in several aspects such as Optical Character Recognition. Support-vector networks provided a precise performance for classifying classes of data, with the ability of margin of maximization and the kernel enhancing capability, namely the (kernel support vector machines), This makes a very important tool for picking up on pattern and classification.

The recent contribution of Babatunde et al. (2023) to the research area, is there is this study about Support Vector Machines (SVM) for the prediction of churn customers, the predictions accuracy of this model is claimed to be 95%. Such result makes SVMs an excellent choice for companies to use in preventing churn, because it does not only assess the risk of churn but rather, helps in showing beneficial insights about customer behavior and feedback, and this strengthens the decision-making process by providing relevant information.

The random forest model then is applied, the random forest is a technique that is an ensemble method, which grows several decision trees on separate random subsets of the data set, then combines which in turn makes it very high in accuracy and able to handle various data scenarios (Biau, 2010). Random Forest, in turn, is a powerful tool which may be used to detect nonlinear patterns as well as interactions between variables, and most importantly it avoids the problem of overfitting, therefore it results in high levels of accuracy.

Random forest is a very flexible classifier that can work with non-linear data in a remarkable way, in a paper done by Adhikary, D., & Gupta, D. (2020), The Regularized Random Forest classifier and Bagging Random Forest classifier were applied on a telecom customer dataset, and both turned out to be very accurate for customer churn prediction in the telecom sector, especially the Bagging Random Forest indicated the best result under imbalanced scenarios testing, and under non nonlinear telecom data.

The gradient boosting machines (GBM) model is an advanced ensemble model that is able to build decision trees in multiple successions, GBM is very unique due to its process of finding accurate prediction using the loss function, which is made by an iterative reduction, meaning each iteration removes the mistakes of the previous iteration.

In a recent 2023 paper by Lukita on xgboost, a comparative analysis was applied on several machine learning algorithms to be tested for their ability to accurately predict customer churn. The experiment included Random Forest, Logistic Regression, Adaboost and XGBoost

on a customer churn dataset, primarily in the paper, there was a big focus on the attributes that had the most impact on, meaning there was detailed and specific focus on a couple of variables. When all the ML algorithms were applied, the XGBoost was the best algorithm in the study, with high accuracy in predicting customer churn. This result particularly is of great meaning to this thesis, because the main focus is on the high-value customers, that means the XGBoost's efficiency can analyze the segmenting and also it can give the appropriate retention strategy for the customer group.

The reason that GBM was chosen over decision trees in this thesis, is because when GBM was compared to the known decision trees algorithm, it can be seen that the implementation of GBM may get findings with higher prediction accuracy. According to a highly cited paper in machine learning by Halibas et al. (2019), it was evident that Gradient Boosting Trees outdid other prediction models in predicting customer churn in telecom companies, with oversampling aiding significantly in rocketing performance.

1.7.11 Evaluation Process:

After applying all the machine learning algorithms that predict churn, there has to be e process of evaluation of the how well did the models do. To be able to recommend the best machine learning algorithm, there must be great deal of reliance on multiple tools of evaluative metrics; accuracy, precision, recall, and F1-score metrics will evaluate the algorithms' success in determining churners among customers of high value. The area under the curve of the ROC (Receiver Operation characteristic) will also be used to test the algorithms' performance. Basically, the measure of F1 score allows to calculate the mean of precision and recall so that its accuracy will consider both values. Moreover, AUC-ROC as a performance metric also has the ability of classifying Churn or not Churn cases as the probability points, which is very beneficial for the analysis.

These evaluation metrics will be implemented after applying the prediction models, then the most suitable algorithm will be identified to predict the high-value churn with precision. This algorithm will be a recommendation to all the telecom sector for future consideration. The algorithm will help in desiging the customer retention programs due to its accuracy and unique abilities. The chosen model will be very beneficial as it will be the main reason for the enhanced profitability and sustainability in the telecommunications.

Chapter 2: Literature Review

2.1 Introduction to Literature Review:

The literature review is the core of the project because it will introduce the research question properly in the research field. This part of the research lays the basics of the thesis, which will focus more on the present and past resources. This paper aims to get in touch with practical usage of prediction models in telecommunication companies to deal with high-value customer churn, the research also recommends proactive approaches to retain high value customers that have been identified by the machine learning algorithms. This literature review offers important evidence to support the research question: "How can telecommunications companies leverage machine learning algorithms to accurately predict high-value customer churn and implement proactive retention strategies to minimize customer loss and optimize revenue??"

This literature review shows some previous studies that have concentrated on the use of predictive models by telecommunications companies to identify customers that are about to churn, and specifically a more intense focus on high-value customers. One of the important goals is to conduct an extensive literature review around the field of customer management, retentions and churn, there will be discussions on the limitations, gaps, and previous and present research in the field. This paper will showcase the present findings and the latest literature discussing customer churn in the field of data analysis, there will be suggestions for the improvement or development of proactive retention of customers.

2.2 The Definition of Customer Churn and its effects on telecommunications sector:

In the literature, there are many customers churn definitions which depend on the choice of research or the field. In a paper by Aeri et al. (2023), they had a unique definition for customer churn, it is the stage where customers end their relationships with businesses or companies offering products or services they care about. Moreover, Neslin et al. (2006) had another remarkable description of churn, they stated that churn rate is an indicator of a customer's tenure and lifetime value with a company and when a company goes through customer retention, the company gains has more profitability. Neslin et al. (2006) had also talked in his research about differentiating between the types of churns to come up with an effective churn management strategy.

Yang et al. (2006) had made another point in their remarkable paper about Knowledge discovery on customer churn prediction, they had stated that involuntary churn is a situation when the customer has no intentions to leave, but it happens involuntarily depending on factors such as budget limit, errors in payment, relocation, or service denial of the provider. Hadden et al. (2007) points out that when the number of customers enroll in a business reaches its highest capacity, it becomes tough and expensive to manage acquire more customers, at this point there

is a phase in the business's life cycle where they should make a shift in focus away from trying to win new customers to retaining the most valuable and existing ones.

Customer churn can be known as the cancellation of services by customers, and the loss of customers when they transfer to the competitors. But after all it can diverse definitions depending on those industries. Mozer et al. (2000) points out that identifying churn customer segments is not only a technical process conducted by company personnel, but it is rather conducted by a set of programs for retaining the churning clients. However, the term churn can simply be looked at from this side as, churn presents the number of customers that a business loses over time with another complex factor such as customer value.

Park et al. (2022) sheds a light on the types of churn, they view it in terms of voluntary and involuntary churn, basically, customers who leave consciously through their free will are defined as voluntary churners, and those who quit due to economic factors are known as involuntary churners. The difference between these types should be addressed and necessary retention strategies should be taken to achieve customer loyalty. (Tatikonda, 2013) note that the stakes are very high when dealing with the churning customers' category. They stated that reaching new clients is close to five times more costly than keeping the current customers, with that fact being said, churn is truly one of the major economic problems that telecommunication companies face till this day.

Customer Churn in the Telecommunications sector:

In this competitive telecommunication field, companies are always combatting churn to survive in the market, users on the other end have a great deal of services to choose from. The studies show that high prices are probably one of the most driving factors that make customers think of choosing another telecom provider (Akmal, 2017). According to research in the customer management, it occurred that high network fees from telecom companies and large subscription bills are the main causes of churn (Mehwish et al. 2017).

With these facts presented, the telecom business shows real pressure with dealing with churn, because customers that tend to frequently change their service provider can perhaps lead and influence other customers into the same action of quitting and switching, resulting in more churn (De Caigny et al. 2018), for telecom providers it is their main mission to solve the causes of churning to prevent customers' churn. In regard to telecommunications, there is an astonishing churn rate of approximately 30-35 percent per year, this is a percentage of a proportion of the customers that are terminating their services within one year (Lu, J. 2002), this in turn, may cause substantial declination in revenue and share of the market for the telecommunication providers. Therefore, there is great necessity for the employment of effective churn prediction methods.

Braun & Schweidel (2011) propose to companies to look out for various predictors of churn like quality of service, pricing policy, and availability of options, These predictors help determine the unsatisfied customer groups precisely, if companies helped pinpoint the churn segment, they would then be able to provide customers with the best solutions like service enhancements, price cuts, and improved features and options in the cell phone industry. Telecom companies must act with fast response and collision-free service delivery in handling customer needs and services; this will instill customer loyalty; So based on a comprehensive paper done in 2011, telecommunication companies really need to work with complex data analysis that helps identifying the customers churning and make successful retention plans, rather than relying on the old ways of studying customers in general, because it drains the company's resources and time.

Effect of customer churn on telecom companies:

Apart from the obvious revenue losses caused by churn, its problems are more than that. Some of these problems may be in the long-term consequences, one example is the crisis of brand value deterioration and weak market competitiveness, According to Reuber et al. (2005) this deterioration of the brand name and value certainly affects how customers perceive the company socially, the effect is often negative because brand image becomes inefficient or not recommendable. As a consequence, It seems that it is not only the financial loss that is to be considered, but also other negative consequences. This group of unsatisfied churning customers ruin the network effect, by making a ripple effect when they terminate their service with a company, because customers get affected by their churn behaviors (De Giorgi, 2020). To be clear, Network effect is a phenomenon where the value of a product or a service rises when more people use it.

The focus on High-Value Customers:

High-value customers are known for generating a big amount of money to firms due to their high usage rates and their usual purchase of premium services. Since valuable customers who discontinue a firm's services affect the firm's cash flow, then, companies in the telecom sector must pay attention to this segment of customers when it comes to churn prediction. Gupta and Lehmann (2005) explain in their paper that it lowers the profitability of a company to lose a high-value customer, that's why predicting and preventing the churn of these clients particularly is very important in order to succeed.

Keiningham et al. (2005) says that valuable churning customers deplete the company of market shares and revenue. Therefore, reaching out to the market to gain other valuable customers will be more expensive. Kumar and Reinartz (2016) state that it can take even 25 times as much money to attract a new customer than to retain the existing ones. Morever, there must be retention strategies ready-to-implement put in place by telecom companies, this way the revenue stream will continue.

2.3 Machine Learning Algorithms for Predicting Churn:

Each industry encounters its own unique different challenges when dealing with customer churn, Brown (2000) in his research emphasized that a lot of different types of data is needed in order to make customized approaches that would suit every customer's unique preference, and in turn that will automatically maximize ROI and manage churn effectively. For instance, the subscription companies deal with customer churn in a simple way, customers just get to churn as expected when their subscription expires. But non-contract businesses deal with customer churn differently, they get to conduct some kind of irregular customer behavior observation and analysis, too further understand what is causing churn in that business. Mirkovic et al. (2022) supports this fact by stating that churn is a multi-faced problem and there needs to be special approaches used. To sum up, telecom companies need to deal with churn based on the nature of their customer data and history, this happens by acquiring the different data and then using analytical tools that are built to satisfy the telecommunications business model and the specific traits of the customers.

According to Hadden et al.'s (2007) words, churn analysis is the analysis of the behaviors of customers with the result of the at risk of churn customers' segment. This analysis ultimately works by predicting churn based on the patterns and insights in the data, it's application for example in the telecom companies is to be readily employed to execute a predictive analysis to identify and counteract any possible churn. Based on that, it is evident in the literature that telecom companies that retain their existing customers using predictive models, and apply personalized offers or marketing are the ones that thrive significantly, and often this leads to the churned customers to come back and rebuild their loyalty once again. Accurate prediction and analysis of customer churn are the basis of starting meaningful, resourceful, and effective retention strategies, which are important to maintain revenue and achieving competitiveness in the highly competitive telecom world (Kisioglu & Topcu, 2011).

The number of researchers and analysts that use machine learning algorithms to predict churn became very high in the literature, because machine learning became very clever when it came to dealing with huge quantities of data, it is just astounding to see how predictive models in the machine learning field successfully disrupted the use of traditional mathematical statistics for data analysis, and it can be debated that the huge amount of research that talks about the easiness of machine learning in churn prediction, could be the major reason of why they're commonly used these days with such straightforwardness (Ovadia et al. 2019). Just the spread of machine learning in every field of research and industry like business, medicine, technology and more illustrates how it is easy to apply machine learning in every field to accurately predict a targeted variable (Dankers, 2019), based on that, this encourages for more use of it and integrating it more in the telecom industry.

Machine learning algorithms are a computer programming models that is able to make computers make decisions and execute tasks automatically without the interference of humans to initiate commands. They work through analyzing data sets, which reveal patterns, trends, and

relationships. Researchers or users then use the patterns which are essentially what drives the algorithms to make predictions or decisions, to then act and construct their methodologies (Kreuzberger et al. 2023). The algorithm works adopts different statistical and mathematical methods, that essentially work by conducting iterative functions that improve the model based on the data, this can be called training process, in which the model adjusts its parameters over time to enhance its overall performance. To be precise, machine learning algorithms enable computers to learn from data and enhance their ableness to do tasks that needed human interference (El Naqa, 2015).

Customer churn primarily in the telecommunication sector has always been among the challenging research subjects, due to its importance in terms of impact on the business and profitability. A lot of papers exist further discussing the design and implementation of the customer churn predictions in telecommunications industry, showing that it's not really a new subject taken into consideration recently, but rather a literature scaling back to the early 2000s (Ullah et al, 2019). Some of the models that have been employed throughtout some of the papers in the literature include: Artificial Intelligence (AI) methods such as Random Forest, Artificial Neural Networks (ANN), XGBoost, Logistic Regression, and Convolutional Neural Networks (CNN), They all share their ability to predict with high accuracy the customers who are likely to churn among the others (Ahmad, 2019).

The customer churn prediction in telecom companies was studied by Umayaparvathi et al. (2016) where they applied machine learning algorithms on big data environments. His research shows the immense value of using machine learning techniques in the context of telecom sector. Furthermore, Kiguchi et al. (2022) have put more effort to this topic, especially with testing a couple of models such as logistic regression, decision trees, and random forest, these types of researches enhance the discussion on the machine learning aspect of churn prediction, because not only the theoritical talk about churn prediction and its effect on business will only bring value, but rather the technical aspect of testing all the solutions for churn.

In order to conduct the churn prediction models more accurately and reliably, Ascarza (2018) advised to report both the metric values obtained (meaning the accuracy results) and the degree of data skewness from telecom datasets visualization. This advice is critically important, especially when applying the prediction models, because it eliminates bias and that subsequently causes more precise forecasts, so essentially Ascarza is calling for thorough descriptive analysis before prediction analysis.

Churn prediction is one of the many problems where a lot of testing is needed, to see which machine learning algorithm is the most effective in identifying customers likely to quit. Among these, logistic regression is on the top of the list in terms of the robustness of concluding binary classification (De Caigny et al. 2018). Logistic regression works by modeling the churn probability in a general logistic curve and, as a result, is very resourceful for predicting the

influence of various customer variables on churn probabilities, Technically, it measures the relationship between each variable with the target variable “churn.”

Logistic Regression has been widely applied in the literature. In a paper conducted in the well-known “Oriental journal of computer science and technology”, Sebastian et al. (2017) had established in their study, that when they used a logistic regression model to predict customer churn in the telecom industry based on past data, the model has proven to be effective in helping businesses prevent customer churn and increase revenue. In fact, their accuracy report showed an impressive 80.02% success rate, this research highlights the importance of using logistic regression to tackle the challenge of customer churn in the telecom sector.

Not only regression alone can make a difference in the research field but along with other machine learning methods, some really valuable insights can be generated, for example, in a paper by (Lu et al. 2014), Boosting was used in as a customer churn prediction technique by segmenting the customers into high-risk clusters and enhanced the accuracy of logistic regression model as the basis learner in the research, therefore boosting was used as a way to enhance the accuracy of logistic regression as the basis learner.

In the literature, SVM had recently taken a lot of attention due to it’s robustness and its easy application to binary data. Quek et al. (2023) introduced a new churn prediction method by combining SVM with attribute selection analysis. This new combined model seemed to work better, it was able to defeat more traditional methods like chi-squared and features-based models. That, in turn, highlights the significance of adopting the best methodologies, which in turn provide a balanced approach to customer acquisition and acquisition strategies. With that being said, the new studies revealed not only the efficiency but also the reliability of SVM in churn prediction, and the need to integrate feature selection, in order to develop models that perform well with easy interpretation.

On predicting churn in (Y, N. Ly, T., and Son, D. 2022), research was carried out in the field of churn prediction within the telecom industry, and the model kernel Support Vector Machines (SVM) was observed. Their newly developed model consisted of SVM, and dimension reduction combined with re-sampling, and it outperformed older SVM models, The F1-score was 99% and the accuracy is 98.9%. When it was compared to older versions, it showed great accuracy, such numbers are truly phenomenal, this highlights its effectiveness in targeted and proactive retention planning.

Xia et al. (2008) study customer churn prediction, and they introduce a new perspective by highlighting the ability of support vector machine (SVM) models over various machine learning methods. Other than the typical metrics of accuracy and hit rate, Xia delves into metrics such as covering rate and lift coefficient. This approach says that SVM accurately predicts churn and also identifies potential churners and maximizes the impact of retention strategies. When these multifaced evaluation metrics are in discussion, the study becomes has more valuable insights

about the capabilities of SVM in addressing customer churn, this way it guides the strategic decision-making processes for telecom businesses.

Since random Forrest is chosen for this research, it is then very important to emphasize on their presence in the literature. Decision trees are valuable and insightful when implemented alongside other models into the decision-making systems of an organization, this integration helps bring clarity into the models to predict churn effectively (Lemmens & Croux, 2006), Based on that, it is important to note that random forest is a form of decision trees, The random forest model is an ensemble methodology in which several decision trees are employed on different random sub-sets of the dataset, then the model combines all the classifiers or trees to get an extremely high accuracy and able to handle different data scenarios (Breiman, 2001).

There was a comparison between random forest and a single decision trees algorithm, and it turned out random forests offered a resilience to model variation and was less prone to overfitting, because it is made up of several decision trees that cast votes on the final result. These models perform particularly well when addressing the complex and high-dimensional nature of client data in large-scale data contexts such as telecommunications (Grushka et al. 2015).

Another unique research conducted by (Yang et al. 2023) had a similar focus on high value customer churn, random forest machine learning model was the main algorithm. After the application of 3 different ML algorithms, such as GBM, random forest, SVM, it showed more preciseness in compared to other methods, the model showed a more balanced set of percentages for the random forest, the Accuracy: 0.60745, Recall: 0.552083, Precision: 0.780696. While other algorithms had extreme unbalanced numbers, for example some models had high precision with low accuracy and vice versa.

At a more advanced level of complexity and precision, Gradient Boosting Machines by far and Random Forest present better ensemble methods that enhance predictive accuracy. GBM constructs an ensemble of deficient prediction models, usually decision trees, in a step-by-step manner, whereby each succeeding model makes corrections for the inaccuracies of the prior ones. The method expertly accounts for a broad range of non-linear customer grappling tools that frequently indicate attrition.

GBM shows great performance from a paper conducted by (Gregory, 2018) from Cornell university, GBM was used to predict customer churn, interestingly this study revealed the usefulness of extreme gradient boosting (XGBoost) in making highly accurate customer churn predictions based on large-scale time series data. Gregory was able overcome the problem of feature engineering that needs to meet the requirement of time sensitivity, which is the key to making models with higher precision, it was surely a challenge. The XGBoost showed its success in achieving the first place in WSDM Cup 2018 Churn Challenge, based on this, XBGoost really proved what the gradient boosting machines are capable of in tackling time-

series data challenges in customer churn prediction, this is really significant since it was dealing with timeseries data which is very big, hard to handle, constantly changing.

The latest study of Chen et al. uses an improved XG-Boost-based model for the forecasting of telecom customer's churn. The model demonstrated its efficiency as churn was predicted with an accuracy as high as 96%, and recall was 70% with precision of 32%. It has promoted increased service provider profitability and better service quality. This highlights the importance of using resources like XGBoost to combat the challenges facing the telcos industry (Chen, et al. 2021). Overall, based on the literature and the numbers, The XGBoost model will likely be able to achieve higher accuracy than other models.

2.4 Data Analytics in Churn Prediction:

Data analytics is very important for tackling the problem of churn in the telecom sector, but first the definition of data analytics must be made clear, Mounika et al. (2016) defines it as the analysis of large volumes of data to reveal patterns, correlations, trends, and customer preferences, data analytics is a tool for making businesses gain benefit and make better decision. Data analytics is very beneficial because it helps marketers, product managers, service providers and other businesses find new revenue opportunities to improve their business and have better customer service and gain competitive advantages in the industry.

Regarding the telecom sector, data analytics has a benefit on the telecom companies in many ways, Data analytics in the field of churn prediction has a positive effect on telecom companies, the effects are an improvement in retention strategies, enhanced quality, and customer service, and the most important of all, customer loyalty, which is the main goal from customer churn (Ben, 2020).

There is a highly cited study conducted by the researchers Kim, Jun, and Lee (2014). The purpose of the study is to explain more about why customers churn and how companies use data analytics to improve customer services. The research responded to churn prediction in a different way by analyzing the pattern of connectedness of customers. It studied people's calling each other, they originated a method to understand how the simple idea of churn can spread between customers in a network. Through this understanding, they have opportunity to know who they should target the retention to, and also use these connections to win customers the customers' trust. By merging this social network data with classic customer data, they were able to tell which segment of customers would churn.

Shirazi and Mohammadi (2019) gave importance to the fact that there is a need for using different data sources and types, in order to construct churn prediction models in the finance industry. The data used in the churn models was mainly customer demographics, account information in the past. However, their experiment illustrates the use of unstructured and unfamiliar types of data such as: Webpage hits and phone conversations' logs and the log of web contents, all these data are collected from 3 million retirees, which can make customer behavior understanding clearer. This makes big data analytics an important tool for preventing churn, which leads to a valuable insight, that it is useful to dissect and analyze the retirement journey of a customer forecasting their planned churn.

The telecommunication sector literature always works on understanding the reasons and variables that play into churn behavior, the following is a glimpse of an overall understanding of what causes churn. Telecom service providers collect a significant volume of data to apply churn prediction models. There are demographic data out there, which include age, location, and income, these variables are helpful in understanding the customer base. Call patterns such as duration of calls, frequency, data consumption, and usage of preferred channels (voice or text) have an impact as well on customer behavior and communication patterns, these insights are helpful for directing a certain retention strategy (Choros, K. 2010).

There are other variables to be measured and analyzed such as: service quality, call drop rates and internet speed, these are possible to directly influence customer experience and churn rates. The customer satisfaction surveys, and social media sentiment analysis give clues into how the customer thinks of the services provided. Other variables like the billing information, the payment history and plan details can reveal the affordability issues of the customers. Therefore, different telecom companies will be able to create effective churn models that are able to give the right retention strategies to such customers in order to gain maximum effect depending on their situation (Quach et al. 2016).

There has been an increase in research about customer churn prediction using Artificial Intelligence as of the years 2022, 2023, and 2024. Banu et al. (2022) are one of the most remarkable rescources in proposing an AI-based Customer Churn Prediction Model exclusively for Telecommunication Businesses. They differ from other models of churn prediction by using a method that is called the Chaotic Salp Swarm Optimization-based Feature Selection (CSSO-FS) and the Fuzzy Rule based Classifier (FRC). So, these two models really come together to help in identification of churn. The AI churn prediction model in the study was seen as the next generation of the typical customer churn prediction models. The accuracy of the new model is even higher than the previous ones and can reach up to 97. 25% and 97. 5%. The study by Banu et al. is an important extension of the normal customer churn prediction research in the telecom industry, with these high accuracy numbers, it truly showed the efficiency of AI models in churn detection.

Siddika and et al. (2021) applied a system that adopts machine learning and deep learning methods to identify the reasons of churn, and by applying this method, they stated that the quality of service, price, and customer satisfaction are the key factors or variables that lead to churn in telco industry. Moreover, Wu (2023) focuses on the phone charges, quality of service, and service variety, which also provoke customers to switch to other service providers. Therefore, telecom companies need to focus on their services based on these findings.

Besides, Yuan (2023) had added logistic regression to neural network as a combo for a more accurate prediction with an 80.5% accuracy rate. The model highlighted the major attributes that affect churn, which are: long-term costs, service duration, network services, and contract types.

Consequently, the topics covered in these papers obviously show that “service” quality related attributes and customer satisfaction appear as major churn drivers. This relationship is a reminder that the importance of service improvement cannot be ignored by other factors, and certainly “service” related variables will help reduce the churn rate. Which therefore will improve B2C relationship, service quality, customer loyalty, and satisfaction, which in the long run led to reduced churn rates and increased customer retention.

Zhu and Liu (2021) applied the famous XGBoost algorithm to predict telecom customers' churn. They ended up finding another set of attributes that affect churn, they stated that users with tenure 1-5, e-check utilizers, those with low total charges and high monthly charges are more likely to churn more than the others. While Khan et al. (2021) state that the "data rate", "call failure rate", "mean time to repair", and "monthly billing amount" are the critical drivers that lead customers to switch Telecom service providers. Their comparison of artificial neural network (ANN) and other ML techniques showed that ANN got an increased efficiency of 79% more than others. In fact, the two studies showed that charges and billing amounts were the deciding factors of churn. It shows that definitely factors of pricing and billing are very important for companies to reconsider changing, because customers and their service providers need to settle on a comfortable financial agreement, where the customer is able to pay for his services without hardship.

Data preprocessing is an important and vital stage of machine learning, data preprocessing is integral since it guarantees the reliability of the data. One of preprocessing procedures includes the cleaning, changing, and the classification of raw data into data that produces the most information. Also, data preprocessing makes sure that data doesn't have inconsistencies, missing values, and outliers. So, when a machine learning algorithm is combined with preprocessing, The results will be both accurate and efficient Y, C et al. (2022). More specifically, in telecommunications churn prediction working with historical customer data, demographic information, usage patterns, quality details, and service history are analyzed to give insights into potential churners (Patil et al. 2023).

To further reinforce that data preprocessing is vital in predicting telecom client churn, Coussement et al. (2017) illustrated that preprocessing and preparation of data significantly improve the model's over-all performance. In the given study, logistic regression modeling was compared with different data augmentation based on telecom data. However, the effects of this preprocessing step were magnificent, it led to an improvement up to 14.5%, and a 34% increase in identifying the most at risk churn segment. This implies that data quality is important to increase accuracy of churn prediction, which leads to overall better real-life strategy application.

To conduct a proper prediction on churn in the telecom sector, one without a doubt should make sure his data is of good quality. As missing values may lead to distortion and make prediction models to underperform or lack accuracy. The author R et al. (2023) has a very high (95%) accuracy in churn prediction using a Hybrid Algorithm and Random Forest model, the author dedicated the success of his model to their very careful data preprocessing, especially in handling the missing values and outliers. This is important because one simple skewness, bias, or even distortion in the data could ruin a telecom business decision making process.

It is also known that applying prediction models on customer data in the telecommunication sector requires intensive pre-processing and segmentation possibly, and this is because of the complexity of customer data as it is highly dynamic and dimensional, and data ranges can significantly vary.

In a highly cited paper by Kim et al. (2014), it was shown that the effectiveness of big data is demonstrated in the case of potential churners of a telecom provider. They apply network analysis approach based on call detail records to reveal how churn information spreads through subscriber networks, The authors stated the prediction process wouldn't go through unless there was data scaling conducted beforehand, the scaled data helped to track dynamic relationships among variables for more accurate prediction of churn. Amin et al. (2019) added another

evidence that data scaling is useful, as they go further to show that distance factors are important in the scaling of data and they also show that using this technique, the prediction accuracy can be improved especially in high-call-volume zones.

Data normalization is the main key to the prediction of customer churn in the telecom industry, this point was well illustrated through the development of different prediction models. Mandić et al. (2019) defined normalization as arranging and limiting the size of the attribute values within a set boundary. range of numbers, they also investigate six data mining models, focusing on soft churn detection, they specified that the normalization of the dataset values was implemented in order to get a proper balance of the influence of all the variables in the data, this of course lays the groundwork for a proper prediction.

Feature engineering is a crucial part of model making where it is necessary to preprocess raw data to get relevant features that actually matter and boost the performance of the model. Feature engineering improves predictive modeling by aiding in data dimensionality reduction, which makes the data and the model complexity improves (Fan et al. , 2019).

In a research about feature engineering, the authors Huang et al. (2012) create a set of features based on a large data for land-line customer churn prediction, the newly engineered feature include demographic data, call details, and billing information, the authors test six prediction techniques and they ended up finding out that these features improved the prediction which demonstrates the effectiveness of meaningful creation of variables. The second paper talks about customer churn, the authors have created a new feature set along with three new methods of window input being used. They found that the input window techniques alongside the newly engineered feature have worked well with their applied prediction models which are decision trees and neural networks to predict land line churn (Huang et al. 2010), this is very insightful for churn prediction, as prediction is not solely tied to machine learning models but also feature selection and engineering.

The paper of Liu and Zhuang (2015) which focused on customer churn prediction in the telecom industry stresses the fact that customer segmentation is essential. The authors had applied segmentation and misclassification cost method applied on a telecom churn data, misclassification cost which is the weights that are applied to the particular outcomes, these are weights that the model considers and may affect the output as a way of preventing loss. After applying the segmentation and the misclassification, the following findings were found, after segmenting the customers based on certain criteria, the decision tree method brought the highest prediction accuracy. Second, the C5.0 model with customer segmentation has proved a higher accuracy and coverage. Ultimately, accuracy is the only perk of C5.0 model with misclassification cost has higher coverage in data.

Hui et al. (2009) have shown in their research that the customer segmentation method is the key in telecom churn prediction. Hui's study uses fuzzy kernel c-means clustering to find the segment of high-value customers, which are beneficial for making the churn prediction. Hui's application of the SAS data mining technology model achieved great accuracy in identifying high value churn customers. Conveniently, Liu and Zhuang (2015) conducted a customer segmentation and misclassification cost methods in the churn prediction process, which is an exercise that enhances the model's ability to recognize churn customers and assist in focusing on a specific segment, the model of Liu and Zhuang is more accurate, more efficient, and more comprehensive than the models that do not use segmentation and do not take the cost of misclassification into account.

Kaggle is a data science and machine learning community that acts as a virtual collaborative space for people to share code and ideas about data, ML model, features, and loss functions (Li et al. 2021). Kaggle is an important tool that improves data analysis research by offering both large collection of datasets and classification methods/algorithms. For instance, (Amjad et al. 2022) showed through data mining on Kaggle data about students, The authors had applied many ML algos, Decision Tree was 90% accurate, AdaBoost 89%, Logistic Regression 88%, SVM 86% and SGD 84%, and Random Forest was the highest with 98%, this further aids the use of data from Kaggle as a reliable repository for research, in this study, the Telco Customer Churn dataset from Kaggle is used for assessing the performance of different ML methods for predictive analytics.

The R programming language is essential for making churn prediction because it has effective packages for developing the work with the big data and creating representations of the results. For example, R language was used to create an improvement in logistic regression model by predicting the telecommunications customers who might churn with 80.02% accuracy as reported by Sebastian et al. (2017). To further explain the significance of R in churn prediction, especially for the insurance companies such as LIC, AVIVA and MAX LIFE, Patil and Chavan (2017) used data from insurance firms to make their prediction. They applied four models to identify potential churners, R was where used to handle big data from these companies, in the research there was numerous graphs and visualizations that helped the companies determine the variables playing into churn, this demonstrates how R can be used in data analysis as well as in building models for churn predictions.

2.9 Model evaluation of machine learning algorithms

Model evaluation or evaluation metrics are a vital part of this study, because it determines the suitability of the Machine learning models to telecom company use, essentially model evaluation in data analysis seeks to determine how well a complex model can be used in decision-making situations and the recommendations that should be made concerning the model (Pendharkar, 2009).

The evaluation models are important for the churn prediction, especially in the case of dealing with imbalanced and big data sets. Wu et al. (2021) have findings in their research that note the importance the F1-score as a metric, in the study, F1-score showed that Adaboost model achieved 77.19% In Dataset 1 (an imbalanced data) and the F1-score is 63.11%. While Random Forest was accurate in Dataset 2 with an accuracy of 93.6% and the F1-score of 77.20% All these elements and scores are crucial when it comes to adjusting parameters of a model and optimizing the customer retention, the authors specificized that the metrics are needed for marketers and decision makers to be able to make retention strategies more precisely.

The usage of evaluation metrics for telecom customer churn prediction is shown in the following paper, according to (Xia et al. 2022). Their study outlined that the gradient boosting

tree algorithm seemed to be the highest performing according to precision, recall, F1 score, and AUC metrics. In this way, they compared the above-mentioned metrics among different kinds of machine learning models to decide the best algorithm for the prediction of churn and said that the evaluation metrics are very useful in the improvement of the prediction preciseness and highlighting the best model. According to the papers presented about model evaluation process, there will be an application of evaluation measures like precision, recall, F1 score, and AUC to measure and assess the 4 selected ML algorithms.

2.10 Proactive retention strategies in telecommunications sector:

Proactive retention strategies is a method that involves using tools such as predictive analysis and analysis that highlight churn customers who should be retained through intervention techniques before they leave. These measures include customized promotions, flexible reward systems, and interaction with the customers for higher level of satisfaction and optimum customer retention. As stated by Xia, Cui & Duan (2022), proactive retention enhances machine learning models in analyzing churn and then act on the targeted segment, which in this study is High value customers, with strategies that meet customers' needs and satisfaction.

In a paper by Ban and Keskin (2021) There was a comparative analysis between multiple regression analysis and logistic regression analysis, the authors explored in their research how prediction models were used for personalized dynamic pricing in in the telecom industry. They showed that when companies designed different individual prices for their customers based on the highly dimensional data, it helped in reducing customer churn in the telecom firms. The findings were: Customers who are on custom optimal pricing plans, that matches their preference, turned out staying with the service provider, which was the complete opposite to those on non-optimal or personalized plans. Such findings verify the important outcome of the customer retention plan, which consists of personalized goals, as well as data-driven strategies.

Applying organizational retention strategies that are supported by the machine learning analysis enhances churn and revenue. For example, there was study by Hargreaves (2019) in which a logistic regression model helped in the reduction of churn by 40% and an improvement in revenue base by $893,908.50 in the sector of telecommunications. There are surely factors that play into churn, in the paper, some of these factors include, FiberOptic, Monthly Contract, DSL, One-Year Contract, and Streaming Movies, these factors were analyzed in order to develop recommendation strategies which would help in the retention of the clients. This simply points to the fact that the use of data analytical tools greatly helps in designing retention strategies.

One of the tools of customer retention is CRM, CRM systems are crucial to telecommunication firms due to high customer churn rate especially in Zimbabwean environment. CRM has the potential of generating improvements in the communication between a company and its customers, Viriri and Phiri (2017) noted that the adoption of CRM in Zimbabwe's telecommunication industry is still very slow, thus has led to low customer loyalty and high churn rates. However, the generalization of the use of the loyalty programs is raising questions about their efficiency, therefore there is a need to develop a stronger framework for CRM strategies that can make retention in customers and growth in companies.

Burez and Poel (2007) show that customer retention through targeted action like sending satisfaction questionnaires to high-risk churners can be very beneficial. To do so, they created and compared different churn prediction models with data from a European pay-TV company, such as Markov Chain Models and Random Forest and Logistic Models. They backed it up with an experiment showing that customer satisfaction questionnaires and other targeted retention actions double profits and therefore it shows the efficiency of CRM strategies in tackling customer churn and boosting revenues.

To support the effectiveness of the tools that assist in predicting churn such as using machine learning algorithms, the model assessment, and strategies for increasing the retention rates. Ben et al. (2020) built an advanced churn prediction model using the Markov Chain Model which is a more flexible model because one can include non-constant retention rates within the model and also has consideration of the behavior of a customer individually, which is really smart. This testing conducted by MATLAB, showed great insights. The study shows the practical use of the model, it aids in identifying optimal forms of churn reduction through retention, which makes the research relevant for researchers and practitioners seeking to create more precise retention strategies.

Chapter 3: Exploratory Data Analysis (EDA)

3.1 Introduction to The Exploratory Data Analysis

This chapter of the thesis mainly covers the analysis of the Telco Customer Churn dataset. There is great importance in describing and dissecting the relationship between features and the nature of the data, also the analysis will help to determine the kind of data to be dealt with to help determining the kind of modeling that will be done. Pattern analysis is also a vital component of this research especially in the predictive modeling part, because it will help in highlighting the patterns and anomalies that may are hard to see when working with huge datasets. The result of going through this analysis is uncovering critical patterns of customer behavior (Cooksey, 2020), the patterns will possibly help in making customer retention strategies for churn and steer the direction of the preprocessing and modelling of the data. There will be different descriptions, figures, visualizations that will illustrate the nature of the data collected about the customer information and behavior.

3.2 An Introduction into the Telco Customer Churn dataset

The Telco Customer Churn dataset was subdivided into the following customer demographics information, this information has factors that influence customer churn, The main goal is to predict customer churn potential by taking into consideration the below features. Here is a brief overview and the structure of the dataset:

Structure of the Dataset:

• customerID: Customer number, the telecom company assigns a unique code to clients in order to identify each client on account of the organization.

• gender: The customer's gender could be Male or Female.

• SeniorCitizen: A variable that shows if the customer is a senior citizen (1, 0)

• Partner: To determine if customer has a partner or not (a partner = yes; no partner = no);

• Dependents: Customer dependency (Yes means Customer has dependents, no means Customer has no dependents.)

• tenure: Is the total amount of months a customer stays with the company.

• PhoneService: Yes/No: Whether the customer has a phone service or not

• MultipleLines: There could be multiple lines for a customer (Yes, No, No phone service)

• InternetService: Customer's internet service provider: (DSL: YES, Fiber optic: YES, NO)

• OnlineSecurity: If the customer has online security service or not (Yes, No or no internet service).

• OnlineBackup: The customer has or not online backup service (Yes, No, No internet service).

• DeviceProtection: If the customer has device protection or not (Yes, No, No internet service)

• TechSupport: if customer has tech support service: Yes, No, No internet service

• StreamingTV: Whether the customer has streaming TV service (Yes, No, No internet service)

• StreamingMovies: Whether the customer has streaming movies services (Yes, No or No internet services).

• Contract: A variable showing the type of contract (Month to month, One year, or Two year)

• PaperlessBilling: Whether the customer has paperless billing? (Yes or no)

• PaymentMethod: The customer's type of payment (Electronic check, Mailed check, automatic bank transfer, Credit card (Automatic)).

• MonthlyCharges: This is the average price which the customer is charged per month.

• TotalCharges: This is the total amount that the customer is charged over lifetime.

• Churn: Whether the customer churned or not (Customer churn was indicated as Yes or No)

3.3 The Summary Statistics of Numerical Variables

Tenure is a numerical variable that measures the time a customer spends with the telecom company in months. MonthlyCharges is the amount due monthly, TotalCharges is the cumulative charges of the clients. The following summary statistics is calculated for by the following statistical variables: mean, median, mode, standard deviation, minimum, maximum, and interquartile range (IQR). The following information was mined by using the summarise function from R.

1. **Tenure**
    - **Mean Tenure**: 32.37
    - **Median Tenure**: 29
    - **Mode Tenure**: 1
    - **Standard Deviation**: 24.56
    - **Minimum Tenure**: 0
    - **Maximum Tenure**: 72

- **Interquartile Range**: 46

2. **Monthly Charges**
   - **Mean Monthly Charges**: 64.76
   - **Median Monthly Charges**: 70.35
   - **Mode Monthly Charges**: 20.05
   - **Standard Deviation**: 30.09
   - **Minimum Monthly Charges**: 18.25
   - **Maximum Monthly Charges**: 118.75
   - **Interquartile Range**: 54.35
3. **Total Charges**
   - **Mean Total Charges**: 2283.30
   - **Median Total Charges**: 1400.55
   - **Mode Total Charges**: 20.20
   - **Standard Deviation**: 2265
   - **Minimum Total Charges**: 18.80
   - **Maximum Total Charges**: 8684.80
   - **Interquartile Range**: 3384.38

Tenure

The first numeric variable is the tenure with a mean of 32.37 months that depicts the number of months the customers have been with the company. This means that, on average a customer stays to use the company's services for about 2.5 years. The median is slightly lower at 29 months so it means that there is a likeliness to get half of the customers with a tenure of less than 29 months and the other half with more than 29 months. The mode is 1 month, pointing to the fact that a large subset of users is likely to cancel their subscriptions within the first month of subscription. The standard deviation of 24.56 is found, It shows that there is a significant distribution of customer tenures starting from at most 0 month up to 72 month. The degree of the variability of this measure is seen in an even greater value in the interquartile range, which is 46 months.

Monthly Charges

MonthlyCharges' mean is centered right about 64.76, it can be concluded that the average customer is charged around $65 every month. And the median monthly charge is $70.35, which means that the charges are slightly right-skewed, which means that more customers are paying below the mean than would be expected. The mode however is much lower at $20.05, which indicated that the most typical total monthly charge is around 20 dollars, which is most probably the price of the base package provided by the company. The standard deviation 30.09 shows that there is a wide variability in the range of monthly charges, it varied from a low of $18.25 with a possibility of reaching $118.75. Also the IQR is 54.35 indicated that there is a wide range of fluctuations that can be seen in the middle fifty percent of customers, from this value it can be concluded that customer are choosing different service plans and using different amounts of services.

Total Charges

The TotalCharges variable which measures the total amount of charges made to customers over the course of their subscriptions, has the following statistics; The mean is $2283, and the median is $1400.55, which tells us that half of the customers have total charges

greater than $1000 but less than $1400. The mode stands out very much with the value $20.20 which implies that a very huge part of the customers have very low affinity, and therefore presumably many of the customers in the distribution will have very low total charges, and that could because they churn early. The standard deviation is 2265, the amount of total charges spread very widely with values ranging as low as $18.80 up $8684.80. The IQR of 3384.38 may suggest a variability that takes over a broad range, shows the high variability of the total charges to customers, which is defining the different amount of usage by customers and their tenure or stay in the company.

3.4 The Descriptive Analysis of Categorical Variables:

The following is the categorical variables in the datai: **gender**, **SeniorCitizen**, **Partner**, **Dependents**, **PhoneService**, **MultipleLines**, **InternetService**, **OnlineSecurity**, **OnlineBackup**, **DeviceProtection**, **TechSupport**, **StreamingTV**, **StreamingMovies**, **Contract**, **PaperlessBilling**, **PaymentMethod**, **Churn**. The frequency distributions of the attributes mentioned are written below, the following statistical calculations were done by using the R summarise function.

The frequency distributions of each categorical variable were analyzed and the results show an insights about the client base and their use of the services. The gender is almost equal with 49.5% female and 50.5% male customers. It is also obvious that the majority of the customers 83. 8% are not senior citizens, while they hold a value of 16.2% as senior citizens. Also, 51.7% of the customers have no partnership. 30% of customers have dependents, while a large percentage of 70% do not have dependents. This information assists in understanding the kind of clients the firm has and therefore helps in designing targeted retention methods of that can prevent churn.

Regarding the services offered by the telecom company, the most popular one is the phone service with 90.03% of customers subscribed to phone service, followed by 9.68% of customers who stated that they do not subscribe to phone service. It also appeared that 48.1% of the customers do not have any multiple lines and 42.2% have multiple lines, and the rest which is 9.68% does not even have phone service. According to the internet services offered by the company, it turned out that 34.4% of customers use the DSL service, and 44% use the Fiber Optic service, and the last 21.7% do not have internet connection at all. It was also observed that value added services such as online security, online backup and device protection, a large number of customers do not have these services or have declined such services with 49.7%, 43.8%, and 43.9% respectively not having these services. This says that the core services are used by most of the customers, but other services have different usage rates which of course has an effect on customer satisfaction and churn.

Other important factors are the contract types and billing preferences, which may reveal many insights about the market. First, a large number of customers 55.0% is subscribed to the month-to-month contracts, which proposes that the customers prefer the flexibility, while flexibility is important to customer satisfaction, it may lead to churn of customers for the lack of commitment, 20.9% are on one-year contract and the other 24.1% are on the two-year contract. The majority 59.2% of the customers have chosen paperless billing where versatility is the new trend. About the payment methods, 33.6% of customers use electronic checks, and the second most popular option is the mailed checks with 22.9%, and the ones that prefer bank transfers have a 21.9% percentage, and the ones with credit card payments are 21.6%. It was

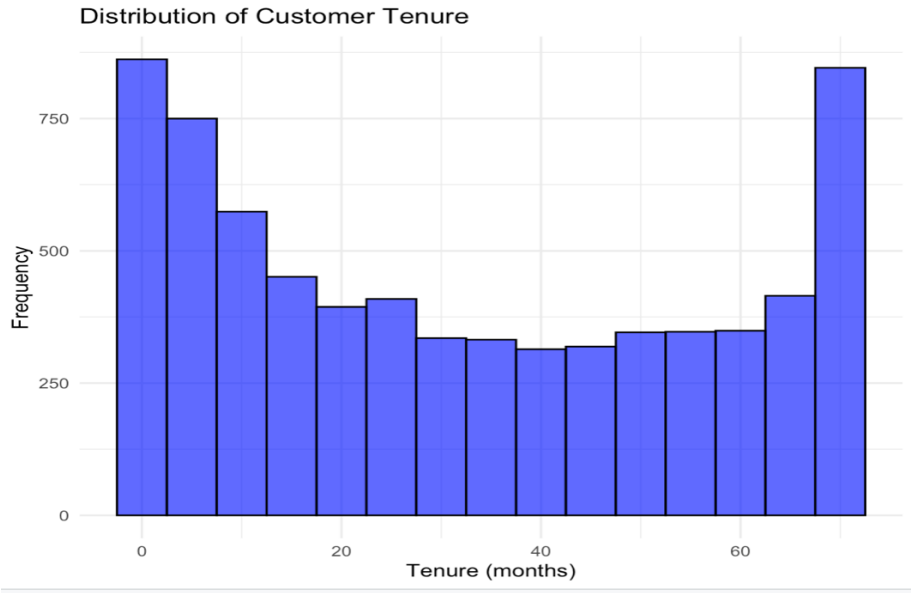
seen in the dataset that the churn attribute had 73.25% customers not churning (no), whereas 26.5% customers churn (yes), which reveals high customer churn percentage among customers, that can only be solved by appropriate segmented, targeted retention campaigns.

Visualizations of Numerical Variables:

Introduction:

In this section, several visualization methods need to be graphed to analyze the correlations between the numerical variables in the Telco Customer Churn dataset. These visualizations include the histograms, box plots, and scatter plots that help to analyze the variables and their connections with churn. These relationships are important to keep in mind when applying the machine learning models for identifying high-value churn, to properly tune the parameters, and make the necessary decisions in data preprocessing techniques.

Figure 1 Histogram examining the Distribution of Customer Tenure

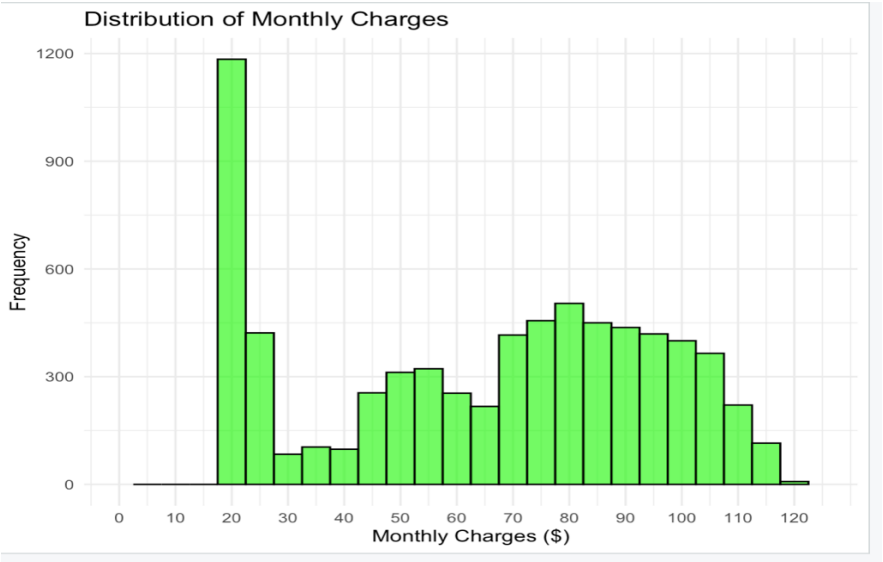


Interpretation of Figure 1:

The provided figure shows many aspects of the customer tenure distribution and how long customers stay with the company. It can be seen in the beginning there is a high peak at the start of the curve which means that there is a large number of customers who get to churn in the first couple of months of their subscription. This goes to show that it is crucial to enhance the customer experience and their level of interaction during this time of their subscription to minimize early churn. After the first peak, it's obvious there is a consistent decreasing in tenure, the distribution shows a normal curve in the middle part of the most frequent months, which means a group of customers in the 20 to 60 tenure range have moderate loyalty. There is also another peak in the end of the tenure (around 70 months) showing many long-term customers who keep being loyal which means they have a great potential and can become a valuable segment because of their great income. These patterns and peaks show that early tenure

customers should be taken care off more, long-term customer loyalty needs to be achieved in order to have the highest retention rates and increase CLV.

Figure 2 Histogram Distribution of Monthly Charges

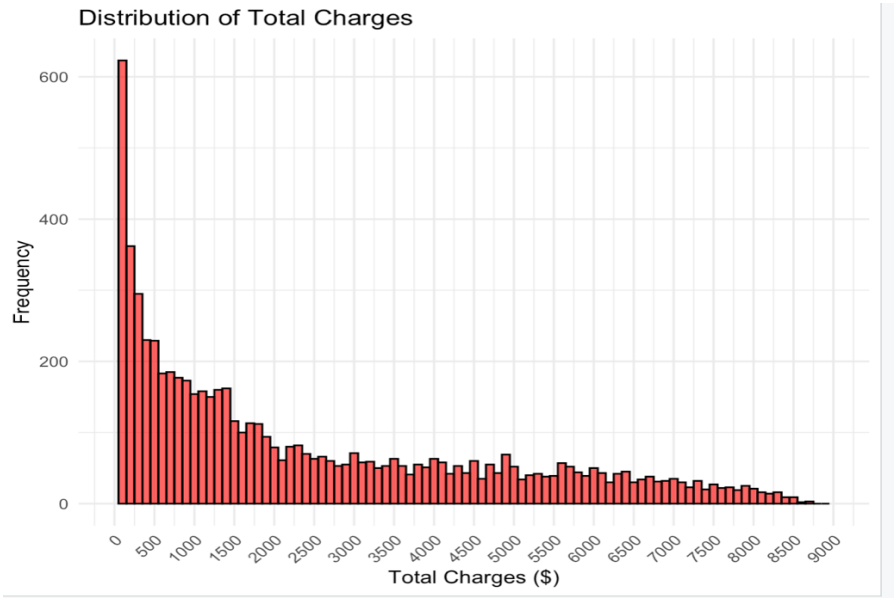


Interpretation of Figure 2:

In the histogram of the monthly charges charged to customers, it is seen that most of the customers' charges are $20 every month, it's clear that there is a large spike in the graph that shows many customers have low charges, it could be a base default package. However, there are still many evenly spread more slightly expensive subscription from $45 to $110, with noticeable premium customer segments of $70-110. These is a clear diversity in the pricing structure between customers in this data, it is alo evident that the 70$ to 110$ subscription group could be identified as the high-value group

The data suggests that the company should implement targeted retention strategies, by offering reasonable prices and also good service quality for the consumers with low charges, in order to capture their attention, while at the same time they should try to improve on the quality of service and benefits offered to high-charged consumers, so that they can retain their clients and maximize the profits financially.

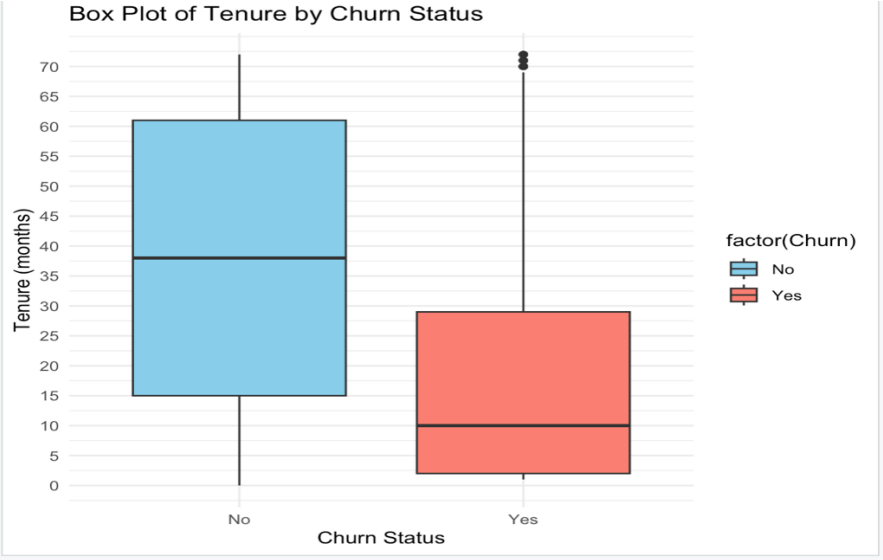
Figure 3 Histogram: Distribution of Total Charges



Interpretation of Figure 3:

This is the distribution of the total charges, here the graph shows that a large number of customers pay little number of total charges, many are concentrated within $0 to $1000, showing low spending customers, which is related to the $20 dollars base package. It is found that as the total charges goes up, the frequency of customers become less throughout the graph and slowly decreases towards $9000. It also points to the fact that there are less customers who are high value but the total value they contribute is substantial. The high charge segment contains customers with total charges above $6000, which consists of a smaller number of consumers. These high-value customers are more likely to stay as users of the long-term, which makes that segment very important for the company's financial health.

Figure 4 Boxplot of Tenure by Churn Status

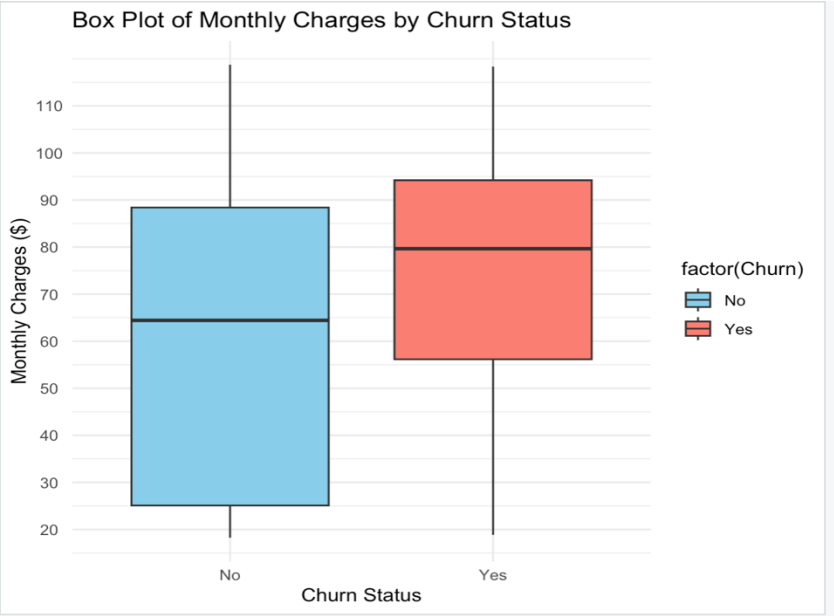


Interpretation of Figure 4:

According to the tenure by churn status boxplot, Customers who churn (Yes) in this case have had a shorter tenure and on average a median of approximately 10 months, the IQR also show most customers churning within the first 27 months. Meanwhile, the customers in the non-churned group show a median value close to about 37 months and their IQR goes up to over 60 months, which shows there is a long-term customer base.

The churn customers in the plot are the ones who have been with the company for a short period of time, The median of tenure is 10 months, The non-churned loyal customers have almost twice of the average of tenures that stand at around 35 and 40 months at the median. It is known that high value customers are prone to stick to a company with more longevity because they do the most frequent spending (Rob et al. 2005). When a company starts retaining existing customers to increase their overall duration of stay with the company, especially if they spend a lot, the company can develop a base of customers that is loyal and maintains a long-time relationship with them. This reduces churn and also optimizes Lifetime customer value.

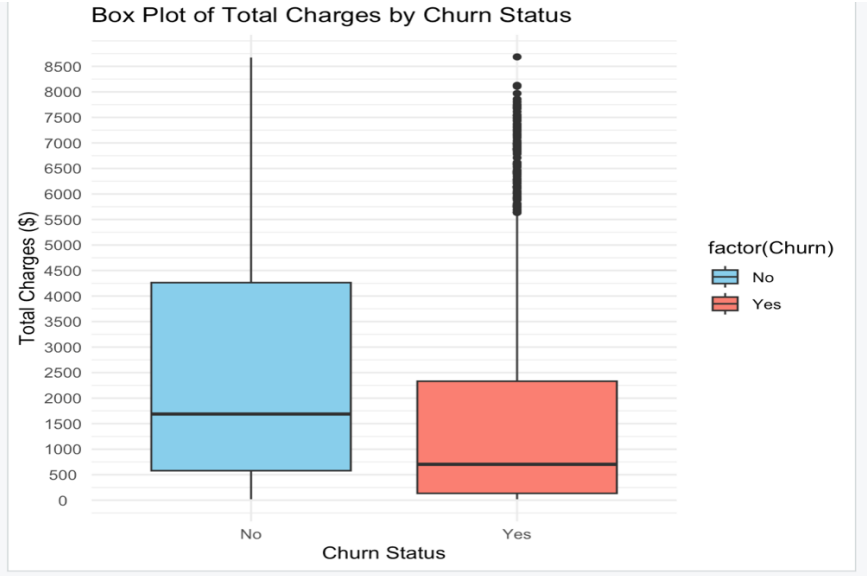
Figure 5: Boxplot of Monthly Charges by Churn Status



Interpretation of Figure 5:

When churn is compared regarding to the monthly charges, it is seen that customers, who churned have expensive monthly charges more than the customers who didn't churn. Churn customer monthly charges have approximately been $85, but for non-churned customers it is $65. The IQR value is tighter for the churned customers, which indicates that the distribution of their mean monthly charges has a short range. It is seen that churn medians are in the upper part of the graph, which means that the higher the charges, the higher the churn. This insight calls for focused retention strategies, so that high-value customers are provided with additional discounts, a decrease in their package payments for their loyalty.

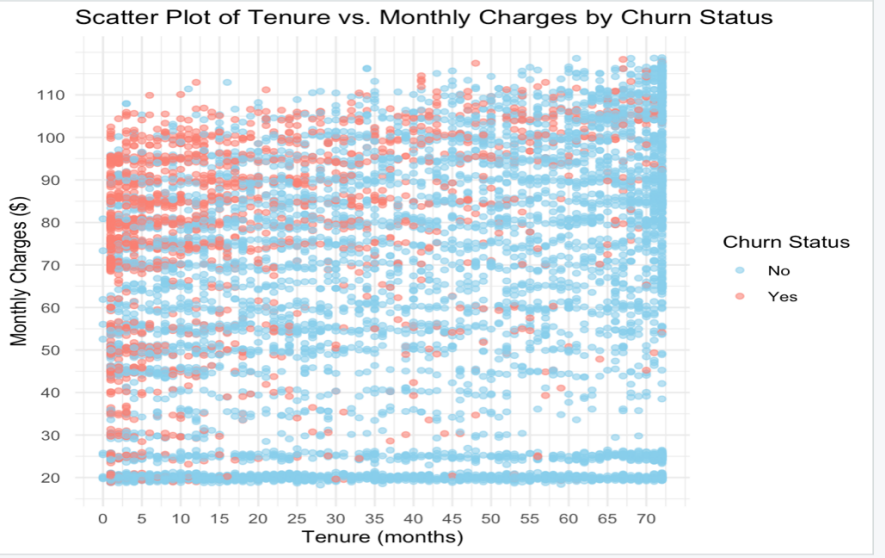
Figure 6: Box plot of Total Charges by Churn Status



Interpretation of Figure 6:

The box plot shows that the number of total charges was higher for non-churned customers. As it can be seen, the customers who have not churned (No) have a higher median total charge of $1600, and the customers who churned (Yes) have a median of about $750. Also, the box plot demonstrates that the IQR of non-churned customers is higher and wider than churned customers, it means that loyal customers continue to use the services and are charged more than $4,000, while churned customers have a low total charge ending at 2250. This means the lower the amount charges to a customer the lower the churn rates, and the consumers that pay more are likely to stay with the company. These insights should persuade telecom companies to invest more in making long-term relationships with high value customers that spend a lot, by offering valued services to enhance customer loyalty.

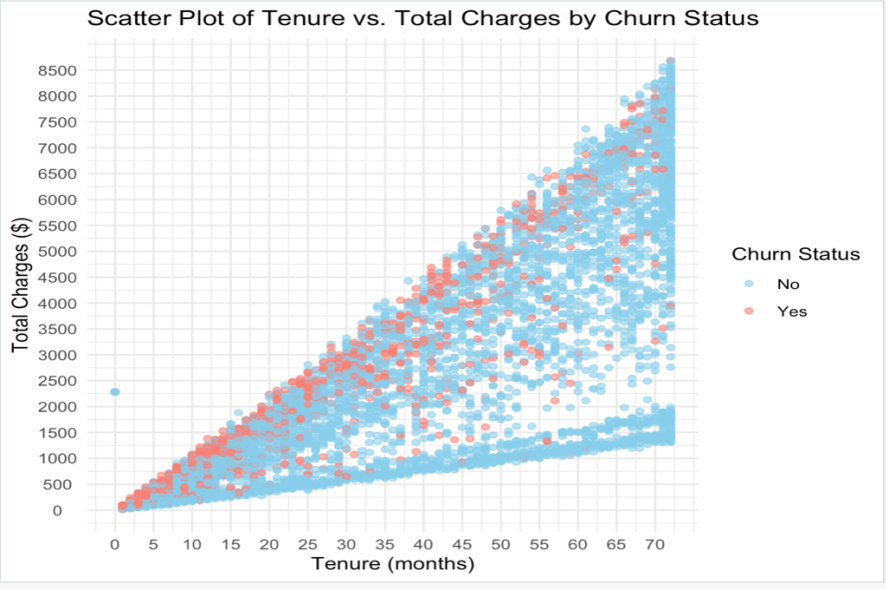
Figure 7 Scatter plot of Tenure vs. Monthly Charges by Churn Status



Interpretation of figure 7:

This scatter plot analyses the effects of tenure on monthly charges, it is clear that non-churned customers are spread out. This is a good point, most of the red churning customer dots are very concentrated when they have high monthly charges in the beginning of their stay. There is a contrast also that notes that non-churned customers are generally characterized by their longer tenure and are found below $60 in the monthly fees. This pattern indicates that customers are satisfied and are more prone to stick to the company when their monthly fees in the beginning are reasonable, because if its high, then customers would feel not committed to their subscription to the company.

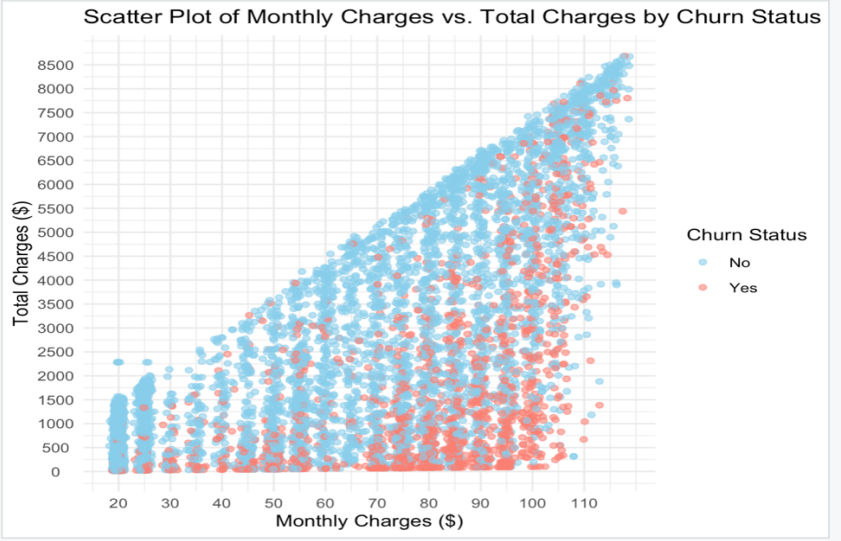
Figure 8 Scatter plot of Tenure vs. Total Charges by Churn Status



Interpretation of figure 8:

First, there is a consistent flow and a correlation between tenure and total charges, which shows that when the number of total charges rises the customer tenure also increases. This is as expected because tenure customers with more months on the plan have more opportunities for charges to accumulate over time. Customers who churned are seen across the tenure distribution but with a little higher concentration in the lower end of tenure (0 to 30 months) and lower total charges range also. Therefore, it's easier to spot that non-churned customers have higher total charges as well as longer tenures, and it suggests that customer lifetime value is directly associated with total charges. This pattern shows us that it is possible to charge a customer more in the future if he or she stays with the telecom company for a long time. This insight helps in the goal of defining high-value customers and segmenting them later in the analysis, and those customers with long tenures and high total charges are most likely to stay and generate flow of income over time.

Figure 9 Scatter plot of Monthly Charges vs. Total Charges by Churn Status



Interpretation of Figure 9:

There is a high correlation between the two. As expected, total charges tend to rise as the monthly charges rises, churned customers seem to belong to the group that pays high monthly charges, typically when the charges are more than $70. It is not right to say that the high monthly charges lead to high total charges because the tenures can be low, or even different. The average of non-churned customers is denser than that of churned customers along the x-axis which represents the total charges, which says that non-churned customers have a tendency to get bigger total charges as their customer tenure goes up. This pattern demonstrates that high monthly rates are potential threats to churn, although loyal customers may have more charges, which shows another reason to retain the high value customers, to make sure they contribute their full potential over a long period (tenure). This interpretation emphasizes on better segmentation strategies to persuade high monthly charge consumers into loyal consumers.

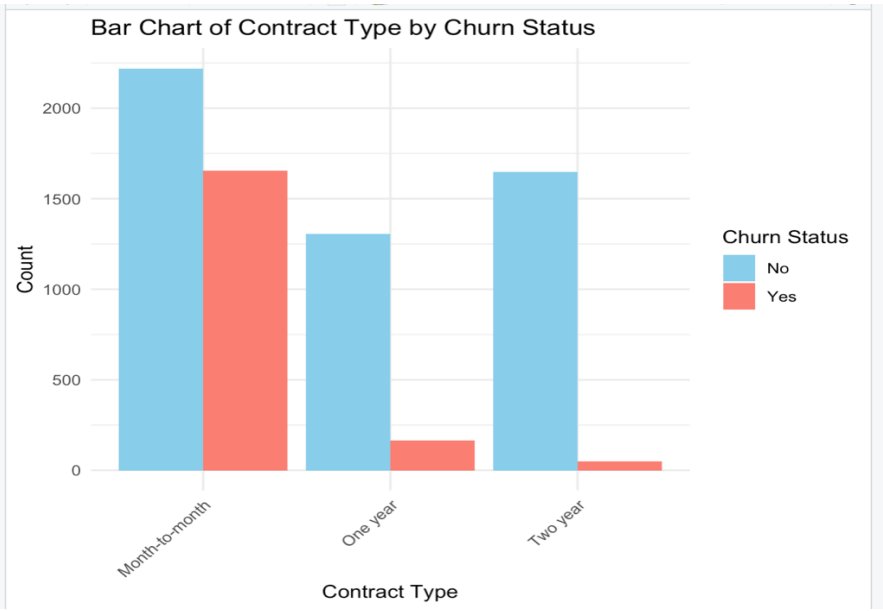
Prioritizing Key Variables in Predictive Modeling:

There are some predictor variables to be prioritized and integrated for customer churn prediction modelling, for example: tenure, monthly charges, and total charges have been considered as variable that are highly important because they are correlated to customer's value and loyalty. Tenure essentially says how long a customer has used the company's services or products, and this will give a clue on potential churn rate. Monthly charges express the revenue per customer. Total charges are also the sum of charges charged over the stay of the customer with the company, this is very linked towards his value to the company. These variables are important for understanding the customer behavior so that we can then apply these variables into the machine learning techniques, including logistic regression, random forest, and gradient boosting and kernel SVM, to highlight the risk of churn for better personalized retaining decisions. Additionally, time is also spent on tuning the model parameters in order to bring the most significant variables into the models, which would help in building the model that is tuned well to understanding the factors that make high value customers churn. However, in the following parts of the thesis, there will be other variables such as the type of contract, the method of payment, and the type of internet service will be included in the model for analysis. Overall,

the focus is to fine-tune the models with respect to tenure, monthly charges, and overall total charges in order to give the best values, among other categorical variables too.

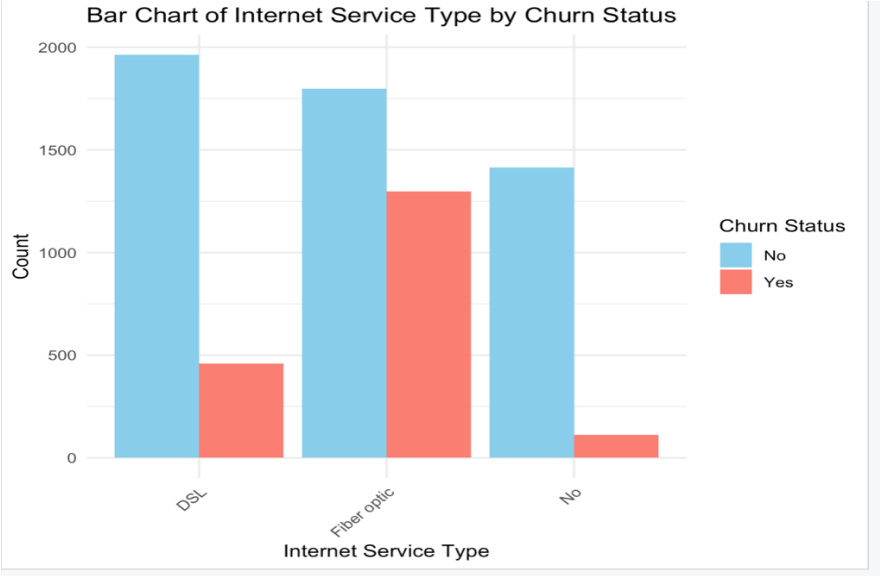
Visualizations of Categorical Variables:

Figure 10 Bar Chart of Contract Type by Churn Status.



In this bar chart about the relationship of contract type by churn status, the overall concept is the reflection of an interesting trend about how customers with different types of contracts behave. So those customers with a month-to-month contract seem to have the highest churn rate, and the number of customers that stay (non-churners) are distributed among all types of contracts, but mostly they seem to be dominant in the one year and two-year contracts and that is probably due to the loyalty and sense of commitment in longer contracts. From this, it could be understood that customers enrolled in month-to-month contracts have a higher churn rate, which would make some sense as these customers are not obliged and have short term contracts. Specifically, the churn rate seen in the two-year contracts is at the lowest level, which suggests that if companies aim to persuade or provide valuable customers long-term contracts, it can help very effectively to keep and retain customers.

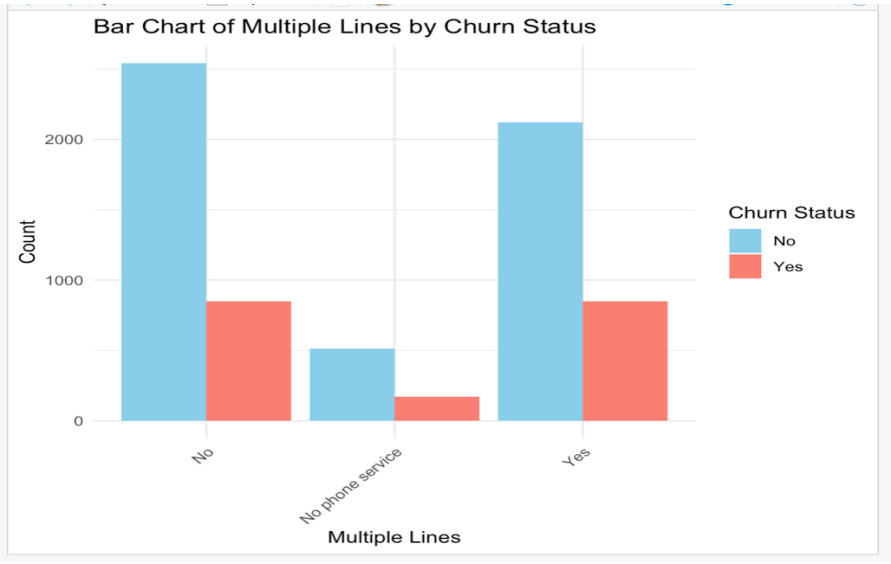
Figure 11 Bar Chart Internet Service by Churn Status



Interpretation of Figure 11:

It is quite clear that customers with fiber optic service get to churn with a lot. The chart says that the even though the fiber optic service may provide better performance, it can also be the cause for people churning, and the reason could be the higher charges that customers have to pay for fiber optic service, or that there could be issues with the service, or even competition pressures. About the churn segment, customers with the DSL service have much lower churn. It is also seen that there is a small group with the "No internet service" has a very low churn level, this group might consist of an uninterested group of customers that do not prefer to switch their Internet Service Provider. All these results highlight that telecom companies need to focus on fiber optic customers, by providing better quality and price of service to improve the chronic churn in this service.

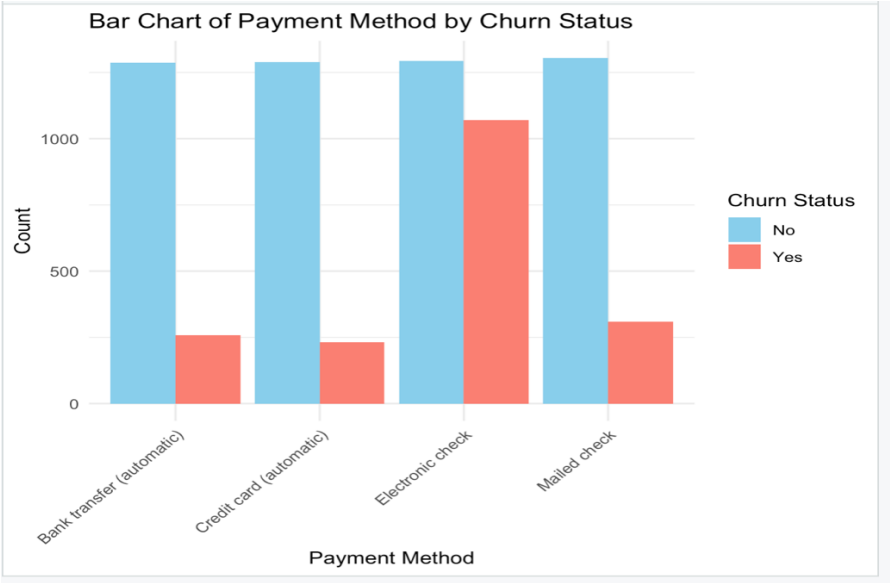
Figure 12 Bar Chart of Multiple Lines by Churn Status



Interpretation of Figure 12:

The bar chart shows the feature of having multiple lines to different customers, the chart indicates that customers with no multiple lines churn more. Especially, many customers that don't have multiple lines stay to use the service, because it is shown by the high proportion of "No" in the red bar. Customers with multiple lines which are almost less than half of the customers as seen by the count, have a relatively low churn rate. This indicates that customers who have subscribed to the various lines are more likely to continue using the service, this could be through brand loyalty by having higher satisfaction levels or through reliance in the services offered. It is seen that the churn ratio for the "No phone service" is the smallest however, the customer count in this area is little and the consumers may not be very active.

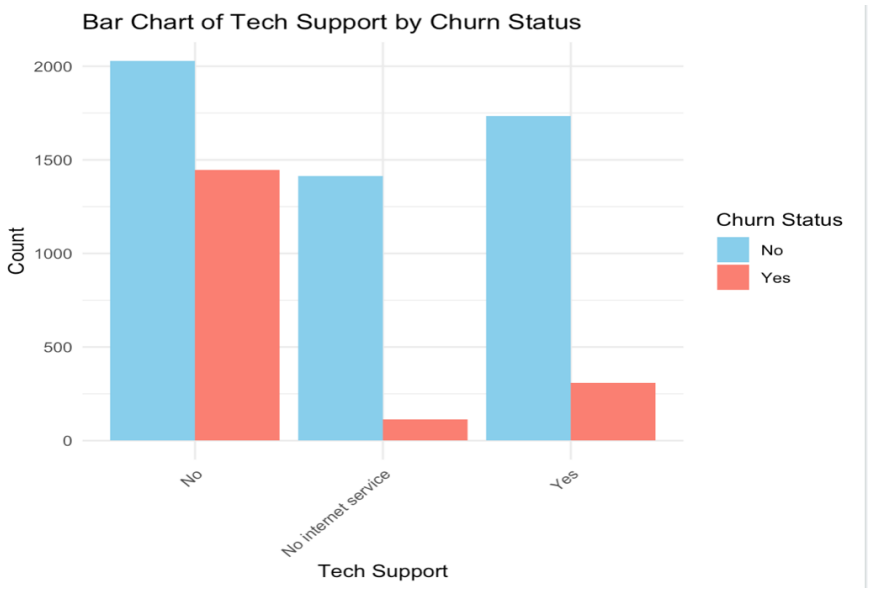
Figure 13 Bar Chart of Payment Method by Churn Status



Interpretation of Figure 13:

In the bar chart that compares the payment methods with the churn status, there can be clear extremes in the levels of churn. First, customers that use electronic checks show they have the highest churn levels, this means that customers who pay by e-check might be dissatisfied with regard to variables such as transaction costs or services availability. On the other hand, customers who pay through techniques like Bank Transfers and Credit/Debit Cards show a lower churn level, which means they can be less likely to leave because of the automated payments that promote a sense of commitment, and this can also be seen with mailed checks. These problems with payments should enable service providers to make specific retention programs, for example, they should give customers extra rewards to change from an e-check to a different type of payment option, or they should look into the problems that customers may encounter in their payment method of e check by using surveys or proactive customer service calls for inquiring about these sorts of problems.

Figure 14 Bar Chart of Tech Support by Churn Status



Interpretation of Figure 14:

By observing the bar chart, it is possible to state that the customers without tech support h leave the company more. On the other hand, the churn rate is much lower among the customers who have taken tech support, this underlines the fact that providing tech support services makes customers happier and less likely to switch. The "No internet service" group has the lowest churn rates for a reason: it probably has a less loyal, less active group of customers. And regarding the incorporation of the tech support variable, it has to be pointed out that its influence on customers' churn is very possible, which means that it is important to include this variable in the exploratory data analysis following later. While most customers think that paid technical support is expensive, there are a few factors that could qualify these customers as more valuable, because if they use tech support, then that is considered as an additional service. As a result, the question of tech support plays an important role both in churn prediction and in the determination of high-value customers' sets.

Combined Feature Importance Using Random Forest and XGBoost:

In order to apply efficient machine learning algorithms for a robust analysis, it is essential to know which attributes to work with, to do that, a feature importance analysis must be applied to the data, in order to show the degree of each variable in influencing the target outcomes (Xue et al, 2016), for instance churn and total charges, total charges variable was used in training of both models (Random Forest and XGboost) used in the parameters because according to the results of the descriptive analysis, the total charges was the best in defining the value of the customers. This analysis helps in directing attention to the factors that have the greatest influence on churn and high value customers, the insights from this analysis will help in segmenting high value customers, and in set the model construction phase, by adjusting parameters accordingly.The following steps outline the process of performing a combined feature importance analysis, which uses both Random Forest and XGBoost models to evaluate feature significance from two perspectives: The two variables used are prediction of customer

churn and total charge. By creating these models, there can be an understanding of the different features for justifying the variables examined in the descriptive analysis:

To conduct a prediction for churn and estimating total charges to segment high value customers, the models that really stand out for their ability to identify important variables for predicting churn are Random Forest and XGBoost models. To do this analysis, first step is to encode the dataset by converting it categorical attributes into numerical ones and ending with training the entire dataset based on the churn and total charges as proxy features and target variables.

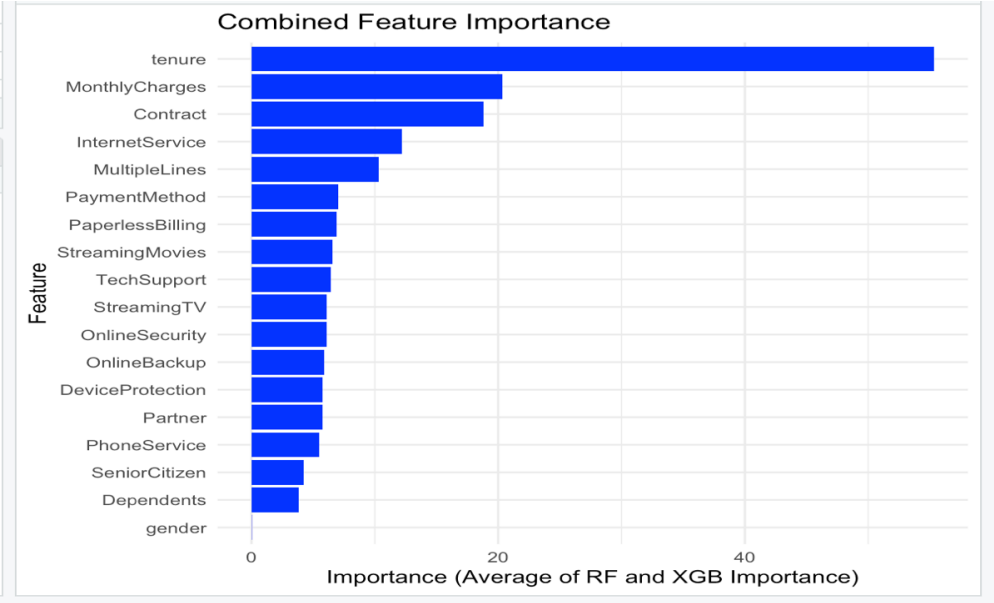
Specifically, Random Forest and XGBoost models are developed for churn where total charges are the variable that all the variables are measured according to. The feature importance scores are derived using a decrease in accuracy for Random Forest model as well as Gain for XGBoost systems. After training the XGBoost boost and random forest separately, each model was trained for each variable (churn, total charges) in this case, to understand the features that important for the analysis based on the target variables, the scores are added up and combined to provide the importance of features shown in the figure below. The ranking of variables based on their importance score is shown in the visualization of bar diagrams. This is done to ensure all key features influencing the churn and the total charges are captured, and to filter out the features which are not very important in order to reduce computational power and dimensionality by focusing on those that will have significant impact only.

The usage of XGBoost and Random Forest Together:

Random Forest was first selected because it is accurate, while XGBoost is selected because it has the possibility to increase the accuracy of Random Forest. Random Forest is better than any other decision trees as it generates more numbers of decision trees in the forest and comes with high accuracy, and less likely to be over-fitting the data set as is in bigger data set like the one used (Alsagri et al, 2020). XGBoost is a gradient boosting framework, and it has the abilities of high accuracy and efficiency in terms of time for the optimization of different types of models for prediction (Yun et al, 2021). By integrating the outcomes of the two models, there is a balanced and a better measurement of the features that cause churn and are related to the total charges. Now the purpose of the combination of the two methods is to make sure of its accuracy and reliability, it can be said that it is more of a validation process that presents a better accuracy of the results, so it is a better approach to segment the customers more clearly especially the high-value ones.

Feature importance scores have been finally obtained for every built model, by using Mean Decrease Accuracy for Random Forest, and Gain for XGBoost. The importance scores for each feature were summed up by using the mean of the feature importance for all the models. The last visual presents the most important features related to churn and customer value, which will impact the data preparation and analysis stage later.

Figure 15 Combined Feature Importance



Interpretation of Figure 15:

The integrated feature importance analysis displays the feature importance level for the variables that best predict customer churn and total charges. Tenure becomes the most important variable, which says that the duration stay with the company has a great impact on the churn levels and is related to the total charges of a customer, meaning it is tied to determining the customer value. One can also observe that monthly charges are related to the total charges, and this indicates that the financial factors of customer contact are also related to churn and high value. Contract type and internet service variables also matter because they are key indicators to the value of service plans and connectivity towards the customers. This plot helps in directing the process of data preprocessing part, as these key variables that are identified will be cleaned and transformed first, then the analysis will start so that the models to be applied are efficient and effective in solely predicting the customer churn segment of the high value customers. The leat important features like gender and dependents may get rejected and excluded from the analysis because they don't contribute a lot to the prediction.

So, the final decision for the variables to be used for this analysis are the top 10 features in the plot, which are:

1. Total Charges
2. tenure
3. Monthly Charges
4. Contract
5. Internet Service
6. Multiple Lines
7. Payment Method
8. Paperless Billing
9. Streaming Movies
10. Tech Support
11. Streaming TV

These features were selected because they all demonstrated the importance in the two models, Random Forest and XGBoost. In order to validate the selection of these variables, some steps need to be taken care of:

Combined Feature Importance Analysis:

Both Random Forest and XGBoost classifiers were learned on the entire dataset. Feature importance scores were extracted and aggregated along with the relative information gain and a brief descriptive analysis was provided of the top 10 features that have a big influence on the churn prediction and customer value. Here is a table comparing the accuracy measures of Logistic Regression, Random Forest, and XGBoost models for both the entire dataset using all features and the top 10 features:

Table 1. Comparing accuracy in prioritising the top 10 features and all the features

Model	Dataset	Accuracy	Precision	Recall	F1 Score
Logistic Regression	All Features	80.4%	84.4%	89.8%	87.07%
Random Forest	All Features	97.9%	97.7%	99.4%	98.5%
XGBoost	All Features	92.4%	93.3%	96.6%	94.9%
Logistic Regression	Top 10 Features	80.4%	84.4%	89.8%	87.07%
Random Forest	Top 10 Features	90.1%	91.5%	95.3%	93.4%
XGBoost	Top 10 Features	89.9%	91.2%	95.4%	93.3%

Model Performance Comparison:

The logistic regression, random forest, and XGBoost models were first trained on all features of the original dataset. Then the models were trained again but using only the top 10 features from the feature importance. The performance of these models in both settings is compared and checked using the measures of accuracy, precision, recall and F1 score. The first analysis testing using all the variables showed a great accuracy. However, The second trial using the top 10 feature had a decreased accuracy, although it's not a big decrease because the average of the decrease was only the deduction of 4 to 5 points, this trade-off is allowed since the model is easier to use and interpret while maintaining a minimum accuracy of the results, this trade off is also allowed because it prevents overfitting and problems with computational power, which is actually advantageous for the analysis

High Value Customer Segmentation:

Introduction:

High-value customers are what makes telecom company sustain their financial success and revenue generation, so it is very important for companies to pinpoint the segment of these customers to be able to personalize advertisements for them and increase their important loyalty. In this part of the analysis, the segmentation of high value customers will take place on the data set telco customer churn, and the method to be used to classify high-value customers is an unsupervised learning algorithm called K-means clustering.

Defining High-Value Customers

Kmeans clustering applies the process of clustering through using a specific algorithm, this algorithm groups data objects compactly around a cluster of data point, and the separation between two clusters is very clear (Huang et al, 2014). The identification of high-value customers is based on the total charges variable, which helps in measuring the worth or contribution of the customer, a customer is considered valuable if his or her total charges are higher than the 25% of the total charges. This is a threshold that was developed with regards to the TotalCharges distribution.

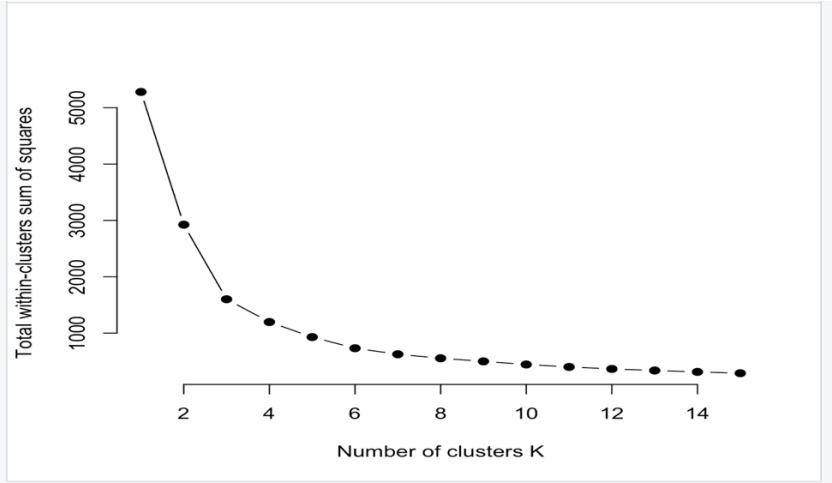
Data Preprocessing

For the clustering analysis, the high value customer subset was considered, and the relevant features were only selected, and these variables are tenure, MonthlyCharges, and TotalCharges. These features were selected because there is a direct correlation between the features and the customer value and customer retention. These were done to standardize the selected features in a way that they all contributed to the clustering process in the same measure. Standardization was done on the features by normalizing the data to have unit variance where the value of each feature is subtracted by the mean and then divided by the standard deviation of the feature.

Applying K-Means Clustering

And to identify the number of clusters, the best way was the Elbow method when WCSS is plotted versus the number of clusters. The number of clusters can be easier determined according to the least slope of the curve, that is the point after which WCSS starts increasing slowly, and this is called the “elbow method”.

Figure 16. The Elbow Method

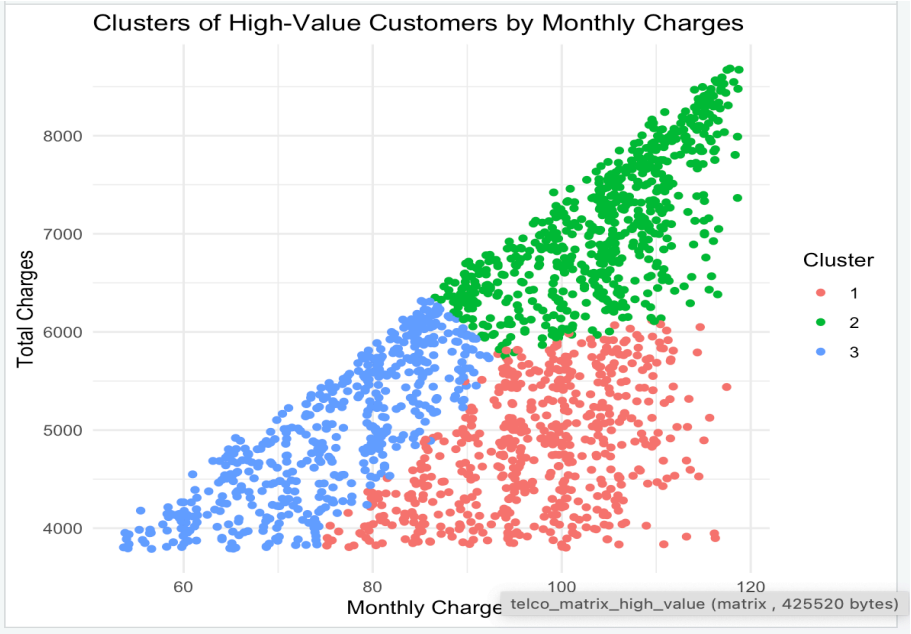


Interpretation of Figure 16:

By plotting the Total Within-Cluster Sum of Squares (WCSS) against the number of clusters (K) on the Elbow method plot, we can better decide how much is the K-value that is best for clustering. The right number of clusters is determined in the elbow point in the graph, and in this case, it is the number 3 in the graph, which means that when added more clusters it will not really be efficient for reducing the WCSS. Three clusters is enough for the best balance between quality and simplicity.

After determining the number of clusters, the K-means clustering method is then applied on the dataset. To be specific, the three clustering methods were applied to distribute each of the high-value customers into one of the three clusters depending on the similarities of their features. These cluster assignments were added to the data to the previous and subsequent analysis steps as a variable.

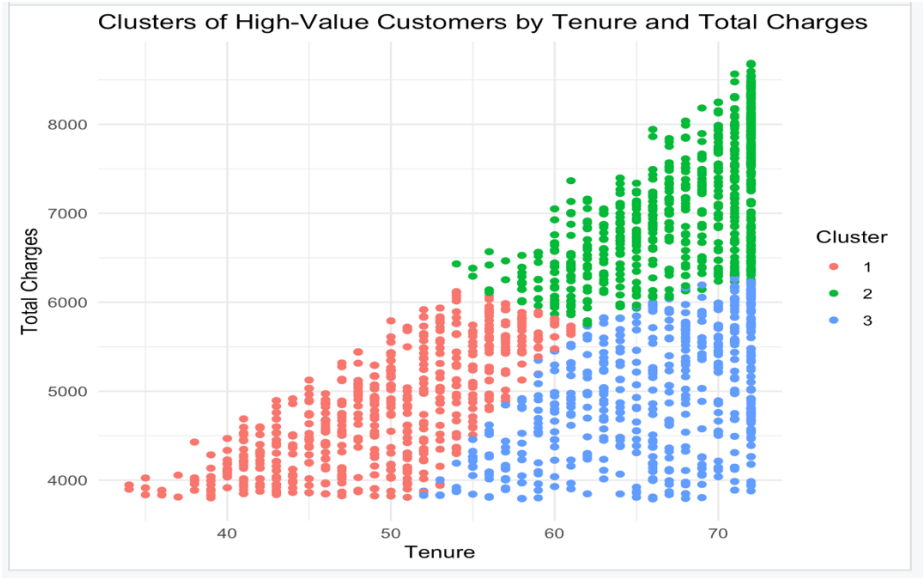
Figure 17 Clusters of High-value Customers by Monthly Charge



Monthly vs. Total Charges

The scatter plot of Monthly Charges vs Total Charges, show different clusters with different vibrant colors, Cluster 1(red) has customers who have high Monthly Charges but low Total Charges therefore these customers less profitable than other clusters in terms of both parameters. Cluster 2 (green) represents the most important customers, that will be focused on and subsetted later for analysis, these are those with high value of monthly charges and total charges at the same time, which their high profitability. The third cluster (blue) has fewer and moderate monthly charges and includes a moderate total charge too, which says that there is different spending behavior. The spatial density of green and blue dots in the upper right quadrant suggests that for each cell, a higher monthly charge makes a significant amount to the total charge; for telecom companies, it is the most beneficial to specifically target this high-value segment and focus on the retention for premium services that are personalized for this segment.

Figure 18. Clusters of High-value Customers by tenure and Total charges



Interpretation of Clusters of High-Value Customers

Tenure vs. Total Charges

The distribution of the Tenure against Total Charges by cluster shows differences between customer types. With the largest total charge figure of about $4,500, Cluster 1 (red) is containing customers with humble short contracts and a total charge approximately around $4,400 on the average. Cluster 2 (green) contains customers with higher ten years of tenure and definitely a higher total charge and can be more than $6000. The third cluster group number 3 (blue) contains High tenure and has a relatively big range of total charges from 4000 to 6000 dollars. From this it is understood that when the tenure increases the total charges increase, even more for Clusters 2 and 3. All these information helps the telecom companies to identify the groups of customers that are valuable in the long run and may serve as loyal customers.

Insights for the Subsequent Analysis:

These clusters are great for understanding customer characteristics and spending contribution for segmenting them and targeting them. From the clustering analysis, it has been

conducted that customer who have long contracts, long tenure, and those who pay the largest monthly bill (Cluster 2) should be at the center of the retention efforts.

First, the customer group with the longest length of stay and the highest monthly fees is Cluster 2 (green), this group is the worthiest segment for companies to focus on. These customers should be given attention in terms of retention strategies that include loyalty bonuses, special services and other features that can help in improving their satisfaction level. Some strategies that help in retention are satisfying the personal preferences of customers, and this could be achieved by marketing creatives that would provide special offers based on previous communication over the preferences of particular high value clients, to let them know about the new services at the very beginning.

It is important to have knowledge about the characteristics of Cluster 2 to apply decent targeted retention and premium services strategies. So, classification and differentiation of customers using tenure, monthly, and total charges will help in developing strategies to increase their loyalty and decrease churn in telecommunications organizations.

Based on these findings, the subsequent analysis should be conducted to examine the application of the machine learning models on Cluster 2 data points. This will help in figuring out the churners and apply suitable retention strategies for such segment. More specifically, the following goals of the analysis are: The creation of Cluster 2 customer segment to uncover deeper insights into its members' usage pattern, their preferences, to help in designing effective marketing strategies and retention tools. There is a need of conducting A/B tests such as different retention and upselling techniques for the customers in Cluster 2, in order to make sure that churn decreases through retention campaigns. Also, the focus on revenue generation improvement and looking at new services, and products that will attract and re high value customers is also a good strategy to be adopted to improve revenue generation.

Here is the structure of the dataset's variables and variables

tenure	MonthlyCharges	Contract	InternetService	MultipleLines	PaymentMethod	PaperlessBilling
58	100.35	1	2	2	3	0
49	103.70	0	2	2	2	1
69	113.25	2	2	2	3	0
71	106.70	2	2	2	2	0
47	99.35	0	2	2	0	1

StreamingMovies	TechSupport	StreamingTV	TotalCharges	Churn
2	1	2	5681.10	0

2	1	2	5036.30	1
2	2	2	7895.15	0
2	1	2	7382.25	0
2	1	2	4749.15	1

Here's a brief paragraph explaining the numerical values in the table:

The numbers in the table are different because they are simply encoded for clustering analysis and later model application, For example, "Contract" has values from 0 for month to month, 1 for one year, and 2 for two-year contracts.

"Internet Service" has 1: DSL, 2: fiber optic and 0: no internet service at all.

"Multiple Lines," "Streaming Movies," "Tech Support", and "Streaming TV" are categorical variables with always being 1 for 'yes' and 0 for 'no'. "Payment Method" has different numerical values for different methods: 0 for electronic check, 1 for mailed check, 2 for bank transfer, and 3 for credit card.

"Paperless Billing" as 1 if the answer is yes and 0 if it is no. "Churn" in which clients who have not churned are 0 and those who have churned are 1.

Chapter 4 Data preprocessing:

Introduction:

Data preprocessing is an important stage for this analysis. It is the process of conducting thorough analysis on the raw data and converting it into a format that is legible for analysis. The steps used in pre-processing are basically the handling of missing values, detecting and processing outliers, categorizing variables, omitting unimportant and redundant Variables, forming high-value and cluster variables and also normalizing/standardizing the data. These steps are important to enhance the quality of the input data and are critical in increasing the efficiency of the learning algorithms.

Handling Missing Values:

This process is very important to prevent data loss and to have accurate results in the model applications, and to have a smooth exploration of the variables and relationships between them. The method followed in this study was simply to fill the missing values using the mean method containing such missing data. First, to ensure no missing values were present in the dataset, R code that identifies how many missing values are in data was ran, and the result was

11 missing values specifically among the categorical variables that had yes and no values, knowing that the amount of missing values was small according to the massive size of the data, it was decided that the mean method was the most appropriate for dealing with such a small amount.

```
> # Check for missing values in the dataset
> missing_values <- colSums(is.na(telco))
> print(missing_values)
```

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
0	0	0	0	0	0	

MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV
5	0	0	0	0	6	0

StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	0	0	0	0	0	

After Dealing with Missing values:

```
> # Counting the total number of NA values in the dataset
> total_missing_values <- sum(is.na(telco_processed))
> print(total_missing_values)
[1] 0
```

Handling Outliers:

Outliers are a threat to the performance of any machine learning models since it can lean the distribution of the data points and therefore make the training process fail. The next step taken to check the data was to consider the Outliers on the key financial variables which are tenure, MonthlyCharges, and TotalCharges, because they are on the only variables that have different varying numbers that are not categorical or binary.

To visualized the outliers, Box plots must be plotted to see any outlier points straying from the other data points in extreme high or low points. No outliers were detected in these variables, as indicated by the absence of red dots in the plots. The IQR was calculated in order to validate if there are any outliers, The IQR method is basically a method that detects outliers as the data points that exist in an area greater than 1.5 times the IQR, and by the first and third quartiles as well. The results confirmed that there are no outliers in the tenure, MonthlyCharges, and TotalCharges variables:

```
# Identify outliers in 'tenure'
> tenure_outliers <- identify_outliers(telco_processed$tenure)
```

```
> print(tenure_outliers)
integer(0)
>
> # Identify outliers in 'MonthlyCharges'
> monthly_charges_outliers <- identify_outliers(telco_processed$MonthlyCharges)
> print(monthly_charges_outliers)
numeric(0)
>
> # Identify outliers in 'TotalCharges'
> total_charges_outliers <- identify_outliers(telco_processed$TotalCharges)
> print(total_charges_outliers)
numeric(0)
```

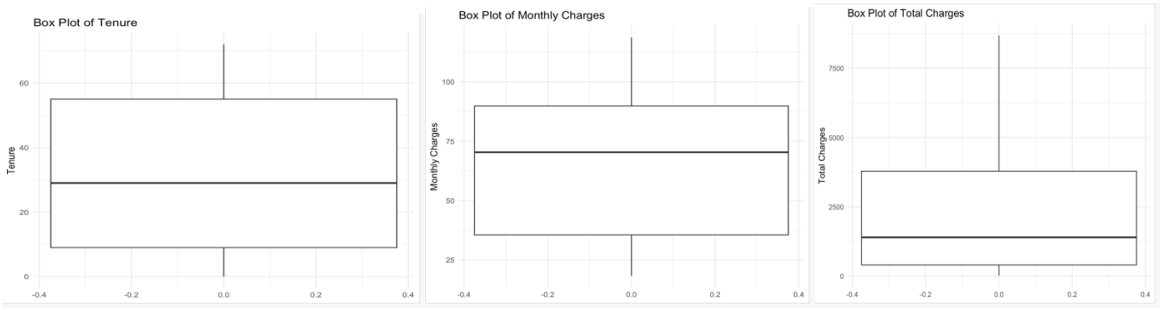
Tenure Outliers: None

Monthly Charges Outliers: None

Total Charges Outliers: None

If the number is zero, then this will help us have a balanced and profound dataset for further training and correct model building.

Figures 19 and 20 and 21 Showing Zero Outliers.



Encoding Categorical Variables:

Categorical variables that include values in character format like words have to be encoded as a part of the data preprocessing stage. For the prediction models to train on the dataset, these categorical variables need to be switched to numerical for effective use in modeling. In this study, categorical features were dummified, meaning the categorical variables were encoded via label encoding method. This method involves the labelling of each factor by an integer value, which labels every word such as yes or no with a assigned number such as 0 or 1, that way, the order in the data points is still reserved and still has a meaning.

The Encoding Methods:

Label Encoding: every category was dedicated a specific unique integer number. The Label encoding method was used on the following variables:

`gender`, `SeniorCitizen`, `Partner`, `Dependents`, `PhoneService`, `MultipleLines`, `InternetService`, `OnlineSecurity`, `OnlineBackup`, `DeviceProtection`, `TechSupport`, `StreamingTV`, `StreamingMovies`, `Contract`, `PaperlessBilling`, `PaymentMethod`, and `Churn`.

For instance, the `gender` variable had value like "Female" and "Male", and now they were converted to 0 and 1, which will convey the same meaning and value to the machine learning models later.

And of course, the other variables were encoded accordingly:

SeniorCitizen: Left as is because it is originally 0 and 1.

Partner: Encoded from "No" "Yes" to 0 and 1.

Dependents: Encoded from "No" "Yes" to 0 and 1.

PhoneService: Encoded from "No" "Yes" to 0 and 1.

MultipleLines: Encoded from "No phone service", "No", "Yes" to 0, 1, 2.

InternetService: Encoded from "DSL", "Fiber optic", "No" to 1, 2, 0.

OnlineSecurity: Encoded from "No", "Yes", "No internet service" to 1, 2, 0.

Contract: Encoded from "Month-to-month", "One year", "Two year" to 0, 1, 2.

PaymentMethod: Encoded from "Electronic check", "Mailed check", "Bank transfer (automatic)", "Credit card (automatic)" to 0, 1, 2, 3.

Churn: Encoded from "No" "Yes" to 0 and 1.

The Table below is basically a subset that depicts the encoded data, it shows how the categorical variables are encoded into numerical variables that are good for machine learning application. This subset of the data shows only ten variables out of the original 21 variables in the telco data just to get a concise overview of the process:

tenure	MonthlyCharges	Contract	InternetService	MultipleLines	PaymentMethod	PaperlessBilling
58	100.35	1	2	2	3	0
49	103.70	0	2	2	2	1
69	113.25	2	2	2	3	0
71	106.70	2	2	2	2	0
47	99.35	0	2	2	0	1

StreamingMovies	TechSupport	StreamingTV	TotalCharges	Churn
2	1	2	5681.10	0
2	1	2	5036.30	1
2	2	2	7895.15	0
2	1	2	7382.25	0
2	1	2	4749.15	1

Variable Elimination and Creation:

In the preprocessing stage, it was very crucial to deal with variables and evaluate their importance, so after conducting th feature importance analysis, some variables were removed to make the dataset more efficient for the model and easier to interpret.

Criteria for Removing Variables:

Irrelevant to Analysis: Variables that don't add any meaningful information to the analysis or modeling.

Feature Importance: Feature Importance Analysis was carried out to measure the importance of variables according to the churn and Total Charges variables, in order to know the variables that will make a difference in predicting customer churn specifically for the high-value segment.

Variables Removed:

customerID: This variable was omitted from the data because its basically an identifier for each customer, and it doesn't contribute to the results of the model application. Overall, having it in the analysis wouldn't have helped in understanding behaviour and variables affecting churn.

The following variables were removed because they have little importance scores and no predictive abilities:

- DeviceProtection, Partner, PhoneService, Dependents, SeniorCitizen and Gender

After removing these variables, the dataset is more efficient for churn prediction and customer segmentation, phone service was removed because the majority of customers had phone service so keeping it in the analysis wouldn't have made a difference, senior citizen variable had very little values that said yes among all the 7043 rows, so that's why it was removed.

Creating Derived Variables

Derived variables are a variable created based on a group of variables in the data, to help the analysis be more insightful.

HighValue: This variable was created to identify high value customers based on the total charges feature.

- Criteria: Customers were segmented as high value if they had total charges in the top 25% of the data points. This is a threshold that has been determined by the 75th percentile of the TotalCharges variable.
- Method: HighValue was encoded for the later analysis, with 1 if the customer's total charges were above the threshold (the top 25%), and 0 if lower than that (meaning 0 is not a high value customer).

Cluster: This is a variable that was created by applying the K-means clustering method to segment high value customers into a specific subset, the cluster was engineered based on several financial variables such as: tenure, monthly charges and total charges.

- Criteria: The number of clusters was determined by the Elbow method, and after plotting the elbow method it occurred that the best number was 3, and K-means clustering applied 3 clusters to segment customers.
- Method: The Cluster variable was added to the dataset, with the cluster assignment for each customer. Cluster 1 and 3 were for the medium tenure and financial contribution to the company, and cluster 2 was the highest tenure and financial contributions, with total charges above 4000$

After creating these derived variables, there is now a more granular outlook of customer segments and levels, these different levels enable telecom companies to target high value customers and improve the analysis as they are very valuable.

During the model application part, R will be used to segment high-value customers based on the high value variable, and this will happen by writing code that's purposed to filter another data frame out of the original dataset, this high value customer data subset will used for a focused machine learning analysis involving Logistic Regression, Kernel SVM, Random Forest, and XGBoost. This subset will be compact just to include the high value ones only to enhance accuracy and customer retention.

Normalization/Scaling:

Min-Max scaling method has been used to normalize the numeric features (the encoded ones), The scaling transforms the values to be either [0 or 1]. It was picked for this analysis because it keeps the relationship between data points whole and eliminates any bias for particular attributes in the model training. Other variables are also being normalized such as: tenure, MonthlyCharges, and TotalCharges. This normalization enhances the effectiveness of machine learning algorithms by removing.

Here is the result of the scaled subset of the data with the top ten features including only the high value customers:

```
# Print the scaled dataset
> print(head(telco_top_ten_high_value_scaled))
   tenure MonthlyCharges Contract InternetService MultipleLines PaymentMethod PaperlessBilling
13 0.6315789     0.7173579        1               2             2             3                 0
```

14	0.3947368	0.7688172	0	2	2	2	1
16	0.9210526	0.9155146	2	2	2	3	0
18	0.9736842	0.8149002	2	2	2	2	0
27	0.3421053	0.7019969	0	2	2	0	1
29	1.0000000	0.5622120	2	1	2	3	1

	StreamingMovies	TechSupport	StreamingTV	TotalCharges	Churn	Cluster
13	2	1	2	0.3864491	0	1
14	2	1	2	0.2547389	1	1
16	2	2	2	0.8387021	0	2
18	2	1	2	0.7339346	0	2
27	2	1	2	0.1960842	1	1
29	2	2	2	0.5270549	0	2

Scaling of the Processed Dataset:

The purpose of scaling the processed dataset is to prepare the whole data for the 4 machine learning algorithms, Scaling is beneficial when attempting to prepare for different subsets of data that will be used in analyses because it maintains uniformity in all features.

Scale the telco_top_ten_high_value dataset:

The purpose of the concentration on high-value customers subset is to make analysis more focused, tighter and within a specific range of customer value.

Chapter 6 Application of Machine Learning Algorithms:

Introduction:

This chapter is dedicated to the application of the four ML algorithms to predict the high value customer churn segment. The chosen algorithms are logistic regression, random forest, kernel SVM, and XGBoost. Every model has different and unique abilities and ways of conducting their predictive analysis. Logistic regression creates a simple and interpretable model, random forest and XGBoost have strong performance and can handle complex interactions between features. Kernel SVM is significantly better than other models in analyzing non-linear relationships. Using these models is intended for achieving high accuracy in churn prediction and for recommending retention strategies.

The following part has details about the model training and testing, the evaluation using the accuracy metrics, and the interpretation process, finally summarized through a comprehensive model performance comparison. The goal is to pick the right model for real implementation in the telecom industry, this selected model will surely achieve customer retention and revenue gain for telecom companies.

Overview of Algorithms: To predict customer churn and improve the retention measures for customers, this chapter employs four viable ML algorithms namely, Logistic Regression, Random Forest, Kernel SVM, and XGBoost.

• Logistic Regression: It is a statistical model that creates a function to predict a binary variable. It is easy to explain and interpret and is suitable for understanding how all the features in the data can influence churn (target variable).

• Random Forest: A classification algorithm that trains on data using many decision trees, and during the training, it outputs the most frequent class out of the trees that were constructed during training. And It works really well on large datasets, and it prevents overfitting from happening.

• Kernel Support Vector Machines (SVM): It is a very effective method of classification; it works by identifying where the hyperplane is located, and it separates the classes in the high dimensional space. It works well in datasets with dimensionality and especially well in working with non-linear relationships by using the so-called kernel trick.

• XGBoost: A fast version Gradient Boosting model, it's based on the tree learning algorithms, by constructing many decision trees, but works by eliminating the inefficient and error trees, it is light, fast, and accurate, and is used a lot in machine learning research for its commendable accuracy.

Description of the subset of high value dataset:

The new data frame that is called "telco_top_ten_high_value_scaled" is derived from the original dataset of the Telco Customer Churn. The subset only includes all the high-value customers that contribute the most financially. This dataset has a total of 1761 observations across the 12 variables. This subset included the top ten most important features which are: Tenure, MonthlyCharges, Contract, Internet Service, MultipleLines, PaymentMethod, PaperlessBilling, Streaming Movies, Tech Support, StreamingTV, Total Charges, and Churn.

The two variables, tenure and MonthlyCharges, describe the usage behavior of the customer to the services, tenure is essentially the number of months the customer has been with the company and MonthlyCharges is the amount of money the customer pays monthly. Contract and InternetService variables have encoded into numerical, they both represent the type of contract the customer signs, and the type of internet service the customer enrolls in. MultipleLines, PaymentMethod, PaperlessBilling, StreamingMovies, TechSupport, StreamingTV are variables that represent the additional services that a customer is subscribes to, and it's been encoded as well for the analysis. The sum of the charges that a customer pay over their stay is the TotalCharges, and Churn is the predictor target variable that says if a customer will churn or not. The features in this subset were preprocessed properly for the predictive modeling and analyses, the encoding of categorical variables, normalization of the numerical data and elimination of the irrelevant features were completed as well.

Basic Summary of High-Value subset

In order to get a brief idea about the telco_top_ten_high_value_scaled dataset, some basic statistical summary results need to be reviewed alongside some plots that visualize the data distribution. This section will provide the summary results as well as plots of the selected variables in a subset in the form of several summary statistics:

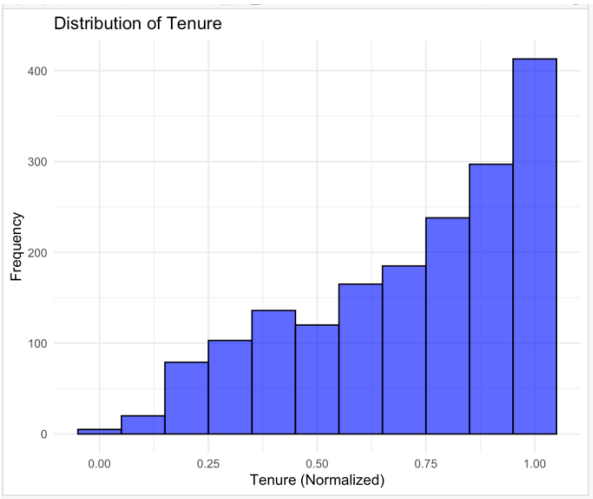
```
summary(telco_top_ten_high_value_scaled)
   tenure       MonthlyCharges    Contract       InternetService MultipleLines  PaymentMethod
Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :1.000   Min.   :0.000   Min.   :0.000
1st Qu.:0.5263   1st Qu.:0.4639   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:0.000
Median :0.7895   Median :0.6429   Median :1.000   Median :2.000   Median :2.000   Median :2.000
Mean   :0.7184   Mean   :0.6122   Mean   :1.243   Mean   :1.699   Mean   :1.728   Mean   :1.673
3rd Qu.:0.9474   3rd Qu.:0.7919   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:3.000
Max.   :1.0000   Max.   :1.0000   Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :3.000
PaperlessBilling StreamingMovies TechSupport     StreamingTV     TotalCharges     Churn
Min.   :0.0000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :0.0000   0:1506
1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:0.1786   1: 255
Median :1.0000   Median :2.000   Median :2.000   Median :2.000   Median :0.3715
Mean   :0.7047   Mean   :1.759   Mean   :1.576   Mean   :1.751   Mean   :0.3936
3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.5883
Max.   :1.0000   Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :1.0000
```

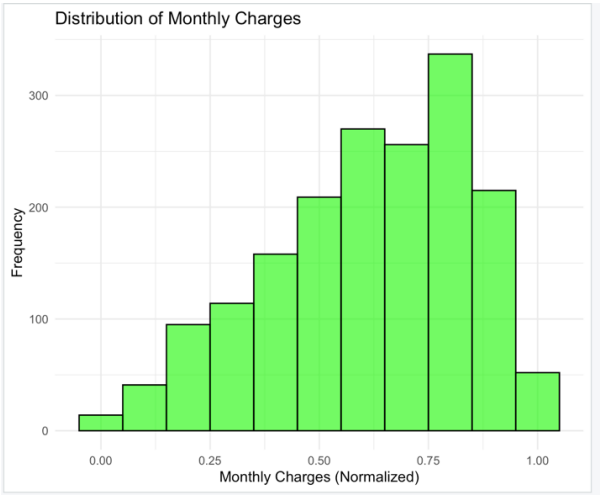
And Here's a structure of how the dataset looks like in its final form for the data analysis:

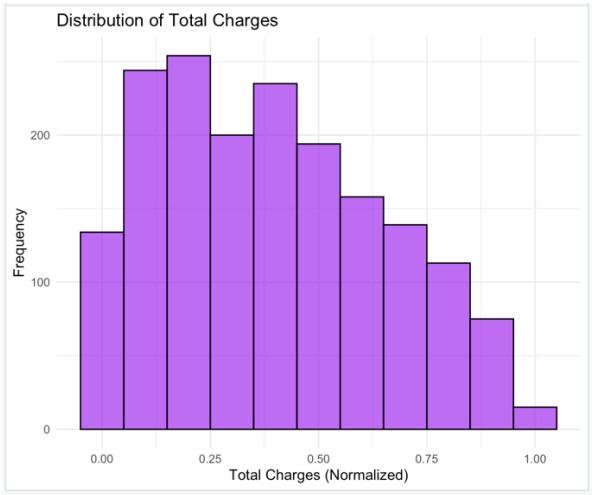
```
> str(telco_top_ten_high_value_scaled)
'data.frame':    1761 obs. of  12 variables:
 $ tenure         : num  0.632 0.395 0.921 0.974 0.342 ...
 $ MonthlyCharges  : num  0.717 0.769 0.916 0.815 0.702 ...
 $ Contract        : num  1 0 2 2 0 2 2 2 2 2 ...
 $ InternetService : num  2 2 2 2 2 1 2 2 1 1 ...
 $ MultipleLines   : num  2 2 2 2 2 2 2 2 2 2 ...
 $ PaymentMethod   : num  3 2 3 2 0 3 3 2 3 3 ...
 $ PaperlessBilling: num  0 1 0 0 1 1 1 0 1 1 ...
 $ StreamingMovies : num  2 2 2 2 2 2 1 1 1 1 ...
 $ TechSupport     : num  1 1 2 1 1 2 2 2 1 2 ...
 $ StreamingTV     : num  2 2 2 2 2 2 1 2 2 2 ...
 $ TotalCharges    : num  0.386 0.255 0.839 0.734 0.196 ...
 $ Churn           : Factor w/ 2 levels "0","1": 1 2 1 1 2 1 1 1 1 1 ...
```

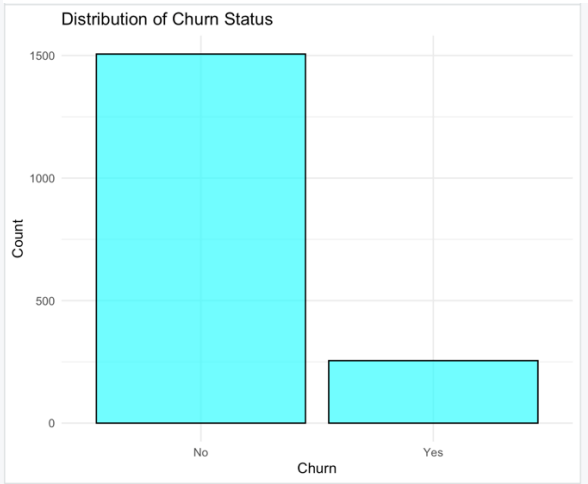
Visualizations of the distribution of the data points across the variables:

Figures 19, 20 21, 22









After visualizing the distribution of these key variables in the dataset, and after the scaling and normalization process, it is clear that these visualizations will be effective during the application of the machine learning algorithms. The first three histograms are tenure, monthly charges, total charges. Tenure variable shows positive left skewness, this shows that the majority of the high value clients stay with the service provider for a long time. The histogram of the monthly charges reveals that there is a normal distribution curve, which makes perfect analysis because the customers are divided in a right manner, where the models will not probably be biased. The total charges histogram is an ideal distribution with a very weak indication of a negative skew, which indicates that most of the customers are likely to be in a middle of the distribution rather than being in the extreme values.

Rationale for Oversampling:

The histogram of the churn status gave crucial information about the imbalance in the data set, where the total amount of the non-churners is much higher than the churners. This imbalance is important to note and take care of, because this will affect the performance of the Models accuracy. Some handling techniques of the datasets like resampling the data or selecting a fit evaluation metric would be good to get a proper actual measure of the performance of the models.

Firstly, while testing the model's performance as an occasional process for checking the accuracy of the models before moving to another part in the analysis, it was found that there was abnormal high accuracy in models like XGBoost and Random Forest with accuracy scores like (XGBoost: 100%, and Random Forest: 98.92%). And of course it is important to have high accuracy levels, but it was unexpectedly high especially when doing a customer churn prediction in a highly complicated model and highly complicated dataset. This had started a confusion about the problems seen in the data, so it was a good idea to check if there are any imbalanced classes like for example, churn has imbalanced levels as seen in Figure 22 above.

Finally, it was concluded the dataset has higher values of no churning customers than the churning customer. This imbalance says that the models trained in a way that made them favor the majority class and therefore, scored unreasonably high accuracy rates. To overcome this problem, there was a need to perform oversampling, in order to balance the classes. It is a technique in which the number of samples of minority class is artificially modified to be closer to the majority class, therefore it would give us a better chance of the data to craft a better and realistic machine learning models that would not be prone to overfitting.

Method and Process

Oversampling was accomplished by using the Random Oversampling Examples (ROSE) method. The method works by generating new samples by using smoothed bootstrap, it enables ROSE to balance the classes since it creates synthetic data points. The applied method was used for oversampling the Churn variable and achieving the balance between customers who churn and customers who do not churn. To confirm the new class distribution, oversampled data set telco_tthvsov (the letters stand for many things that represent oversampling and scaling and many preprocessing steps) had a more stable count, the clients who did not churn were close to the customers who churned, 889 non-churners and 872 churners.

To employ the ROSE technique to our data, the following parameters in the ROSE method have been used:

- formula = Churn ~ .
- data = telco_top_ten_high_value_scaled
- seed = 123

Due to the oversampling, the variables had some perturbed values, for example the financial features such as tenure, monthly charges, and total charges, had some negative values which is not appropriate for the analysis, and the categorical variables that were encoded had some problems with being in decimal format although originally, they are in integer format originally. For evidence, one car refer to the summary of how the values' nature were after oversampling.

```
summary(telco_tthvsov)
    tenure      MonthlyCharges     Contract      InternetService MultipleLines   PaymentMethod
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :1.000   Min.   :0.000   Min.   :0.0000
 1st Qu.:0.4348   1st Qu.:0.4972   1st Qu.:0.1969   1st Qu.:1.708   1st Qu.:1.610   1st Qu.:0.2747
 Median :0.6983   Median :0.6770   Median :0.9668   Median :1.940   Median :1.924   Median :1.6751
 Mean   :0.6531   Mean   :0.6453   Mean   :0.9752   Mean   :1.756   Mean   :1.699   Mean   :1.5186
 3rd Qu.:0.9031   3rd Qu.:0.8253   3rd Qu.:1.7172   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.6363
 Max.   :1.0000   Max.   :1.0000   Max.   :2.0000   Max.   :2.000   Max.   :2.000   Max.   :3.0000
 PaperlessBilling StreamingMovies   TechSupport      StreamingTV     TotalCharges      Churn
 Min.   :0.0000   Min.   :0.1861   Min.   :0.1474   Min.   :0.2917   Min.   :0.0000   0:889
 1st Qu.:0.4787   1st Qu.:1.6066   1st Qu.:0.9760   1st Qu.:1.5571   1st Qu.:0.1679   1:872
 Median :0.8664   Median :1.9132   Median :1.4213   Median :1.9156   Median :0.3612
 Mean   :0.6969   Mean   :1.7141   Mean   :1.4118   Mean   :1.6976   Mean   :0.3934
 3rd Qu.:1.0000   3rd Qu.:2.0000   3rd Qu.:1.9645   3rd Qu.:2.0000   3rd Qu.:0.5845
 Max.   :1.0000   Max.   :2.0000   Max.   :2.0000   Max.   :2.0000   Max.   :1.0000
```

Returning the encoded categories back to their original integer form, was essential for the model applications to work.

- str(trainData_tthvsov)

```
'data.frame':       1410 obs. of  12 variables:
 $ tenure         : num  0.34 1 0.903 0.911 0.943 ...
 $ MonthlyCharges  : num  0.663 0.693 0.547 0.309 0.557 ...
 $ Contract       : int  1 1 2 2 2 2 0 1 1 1 ...
 $ InternetService : int  1 1 1 1 1 1 1 2 1 2 ...
 $ MultipleLines   : int  1 1 2 1 2 0 1 0 2 2 ...
 $ PaymentMethod   : int  3 3 0 3 3 1 2 0 2 0 ...
 $ PaperlessBilling: int  1 0 1 0 0 1 1 0 0 0 ...
 $ StreamingMovies : int  1 2 1 1 2 1 2 1 1 2 ...
 $ TechSupport     : int  1 1 1 2 2 2 2 2 1 1 ...
 $ StreamingTV     : int  2 1 0 0 2 2 2 1 1 2 ...
```

```
$ TotalCharges    : num  0.182 0.851 0.871 0.465 0.435 ...
$ Churn           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

The following is the final result from the summary function, that shows the count of the zero and one levels of the churn variable, and as it can be seen here that the zero which represents no churn has a count of 712, and the 1 level which means yes churn, shows that it has a count of 698, which is approximately a ratio of 1:1 that is good for further data analysis:

```
> summary(trainData_tthvsov$Churn)
  0   1
712 698
```

The Application of Logistic Regression:

Training Process and Parameter Tuning

Logistic Regression was used as one of the main algorithms to work with customer churn prediction. The model was used to train on a subset training data called trainData_tthvsov, where Churn was set as the binary dependent variable for its nature of value and that it was the target variable. The estimates for the model were obtained using the glm function in R with binomial family since the response variable is binary, the following method was used:

Churn~tenure+MonthlyCharges+Contract+InternetService+MultipleLines+PaymentMethod+ PaperlessBilling+StreamingMovies+TechSupport+StreamingTV+TotalCharges

For the analyzing of logistic regression, no hyperparameter tuning was performed since the main concern was to interpret the coefficients and understand the association between the predictors and the response variable, there was no hyper para meter because also all of the tuning was done to the data itself where features were eliminated, and the high value customer segment was targeted.

Model Evaluation

Before interpreting the results of the study, some metrics were chosen to determine the performance of the logistic regression model alongside the other models. These metrics are

accuracy, precision, recall, F1 score and the Area under Curve, and the Receiver Operating Characteristic curve (AUC-ROC). The following steps show how the evaluation process was carried out:

- The model produced probability of the Churn outcome as put on the test dataframe testData_tthvsov.

- These probabilities were then categorized into binary by class predictions with a threshold of 0. 5.

Evaluation Metrics:

Accuracy: The percentage of how much data from the total amount of data that was classified correctly.

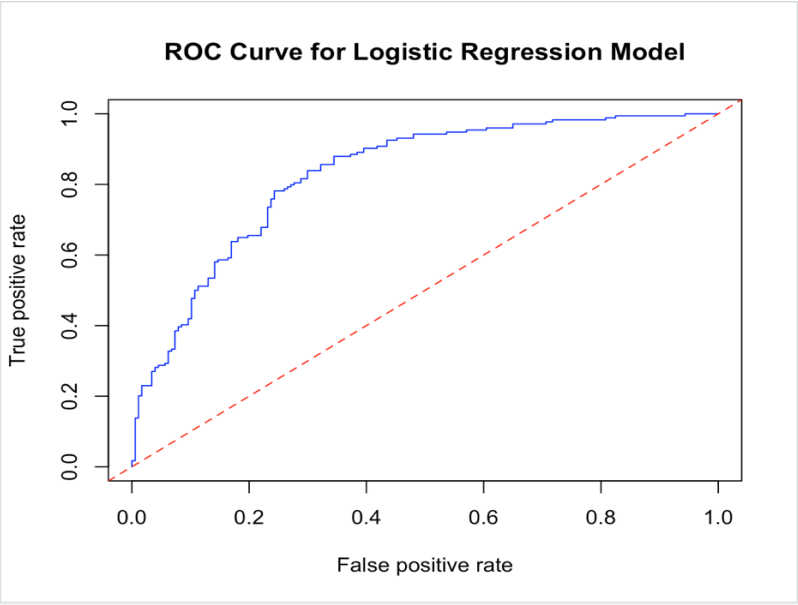
Precision: Out of all the cases classified as positive by the model, the precision tool shows what percentage are actually true positive.

Recall: A measure of the true positive predictions to the total number of cases that are actually positive.

F1 Score: A tool that considers precision and recall, that combines both the tools.

AUC-ROC: Receiver Operating Characteristic is defined as the probability of the right classifications made per unit. And the Area under the curve measures the model's ability to distinguish between positive and negative classes.

Figure 22. ROC curve for logistic regression Model



Results and Interpretation:

Accuracy: In the end, the logistic regression model yielded an accuracy of 76%.

Precision: Regarding the accuracy level in detail, the precision score was 77.9%, which means that if the model was predicting that a customer would churn, it was right about 78% percent of the time.

Recall: In general, the result of the recall score was 75.7%, which means that according to the given model resources, the model was able to correctly predict 75.7% of all the churn instances.

F1 Score: F-measure = 2TP / (2TP + FP + TN) = 1.

This values of Precision = 1Recall = 1The F1 score was 0. 768 in order to maintain the right amount of precision/recall ratio.

AUC-ROC: The AUC value was of 82.8%, which indicates that the model was fair in separating the churning and non-churning customers.

Confusion Matrix:

Confusion matrix was use that as part of the evaluation metrics, to help identify the relationship between the classes and the predicted classes of the test set. This confusion matrix shows the results into true positives, true negatives, false positives, and the false negatives.

		Reference	
		0	1
Prediction	**0**	134	38
	1	43	136

Looking at the confusion matrix, it shows that there are 134 customers that the model classified as non-churn, while 136 that were classified as churn. However, it classified 38 churn cases wrongly as non-churn and 43 non-churn cases wrongly as churn.

Model Coefficients

```
> print(coefficients)
                  Estimate Std. Error    z value     Pr(>|z|)
(Intercept)      -4.91761611 0.77781743 -6.3223270 2.576534e-10
tenure           -0.27475624 0.30986343 -0.8867011 3.752399e-01
MonthlyCharges    0.27859527 0.44166561  0.6307832 5.281823e-01
Contract         -0.79157287 0.10291730 -7.6913491 1.455916e-14
InternetService   1.76053476 0.28339281  6.2123480 5.219868e-10
MultipleLines     0.66751491 0.16105699  4.1445883 3.404247e-05
PaymentMethod    -0.09880177 0.05663922 -1.7444056 8.108839e-02
PaperlessBilling  0.08075482 0.17709581  0.4559951 6.483935e-01
StreamingMovies   0.41243960 0.17447491  2.3638905 1.808415e-02
TechSupport      -0.07058538 0.13969824 -0.5052704 6.133690e-01
StreamingTV       0.77223380 0.17079182  4.5214917 6.140538e-06
TotalCharges     -1.17496722 0.35095524 -3.3479118 8.142291e-04
```

Interpretation:

The intercept (-4. 918) shows that the baseline log-odds of churn when all predictors are at 0. The coefficient for tenure is (-0. 275) which means it cannot be accepted, meaning that tenure does not impact churn significantly. Similarly, the monthly charges which are approximately 0.279 do not appear to have any influence on the churn probability. On the other hand, if a customer signs a contract (-0. 792) their churn probability is greatly reduced and the longer the contract the lower the churn rate. And here the Internet service type (1.761) shows that customers with some specific types of internet are likely to churn. Likelihood of churn according to multiple lines is (0.668) which is very correlated and accelerates even the customer churn

rates. The coefficient for payment (-0. 099), paperless billing (0. 081) and tech support (-0. 071) seem to be less significant, they have little effect on churn. Streaming movies (0.412) and streaming TV (0.772) show that customers Streaming Movies and tv are likely to churn meaning that these services can lead to churn. At last, the total charges (-1.175) decrease the chances of churn which means the more money the customers spend, the less likely they will churn.

In conclusion, the logistic regression model shows good prediction accuracy of customer churn, involving balanced precision-recall measures and a fairly high AUC-ROC, that showed the ability to provide a solid differentiation between churn and no-churn cases.

Random Forest:

The Random Forest model was used to make another prediction of customer churn, essentially based on the group method of learning. During the training, several decision trees are built and at the time of prediction, the mode of the classes (in case of classification) or mean of the predictions (in case of regression) is given as an end result of the prediction.

As for the method of training, there was a big group of 500 trees (ntree = 500). And the Maxnodes were 10 and nodesize was set to 10, which means that no terminal nodes with more than 10 nodes were allowed, and no node was allowed to split if it contained less than 10 cases. These above parameters were modified to essentially make the model less complex for more interpretation, and to also avoid that it gets overfitted.

Evaluation Metrics: The performance is as follows:

Accuracy: 0. 7578

An accuracy of 75.78% means that the model provides a good accuracy in differentiating churn status for approximately 76% of the customers based on the complexity of the dataset for churn prediction model.

Precision: 0. 7805

78.05% refers to how often the model accurately predicted churn, which means that when the model predicts customer churn it is right in 78.05% of the time.

Recall: 0. 7232

This may be crucial in order to get as many actual churners as possible in order to act before these clients leave.

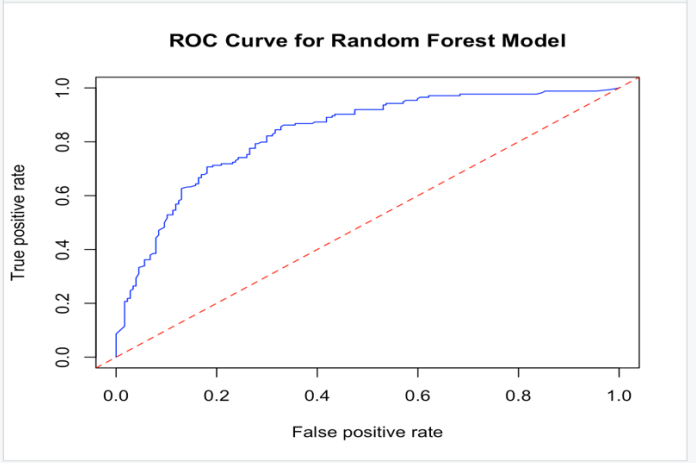
F1 Score: 0. 7507

F1 score got this score of 75%, that means it balances precision and recall as numerator and denominator, with false positive and false negatives contributing to the overall evaluation.

AUC: 0. 8305

The Retrieved AUC 83.05% shows that the model has a good level of discrimination between the churn and non churn

Figure 23



Interpretation:

The smoother the curve sticks to the left side of the ROC plane, the better the performance of the test. In this case the ROC curve proves that the model is highly sensitive with low probability of having false positives which is a proof that the model colonizes between the churning customers and those that do not churn effectively.

	Reference 0	**Reference 1**
Prediction 0	128	36
Prediction 1	49	138

The confusion matrix provides a detailed breakdown of the model's performance:

True Positives (TP): 28 customers were predicted not to churn in the train phase and among the test customers, 128 customers were correctly predicted not to churn.

True Negatives (TN): Ultimately, the number of accurately classified customers with probability of churn is 138.

False Positives (FP): Out of the total customers, 49 of them were classified as churner customers even when they are not.

False Negatives (FN): It is also important to understand that 36 customers were misclassified as non-churners.

SVM Model:

The Support Vector Machine model was built by using th edefault settings and with radials kernel, which is a type of kernel in the SVM machine learning model, In the training, the training data and the test data sets were developed with a split of 80/20 for the training and testing sets. For the SVM model, the default values for the cost parameter have been set at 1 while, anad for the gamma parameter, it was set at "scale. "

Evaluation Metrics:

Accuracy: The accuracy of the SVM model is 77.8% and this means that it correctly classified seventy-seven point eight percent of the instances which points to a highly accurate and dependable customer churn prediction model.

Precision: In terms of precision, the model has got an ability of about 78.06% to identify churners without having too many false positives.

Recall: By capturing almost all actual churners, recall is quite high at 77.95%. This is important in situations where one would want as many churners as possible.

F1 Score: The trade-off between precision and recall results in an F1 score of 78.00%, which gives a holistic measure for performance of the model.

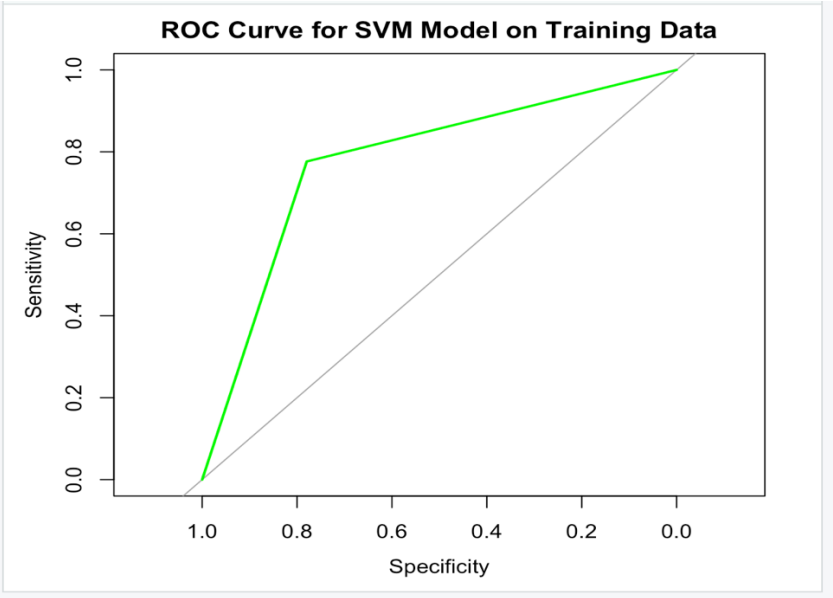
AUC-ROC: A value for AUC-ROC at 0.778 indicates that this is a robust model that can easily discern between non-churners from churners with fairly good accuracy.

Reference Prediction 0 1

0 0 555 156

1 1 157 542

- **True Positives (TP)**: 555
- **False Positives (FP)**: 157
- **True Negatives (TN)**: 542
- **False Negatives (FN)**: 156



ROC Curve:

The ROC curve shows in a graph how good the model is at sorting different outcomes. If the curve is near the top-left, it means the model can guess right who will leave or stay, without making many mistakes.

Conclusion:

The basic SVM tool with a radial setup works well to guess if customers will leave. It's good at being right a lot and balancing its guesses. Its ability to stay high on the ROC chart shows it can clearly tell who might leave from those who will stay.

XGBoost Model

Training Process and Parameter Tuning:

The XGBoost model used default parameters with a logistic regression goal and evaluated the use of log-loss. The learning process that the model goes through is heavily concerned with converting the data into a matrix format. The range of the rounds for conducting a viable prediction is set to 500 to make sure of a thorough learning of the whole data training set.

- Accuracy: 0.6952

Accuracy: The XGBoost model achieved an accuracy of 69.52%, meaning it correctly classified 69.52% of the instances. This value is slightly below that achieved by other models but it is still good in terms of predictive power.

- Precision: 0.6823

Precision: With an accuracy rate of 68.23 , the model shows a very high degree of reliability in identifying churners while not making too much use of extraneous factors.

- Recall (Sensitivity): 0.7401

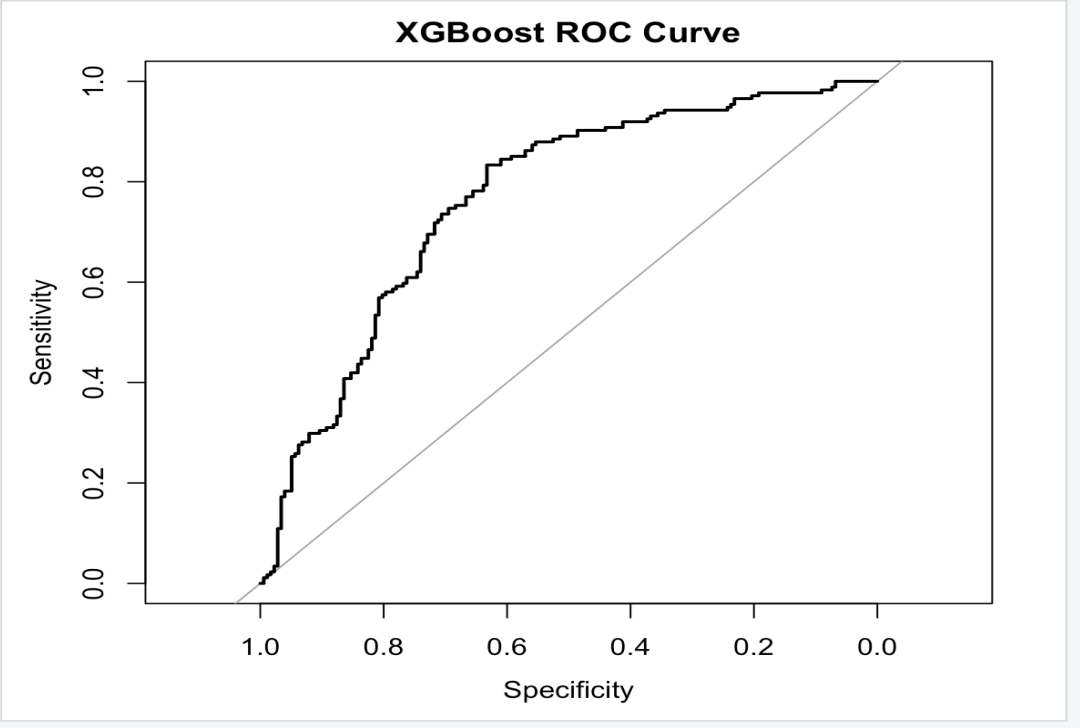
74.01% Recall means the Model has predicted Yes, 74.01% of Actual Churners correctly.

- F1 Score: 0.7100

F1 Score: The F1 score of 71.00% balances precision and recall, by giving a total combined measure of both.

- AUC-ROC: 0.765

AUC-ROC: An AUC-ROC of 76.5% says that this is a robust model that is capable of distinguishing between churners and non-churners effectively.



The ROC curve shows how well the model works at different settings. If the curve is close to the top-left corner, it means the model correctly identifies more true positives and fewer false positives, showing it's good at classifying.

The XGBoost model does a good job at predicting when customers might leave. Even though it's not the most accurate model, it's good at both correctly identifying positive cases (precision) and not missing many positive cases (recall), which makes it dependable for spotting

customers who might leave. The high AUC-ROC score also shows that the model is good at telling apart customers who will leave from those who won't.

Chapter 7: The Findings:

The findings from this research are very significant, and will further direct the developing of tailored retention strategies for high-value customers in the telecommunications industry. The correct retention strategies will be chosen and designed by using the right set of machine learning algorithms such as logistic regression, random forest, support vector machines(SVM), and according to the data analysis machine learning model applications, it is evident that XGBoost is the most balanced for telecommunication enterprises to help solve the problem of churn. Here are some main points.

The research question of this thesis, essentially revolves around this, How does telecommunications use machines learning to predict key figures that show whether important customers want to leave? And how should they put in place targeted retention measures that reduce churn levels and maximize revenues?'. The study objectives were to compare the performance of various machine learning models in predicting customer churn and to identify the most effective model for practical implementation.

Best Performing Model

Among all the evaluated models, the Random Forest model was the best performing model, because of the remarkable balance of high accuracy, precision, recall, F1 score, and AUC. The Random Forest model achieved an accuracy of 75.78% and an AUC of 0.831, showing a strong ability to distinguish between customers who are likely to churn and those who are not. The high f-score of 0.7507 also shows how well this model balances between percison and recall.

The feature importance analysis from the Random Forest model was very significant to the thesis, as it helped to reveal the importance and significance of the variables such as contract type, internet service quality, and multiple lines, had significant contributions to the prediction of customer churn. That insight helps telecommunications companies to make better-informed choices when deciding whether or not to retain their subscribers by advertising certain things or not, because as it was clear in the thesis, these variables had importance in the analysis, but in the model application, it occurred from the logistic regression application' coefficient table, that these variables certainly made customer churn, and that could be because of the lack of satisfaction with the services of internet or the streaming services made by the telecom company.

And rather the important variables that really had an affect on churn where the financial variables which are tenure, monthly charges, total charges, and it was evident that these variables positively affect the churn in a way where customers to stay long like to not churn.

Discussion of the Retention Strategies

Based on the findings of the thesis, Telecom companies should develop retention strategies that are suitable for the specific high value customer segments, telecoms can make the retention strategies by ultimately using the Random Forest model, which was determined as the most effective machine learning model to predict customer churn. Here are several strategies about the insights of the study:

The feature importance showed that customers churn is affected notably by factors like contract type, internet service quality, and having multiple lines. It helps them derive data based on which telecom companies can come up with retention offers of their preference. For instance, those client base who are seen to be at the risk of churning due to month to month tariffs could be offered affordable long term deals or a customized package of services that would make them reconsider to continue to be clients of the firm. The dissatisfied customers who valued the internet service to be of high quality might be provided with additional support services or provided with the advanced package services.

Proactive Customer Engagement

Using the Random Forest model that has a great ability to predict the churn, telecom companies can approach the high risk customer beforehand of his/her decision to churn. Being in touch with the customers on a routine basis, communicating with them using their names, and presenting probable problems with their probable solutions also goes a long way in making a customer's experience a positive one. For instance, if the model recognises churn occurring because of multiple complaints of bad services, the company can prevent this by assigning a special team to handle the complaints in a professional and fast manner.

Early Intervention Programs

The model can point out the potential customers that are on the verge of leaving the company's customer lifecycle. In the same way, education programs for young children may be developed to attend to particular related requirements throughout the various phases. In the case of new customers, orientation programmes are also effective especially those that aim at ensuring that the customers are making the best out of the services they have subscribed to. For regular clients, it is possible to use bonus programs that stimulate the constant use of the service. Also, the dissatisfied customers who have lately been downgrading their services can also be offered retention offers to satisfy them.

Enhancing Service Quality

The study highlighted that internet service quality is a crucial factor in customer retention. It's also possible for telecommunications providers to gain insights into their networks by analyzing these data - they may be able to identify bottlenecks or other causes of disturbances more easily. Regular care, updating networks, and clear information on services improve customer loyalty in many ways.

Customized Communication Strategies

The different clients like the conversation differently. But by knowing which factors are decisive for churning, telecommunications providers have an opportunity to influence them. This could mean that technically skilled clients would rather use modern communication media such as e-mails or smartphone applications for inquiries about products or services, but others are also happy with personal telephone conversations or direct contact. The use of multichannel communications makes sure that all customers' demands and wishes for contact with us are catered to.

Loyalty and Reward Programs

A means for retaining faithful clients consists in making use of coupons, bonuses, sales at reduced rates. The predictive model can identify loyal customers who are at risk of churning and target them with exclusive offers, discounts, and rewards for their continued patronage. The usual ones include points bonus on buying something via the App, specific discounts like free trial periods for extra services, or individually negotiated prices.

The application of the Random Forest model in predicting customer churn has provided valuable insights for developing targeted retention strategies in the telecommunications industry. The telecommunications industry can achieve significant reductions in customer turnover by using personalization with regard to targeted offerings, pro-active approach, timely interventions, enhanced services, individualized communications and loyalty programmes. The tactic does not only make customers loyal, but also makes them useful in the longer term. That's why this investigation is so important for businesses that want to compete in today's communications market.

Chapter 8: Conclusion

Recap of Key Findings

The study revealed that the Random Forest model outperformed other machine learning algorithms in predicting high-value customer churn within the telecommunications industry. The model reached an exactness of seventy-five point seven eight percent, a precision of seventy-eight point zero five percent, a recall of seventy-two point three two percent, and an f-measure of seventy-five point one zero seven. It also demonstrated a strong ability to

discriminate between churn and non-churn classes with an AUC-ROC score of 0.831. Key factors influencing churn included contract type, internet service quality, and the presence of multiple lines, highlighting critical areas for targeted retention efforts.

Contributions to Knowledge

The results show that this study makes an important contribution to the area of customer retention by proving the efficiency of the random forest algorithm for predicting high-risk potential churners. It also provides a comprehensive evaluation of various machine learning models, offering valuable insights into their comparative performance. The study underscores the importance of feature importance analysis in understanding the key drivers of customer churn, thereby informing more effective and personalized retention strategies. These contributions add to the existing body of knowledge by bridging the gap between predictive modeling and actionable business strategies in the telecommunications industry.

Future Research Directions

While this study has yielded valuable insights, it also highlights several avenues for future research to further enhance customer churn prediction and retention strategies.

Exploring Advanced Machine Learning Techniques

It would be interesting in future studies to try out different artificial intelligence methods, for example those based on so-called deep learning or ensembles. The examination of neural network methods for modeling, gradient boosting machines or hybrids would allow new insight into consumer behaviour and fluctuations.

Incorporating Additional Data Sources

The use of further explanatory variables in the model would be expected to increase accuracy. Future research might use information about client contacts, public opinion on the Internet or in newspapers, and statistics for the complaint offices as an example. For example, incorporating social or geographical information would make possible an even finer analysis by segmenting customers.

Longitudinal Studies

The company would be able to conduct longitudinal surveys in which individual customers are followed through their life as members of the company. This approach would help in identifying

early warning signs of churn and understanding how changes in customer behavior and preferences influence their decision to stay or leave.

Real-Time Churn Prediction

The future may see new efforts in this direction aimed at devising predictive models for churn behavior which would allow telecommunications firms to identify those subscribers who are likely to leave them. By implementing actual time series analyses for the processed data in realtime one might be able to achieve an improvement both with respect to timing as well as efficiency.

Personalization and Customization

The effect that individualized retention efforts have on customer satisfaction and commitment would also make an interesting subject for further study. The purpose of this research must therefore be to determine by means of appropriate experiments whether any particular measure has an influence upon customer retention. The decisive factor for this would probably be whether or not individualized services are offered.

Ethical and Privacy Considerations

As client information becomes more widely available for prediction purposes, further studies are needed on issues such as ethics or confidentiality. Understanding customer attitudes towards these interventions is also important here. "There's still room for improvement in balancing prediction precision against personalization," says Klaas van der Linden, who heads up the team at Deutsche Telekom that works on this technology. The team has already achieved some success: In Germany alone they have reduced network congestion by more than ten percent thanks to their algorithms.

Final Thoughts

The results show that there are very promising opportunities for using machine-learning algorithms (in particular random forest) when trying to predict which customers would be worth retaining. This makes it possible in particular to predict future behavior on the part of consumers with high precision, which allows telecommunications companies to actively counteract any potential risks by offering appropriate services. This work provides new knowledge that offers great opportunities for further development and application in telecommunication technology. It shows how important an exact analysis of facts and observations really is.

References

Beers Brian, Dr. Brown, JeFreda R. Brown. (2023, May 8). *Investopedia*. Retrieved from Telecommunications Sector: What and How To Invest in It: https://www.investopedia.com/ask/answers/070815/what-telecommunications-sector.asp

Frederick F. Reichheld and W. Earl Sasser, Jr. (n.d.). *Zero Defections: Quality Comes to Services*. Retrieved from Harvard Business Review: https://hbr.org/1990/09/zero-defections-quality-comes-to-services

Kumar, V., Venkatesan, R., Bohling, T., & Beckmann, D. (2008). Practice Prize Report—The power of CLV: Managing customer lifetime value at IBM. Marketing science, 27(4), 585-599.

Banasiewicz, A. (2004). Acquiring high value, retainable customers. Journal of Database Marketing & Customer Strategy Management, 12, 21-31.

Christensen, C., & Raynor, M. (2003). *The Innovator's Solution: Creating and Sustaining Successful Growth*. Harvard Business School Publishing. ISBN 978-1-57851-852-4.

Bughin, J., Doogan, J., & Vetvik, O. J. (2010). A new way to measure word-of-mouth marketing. McKinsey Quarterly, 2(1), 113-116.

Abbasimehr, H., Setak, M., & Soroor, J. (2013). A framework for identification of high-value customers by including social network based variables for churn prediction using neuro-fuzzy techniques. *International Journal of Production Research*, 51, 1279 - 1294. https://doi.org/10.1080/00207543.2012.707342.

Research Methodology:

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1-26.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. Journal of Open Source Software, 4(43), 1686.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), 18-22.

Grolemund, G., & Wickham, H. (2016). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media.

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Syst. Appl.*, 38, 15273-15285. https://doi.org/10.1016/j.eswa.2011.06.028.

Surya, P., & Anitha, K. (2022). Comparative Analysis of Accuracy and Prediction of Customer Loyalty in the Telecom Industry using Novel Diverse Algorithm. *2022 International Conference on Business Analytics for Technology and Security (ICBATS)*, 1-7. https://doi.org/10.1109/ICBATS54253.2022.9759079.

Y, N., Ly, T., & Son, D. (2022). Churn prediction in telecommunication industry using kernel Support Vector Machines. PLoS ONE, 17. https://doi.org/10.1371/journal.pone.0267935.

Babatunde, R., Abdulsalam, S., Abdulsalam, O., & Arowolo, M. (2023). Classification of customer churn prediction model for telecommunication industry using analysis of variance. *IAES International Journal of Artificial Intelligence (IJ-AI)*. https://doi.org/10.11591/ijai.v12.i3.pp1323-1329.

Quek, J., Pang, Y., Lim, Z., Ooi, S., & Khoh, W. (2023). Customer Churn Prediction through Attribute Selection Analysis and Support Vector Machine. *Journal of Telecommunications and the Digital Economy*. https://doi.org/10.18080/jtde.v11n3.777.

Biau, G. (2010). Analysis of a Random Forests Model. *J. Mach. Learn. Res.*, 13, 1063-1095. https://doi.org/10.5555/2503308.2343682.

Adhikary, D., & Gupta, D. (2020). Applying over 100 classifiers for churn prediction in telecom companies. *Multimedia Tools and Applications*, 80, 35123 - 35144. https://doi.org/10.1007/s11042-020-09658-z.

Lukita, C. (2023). Predictive and Analytics using Data Mining and Machine Learning for Customer Churn Prediction. *Journal of Applied Data Sciences*. https://doi.org/10.47738/jads.v4i4.131.

Halibas, A., Matthew, A., Pillai, I., Reazol, J., Delvo, E., & Reazol, L. (2019). Determining the Intervening Effects of Exploratory Data Analysis and Feature Engineering in Telecoms Customer Churn Modelling. *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)*, 1-7. https://doi.org/10.1109/ICBDSC.2019.8645578.

LITERATURE REVIEW

Aeri, M., Dhondiyal, S., Rana, Y., Rawat, S., Kothari, P., & Adhikari, R. (2023). Customer Churn Prediction in Telecom Services. *2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET)*, 809-811. https://doi.org/10.1109/ICSEIET58677.2023.10303463.

- Park, W., & Ahn, H. (2022). Not All Churn Customers Are the Same: Investigating the Effect of Customer Churn Heterogeneity on Customer Value in the Financial Sector. *Sustainability*. https://doi.org/10.3390/su141912328.
- Tatikonda, L. (2013). The Hidden Costs of Customer Dissatisfaction. *Management Accounting Quarterly*, 14, 34.
- Gupta, S., & Lehmann, D. R. (2005). Managing Customers as Investments: The Strategic Value of Customers in the Long Run. *Wharton School Publishing*.
- Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer Assisted Customer Churn Management: State-of-the-art and Future Trends. *Computers & Operations Research*, 34(10), 2902-2917.
- Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journal of Marketing Research*, 43(2), 204-211.
- Vafeas, M., Hughes, T., & Hilton, T. (2016). Antecedents to Customer Network Effects in the Mobile Telecommunications Industry. *Journal of Business & Industrial Marketing*, 31(6), 751-764.
- Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry. *IEEE Transactions on Neural Networks*, 11(3), 690-696.
- Brown, S. A. (Ed.). (2000). Customer Relationship Management: A Strategic Imperative in the World of e-Business. Wiley. [ISBN: 0471644099, 9780471644095].

- Mirkovic, M., Vučković, T., Stefanović, D., & Anderla, A. (2022). Customer churn prediction in B2B non-contractual business settings using invoice data. *Applied Sciences, 12*(10), 5001. https://doi.org/10.3390/app12105001

- Larivière, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications, 29*(2), 472-484.
- Yang, L. S., & Chiu, C. (2006, November). Knowledge discovery on customer churn prediction. In *Proceeding of the 10th WSEAS international conference on applied mathematics, Dallas, texas, USA*.

Telecom and churn

Akmal, M. (2017). Factors Causing Customer Churn: A Qualitative Explanation Of Customer Churns In Pakistan Telecom Industry.

Mehwish, F. A., Zaffar, A. S., & Sumaira, J. M. (2017). Research Article Autonomous Toolkit to Forecast Customer Churn. International Journal of Current Research. Vol. 9. Issue 12. p.62999-63006.

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A New Hybrid Classification Algorithm for Customer Churn Prediction Based on Logistic Regression and Decision Trees. European Journal of Operational Research. p.1–13.

Sebastian, H., & Wagh, R. (2017). Churn Analysis in Telecommunication Using Logistic Regression. *Oriental journal of computer science and technology*, 10, 207-212. https://doi.org/10.13005/OJCST/10.01.28.

Lu, N., Lin, H., Lu, J., & Zhang, G. (2014). A Customer Churn Prediction Model in Telecom Industry Using Boosting. *IEEE Transactions on Industrial Informatics*, 10, 1659-1665. https://doi.org/10.1109/TII.2012.2224355.

Xia, G., & Jin, W. (2008). Model of Customer Churn Prediction on Support Vector Machine. *Systems Engineering - Theory & Practice*, 28, 71-77. https://doi.org/10.1016/S1874-8651(09)60003-X.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5-32. https://doi.org/10.1023/A:1010933404324.

Grushka-Cockayne, Y., Jose, V., & Lichtendahl, K. (2015). Ensembles of Overfit and Overconfident Forecasts. *Capital Markets: Asset Pricing & Valuation eJournal*. https://doi.org/10.2139/ssrn.2474438.

Yang, P. (2023). Data Visualization and Prediction for Telecom Customer Churn. *Highlights in Science, Engineering and Technology*. https://doi.org/10.54097/hset.v39i.6711.

Lu, J. (2002). Predicting customer churn in the telecommunications industry—An application of survival analysis modeling using SAS. *SAS User Group International (SUGI27) Online Proceedings*, *114*, 27.

Reuber, A. R., & Fischer, E. (2005). The company you keep: How young firms in different competitive contexts signal reputation through their customers. *Entrepreneurship theory and practice*, *29*(1), 57-78.

De Giorgi, G., Frederiksen, A., & Pistaferri, L. (2020). Consumption network effects. *The Review of Economic Studies*, *87*(1), 130-163.

Keiningham, T. L., Cooil, B., Aksoy, L., Andreassen, T. W., & Weiner, J. (2007). The value of different customer satisfaction and loyalty metrics in predicting customer retention, recommendation, and share-of-wallet. *Managing service quality: An international Journal*, *17*(4), 361-384.

Kumar, V., & Reinartz, W. (2016). Creating enduring customer value. *Journal of marketing*, *80*(6), 36-68.

Kisioglu, P., & Topcu, Y. I. (2011). Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey. Journal of Business Economics and Management, 12(3), 438-455.

SAS Institute Inc. (2021). Predictive analytics and customer churn in the telecom industry. Retrieved from https://support.sas.com/resources/papers/proceedings/proceedings/sugi27/p114-27.pdf

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B., & Snoek, J. (2019). Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *Neural Information Processing Systems*.

Dankers, F. J., Traverso, A., Wee, L., & van Kuijk, S. M. (2019). Prediction modeling methodology. *Fundamentals of clinical data science*, 101-120.

Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine learning operations (mlops): Overview, definition, and architecture. *IEEE access*.

El Naqa, I., & Murphy, M. J. (2015). *What is machine learning?* (pp. 3-11). Springer International Publishing.

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. u., & Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. IEEE Access, 7, 60134-60149. https://doi.org/10.1109/access.2019.2914999

Ahmad, A., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0191-6

Umayaparvathi, V., & Iyakutti, K. (2016). A survey on customer churn prediction in telecom industry: Datasets, methods and metrics. *International Research Journal of Engineering and Technology (IRJET)*, *3*(04).

Kiguchi, M., Saeed, W., & 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*, 108491.

Ascarza, Eva. (2018). Retention Futility: Targeting High-Risk Customers Might be Ineffective. Journal of Marketing Research , 55 , 80 - 98 . http://doi.org/10.1509/jmr.16.0163

Ban, G. Y., & Keskin, N. B. (2021). Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science*, *67*(9), 5549-5568.

Kaur, H., & Singh, D. (2020). Predictive analytics in telecommunications: Revisiting strategies in a competitive digital age. *International Journal of Information Management, 53*, 102103.

Lemmens, A., & Croux, C. (2006). Bagging and Boosting Classification Trees to Predict Churn. *Journal of Marketing Research, 43*(2), 276-286.

Utami, Y. (2020). Penerapan Algoritma C4.5 Untuk Prediksi Churn Rate Pengguna Jasa Telekomunikasi. , 8, 69-76. https://doi.org/10.23960/KOMPUTASI.V8I2.2647.

Idris, A., Iftikhar, A., & Rehman, Z. U. (2017). Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling. Cluster Computing. p.1–15.

Gregory, B. (2018). Predicting customer churn: Extreme gradient boosting with temporal data. *arXiv preprint arXiv:1802.03396*.

Chen, H., Tang, Q., Wei, Y., & Song, M. (2021). Churn Prediction Model of Telecom Users Based on XGBoost. *Journal on Artificial Intelligence*. https://doi.org/10.32604/jai.2021.026851.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20, 273-297. https://doi.org/10.1023/A:1022627411411.

Mounika, J., & Sowmya, K. (2016). Data Analytics and Its Usage in Various Sectors. *Indian Journal of Science*, 23, 302-305.

Cao, M., Chychyla, R., & Stewart, T. (2015). Big Data Analytics in Financial Statement Audits. *Accounting Horizons*, 29, 423-429. https://doi.org/10.2308/ACCH-51068.

Ben, A., , A., , O., , K., & Seun, E. (2020). Enhanced Churn Prediction in the Telecommunication Industry. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3577712.

Kim, K., Jun, C., & Lee, J. (2014). Improved churn prediction in telecommunication industry by analyzing a large network. *Expert Syst. Appl.*, 41, 6575-6584. https://doi.org/10.1016/j.eswa.2014.05.014.

Shirazi, F., & Mohammadi, M. (2019). A big data analytics model for customer churn prediction in the retiree segment. *Int. J. Inf. Manag.*, 48, 238-253. https://doi.org/10.1016/J.IJINFOMGT.2018.10.005.

Choros, K. (2010). Real Anomaly Detection in Telecommunication Multidimensional Data Using Data Mining Techniques. , 11-19. https://doi.org/10.1007/978-3-642-16693-8_2.

Quach, T., Jebarajakirthy, C., & Thaichon, P. (2016). The effects of service quality on internet service provider customers' behaviour: A mixed methods study. *Asia Pacific Journal of Marketing and Logistics*, 28, 435-463. https://doi.org/10.1108/APJML-03-2015-0039.

Banu, J., Neelakandan, S., Geetha, B., Selvalakshmi, V., Umadevi, A., & Martinson, E. (2022). Artificial Intelligence Based Customer Churn Prediction Model for Business Markets. *Computational Intelligence and Neuroscience*, 2022. https://doi.org/10.1155/2022/1703696.

Wu, S. (2023). Customer Churn Prediction in the Telecommunication Industry. Advances in Economics, Management and Political Sciences. https://doi.org/10.54254/2754-1169/4/20221017.

Siddika, A., Faruque, A., & Masum, A. (2021). Comparative Analysis of Churn Predictive Models and Factor Identification in Telecom Industry. 2021 24th International Conference on Computer and Information Technology (ICCIT), 1-6. https://doi.org/10.1109/ICCIT54785.2021.9689881.

Yuan, X. (2023). Telecom customer churn prediction in context of composite model. , 12597, 125972P - 125972P-6. https://doi.org/10.1117/12.2672716.

Zhu, M., & Liu, J. (2021). Telecom Customer Churn Prediction Based on Classification Algorithm. 2021 International Conference on Aviation Safety and Information Technology. https://doi.org/10.1145/3510858.3510945.

Khan, Y. (2021). COMPARATIVE ANALYSIS OF PREDICTION TECHNIQUES ON THE BASIS OF TELECOM CUSTOMER CHURN. JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES. https://doi.org/10.26782/jmcms.2021.09.00002.

Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decis. Support Syst.*, 95, 27-36. https://doi.org/10.1016/j.dss.2016.11.007.

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research*. https://doi.org/10.1016/J.JBUSRES.2018.03.003.

R, N., & N, L. (2023). Machine Learning Methods for Predictive Customer Churn Analysis in the Telecom Industry. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1-7. https://doi.org/10.1109/ICCCNT56998.2023.10306395.

Mandić, M., Kraljević, G., & Boban, I. (2019). Performance comparison of six Data mining models for soft churn customer prediction in Telecom. *IJEEC - INTERNATIONAL JOURNAL OF ELECTRICAL ENGINEERING AND COMPUTING*. https://doi.org/10.7251/ijeec1801029m.

Fan, C., Sun, Y., Zhao, Y., Song, M., & Wang, J. (2019). Deep learning-based feature engineering methods for improved building energy prediction. *Applied Energy*. https://doi.org/10.1016/J.APENERGY.2019.02.052.

Huang, B., Kechadi, M., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Syst. Appl.*, 39, 1414-1425. https://doi.org/10.1016/j.eswa.2011.08.024.

Huang, B., Kechadi, M., Buckley, B., Kiernan, G., Keogh, E., & Rashid, T. (2010). A new feature set with new window techniques for customer churn prediction in land-line telecommunications. *Expert Syst. Appl.*, 37, 3657-3665. https://doi.org/10.1016/j.eswa.2009.10.025.

Liu, Y., & Zhuang, Y. (2015). Research Model of Churn Prediction Based on Customer Segmentation and Misclassification Cost in the Context of Big Data. *Journal of Computational Chemistry*, 03, 87-93. https://doi.org/10.4236/JCC.2015.36009.

Wu, S., Yau, W., Ong, T., & Chong, S. (2021). Integrated Churn Prediction and Customer Segmentation Framework for Telco Business. *IEEE Access*, 9, 62118-62136. https://doi.org/10.1109/ACCESS.2021.3073776.

Hui, H. (2009). Research of customer-churn based on customer segmentation. *Computer Engineering and Design*.

Xiaobin, Z., Feng, G., & Hui, H. (2009). Customer-Churn Research Based on Customer Segmentation. *2009 International Conference on Electronic Commerce and Business Intelligence*, 443-446. https://doi.org/10.1109/ECBI.2009.86.

DESCRIPTIVE ANALYSIS

Cooksey, R. (2020). Descriptive Statistics for Summarising Data. *Illustrating Statistical Procedures: Finding Meaning in Quantitative Data*, 61 - 139. https://doi.org/10.1007/978-981-15-2537-7_5.

Rob, R., & Fishman, A. (2005). Is Bigger Better? Customer Base Expansion through Word-of-Mouth Reputation. *Journal of Political Economy*, 113, 1146 - 1162. https://doi.org/10.1086/444552.

Alsagri, H., & Ykhlef, M. (2020). Quantifying Feature Importance for Detecting Depression using Random Forest. *International Journal of Advanced Computer Science and Applications*, 11. https://doi.org/10.14569/ijacsa.2020.0110577.

Yun, K., Yoon, S., & Won, D. (2021). Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process. *Expert Syst. Appl.*, 186, 115716. https://doi.org/10.1016/J.ESWA.2021.115716.

Xue, B., Zhang, M., Browne, W., & Yao, X. (2016). A Survey on Evolutionary Computation Approaches to Feature Selection. *IEEE Transactions on Evolutionary Computation*, 20, 606-626. https://doi.org/10.1109/TEVC.2015.2504420.

Huang, X., Ye, Y., & Zhang, H. (2014). Extensions of Kmeans-Type Algorithms: A New Clustering Framework by Integrating Intracluster Compactness and Intercluster Separation. *IEEE Transactions on Neural Networks and Learning Systems*, 25, 1433-1446. https://doi.org/10.1109/TNNLS.2013.2293795.