A
PROJECT REPORT
ON

SUBMITTED
TO
CENTRE FOR ONLINE LEARNING
Dr. D. Y. PATIL VIDYAPEETH, PUNE

IN PARTIAL FULFILMENT OF DEGREE OF
MASTER OF BUSINESS ADMINISTRATION
BY

Student Name: Smita Dinkar Shinde
PRN No.: 23050203714
ERP ID: 231102013
Specialisation Name: Business Analytics
Address: F204, Puneville, Punawale, Pune – 411033
Mobile No.: 8698232644
Institute: Dr. D. Y. Patil Vidyapeeth, Pune
Academic Year: 2024

PROJECT REPORT – SMITA SHINDE

TOPIC - **PREDICTIVE ANALYTICS FOR CUSTOMER CHURN PREDICTION IN THE TELECOM INDUSTRY**

Dr. D.Y. Patil Vidyapeeth's
CENTRE FOR ONLINE LEARNING,
Sant Tukaram Nagar, Pune.

Certificate / Declaration Regarding Originality
Declaration by the Learner

This is to declare that I have carried out this project work myself in part fulfillment of the MBA Program of Dr. D. Y. Patil Vidyapeeth Center for Online Learning, Pimpri Pune.

The work is original, has not been copied from anywhere else, and has not been submitted to any other University/Institute for an award of any degree/diploma.

Date: 16/01/2025
Signature:
Place:Pune
Name:Smita Shinde

CERTIFICATE

This is to certify that Mr./Ms. –Smita Dinkar Shinde**,** PRN 23050203714 has completed his/her internship at **Alpha Predictions LLP** Starting from **1st July 2024** to **30th December 2024** Her project work was a part of the MBA (ONLINE LEARNING). The project is on **PREDICTIVE ANALYTICS FOR CUSTOMER CHURN PREDICTION IN THE TELECOM INDUSTRY** Which includes research as well as industry practices. He was very sincere and committed in all tasks.

Project Guide **Director**
Anand Irrabatti Snehalkumar Kadam
Date – 12/01/2025

DECLARATION BY LEARNER

This is to declare that I have carried out this project work myself in part fulfillment of the M.B.A Program of Centre for Online Learning of Dr. D.Y. Patil Vidyapeeth's, Pune – 411018

The work is original, has not been copied from anywhere else, and has not been submitted to any other University / Institute for an award of any degree / diploma.

Date: - 16/01/2025 **Signature: -**

Place: - Pune Name: - Smita Shinde

ACKNOWLEDGEMENT

It gives us great pleasure in presenting the preliminary project report on **PREDICTIVE ANALYTICS FOR CUSTOMER CHURN PREDICTION IN THE TELECOM INDUSTRY**
We would like to express our deep and sincere gratitude to my guide, **Mr. Anand Irrabatti** of Business Analytics for his unflagging support and continuous encouragement throughout the project work. Without her/ his guidance and persistent help this report would not have been possible.

We would like to express our gratitude towards Head of Department for his kind cooperation and encouragement which helped us in completion of this project. Furthermore, we would like to acknowledge with much appreciation the crucial role of the staff of DIT, Pimpri, who gavethe permission to use all required equipment and the necessary materials to complete our project.

TABLE OF CONTENTS

Executive Summary

This project focuses on developing a predictive model for customer churn in the telecom industry utilizing advanced analytics techniques. Customer churn, defined as the rate at which customers discontinue their service with a provider in favor of a competitor, presents a significant challenge for telecom companies, impacting their revenue and growth.
The core objective of this project is to harness historical customer data to uncover patterns and identify key factors that contribute to customer churn. By analyzing variables such as customer demographics, service usage, billing information, and customer service interactions, the project aims to build a robust model that can accurately predict future churn events.
Utilizing machine learning algorithms, the predictive model will be trained and validated to ensure its accuracy and reliability. This model will serve as a powerful tool for telecom companies, enabling them to proactively identify at-risk customers and implement targeted retention strategies.
The ultimate goal of the project is to provide actionable insights that will help telecom companies reduce churn rates, thereby improving customer retention and enhancing overall business performance. Through this predictive analytics approach, telecom companies can make data-driven decisions, optimize their customer relationship management efforts, and achieve a competitive edge in the market.

Introduction
In the highly competitive telecommunications industry, customer retention is vital for sustaining and growing a firm. Customer churn, the phenomenon when customers discontinue their utilization of a company's services, is a significant impediment. It is imperative for telecommunications companies to comprehend the factors that lead to client attrition in order to devise effective retention tactics. This studyexplores the several factors thatcontribute to customer churn, using a data-driven approach to reveal the intricacies of client retention and service termination. The research endeavors to identify the key determinants of churn in the telecommunications sector and offer practical suggestions to tackle this prevalent issue, using a thorough dataset analysis.
Research Questions:
What are the main variables that affect the likelihood of Churn in the telecom industry?

How can focused marketing techniques in the telecom industry boost customer retention by addressing the primary predictors of customer churn?

The telecommunications industry across the world is becoming one of the major sectors and consequently the technical growth and the ever-developing operator number increased the level of competition. Telecom firms are making an effort to subsist in this rivalry market and some measures have been formulated to bring in huge amount of revenues. To enhance the retention time of customers it is important for the companies to lessen the possibility of churn of customer, referred to as “the movement of customer from one service provider to another service provider (Ascarza, et al 2016). The churn of customers’ is considered a major issue in service fields with increased cutthroat services (Ahmad, et al 2019). Many studies (Umayaparvathi and Iyakutti 2016; Jutla and Sivakumar 2005) emphasized that machine learning applications are increasingly effective to predict this situation.
The underlying principle of customer churn prediction in terms of telecom industry is to calculate subscribers approximately who literally feel like to leave from a company they used so far and suggest solutions to prevent considerable churns. Recently, making an estimation of churners before they quit has become necessary in the environment of stiff competition amongst companies. The major role that the telecom industry plays made it all the more significant to build prediction mechanisms alongside the lines of churn prediction. Few studies exhibit the significance of the user retains in this industry. One of the studies exhibits that 1% intensification in the customer retains movement might essentially give rise to the 5% increase in the entire shares of the companies (Kisioglu and Topcu 2010). Huang, et al (2012) predicted that in terms of telecommunication sector, the monthly ratio of customer churn is 2.2% (Yildiz and Albayrak 2018) and the yearly rate of customer churn was 27%.
Customers’ retain in a telecom industry has already become a nightmare as a result of increased competitive services. Brandusoiu and Toderean (2016) proposed an advanced data mining method in order to find customer churn using algorithms of machine learning namely NN (Neural Networks) and SVM (Support Vector Machine). The findings

emphasized that machine learning algorithm performs better in predicting customers' churn. He et al. (2009) used a machine learning model to resolve the issue of churn among customers in big telecommunication firms. Huang et al. (2015) investigated the issue of churn among customers' in the platform of big-data, which intended to emphasize that big data significantly improve the progression of estimating the churn based on the variety, velocity and volume of the records. Ahmad, et al (2019) specified that the social network analysis application attributes improve the outcomes of estimating the churn in telecommunication sector.

Amongst all other sectors, over the decades telecom sector are facing the maximum yearly rate of churn ranging from 20 to 40 percent (Ahna, et al 2006). This essentially has financial consequences on a firm, since it bears five to ten times much to bring in a new user than retaining an old user within the firm (Junxiang Lu, 2003). Nowadays companies intend to build strong relationship with their users (Chen et al., 2012). Consequently, it has become a conviction that the best promotional policy is to retain the old users or more simply to deal with customer churn (Mohammadi et al., 2013). Hashmi, et al (2013) emphasized that machine learning approaches work with high dimensional, big, non- linear datasets with improved prediction accuracy however considered complicated with regards to real-world applications.

Artificial Neural Networks (ANN) is a widespread machine algorithm approach used to deal with complex issues, for example the churn prediction problem (Mozer, et al 2000). Support Vector Machines (SVM), developed by Boser, et al (1992) is supervised algorithm of learning that assess data and identify patterns, specifically applied for regression and classification analysis. Decision Trees (DT) is another application of machine learning which have no impressive performance on acquiring non-linear and complex relations among the characteristics. Au et al (2003) emphasized that comparatively neural networks are likely to present better performance than that of Decision Trees. But, in terms of the problem of churn among customers, the level of accuracy of a Decision Tree could be better, based on the data form (Hadden, et al 2006). Similarly, the Naïve Bayes classifier has shown good results in terms of the problem of churn prediction for the telecom sector (Nath, et al 2003) and comparatively it could also reach better prediction rates (Asthana, 2018) than other extensively used algorithms, such as DT. The KNN and logistic regression algorithms are chosen in this study because KNN algorithm can perform the task of classification without prior knowledge about data

distribution whereas logistic regression is used to estimate the probability of churn as a function of variables set or customer characteristics. The comparison of both the algorithm will help to predict the churn of customer accurately as well as resolving the factors that leads to retention of customer. K-Nearest Neighbour and Logistic Regression algorithm outperforms other algorithms based on several benchmark sets of data.
A comparison between machine algorithm models has been carried out by Khan, et al (2010). The research identified that neural network achieves a little higher than the decision trees and logistic regression. Vafeiadisa, et al (2015) carried out a comparison analysis of customer churn prediction and the findings exhibited that decision trees along with BPN reached accuracy level of 94%, followed by SVM (93%) whilst Logistic Regression achieved low accuracy level of 86%. For customer churn prediction, SVM methods have been broadly examined and estimated to be of increased predictive performance (Coussement and Poel 2008) For churn problem, Naïve Bayes estimates (Sabbeh, 2018) the possibility that a user will continue with his/her service provider or change to another one.

Problem statement

The retention and acquisition of users are the major concerns in telecom industry. The fast growth of marketplace in every business is giving rise to increased subscriber base. Accordingly, companies have recognized the significance of retaining the customers who is on hand. It has become necessary for service-providers to reduce the churn rate of customers since the inattention might negatively influence profitability of the company. Churn prediction contributes to identify those users who are likely to switch a company over another. Telecom is enduring the problem of ever-increasing churn rate. Accordingly, the current study employs machine learning algorithm on big-data platform. Machine learning algorithm techniques facilitate these telecom firms to be protected with efficient approaches for lessening the rate of churn. Silent churn is one type which is considered complicated to predict since there might have such kind of users who might probably churns in the near future. It must be the aim of the decision-maker and advertisers to lessen the churn ratio since it is a recognized fact that comparatively

existing customers are the most beneficial resources for companies than acquiring new one.

Aims and objectives

The primary and secondary objectives of the study are as follows:

Primary objectives

i. To explore the customer churn prediction in telecom using machine learning in big data platform

Secondary objectives

To investigate the impact of customer churn in telecom industry as a whole
To discuss the significance of customer churn models in telecom industry
To compare the algorithms that are effective in reducing churn rate in telecom companies

Research Questions

What is the role of customer churn prediction in telecom using machine learning in big data platform?
What are all the significances of the customer churn models in telecom industry?
What are all the overall impacts of churn rate in telecom industry?
Which algorithms are all effective in reducing churn rate in telecom companies?

Significance of the study

The major contribution of the current study is to present a model of churn prediction that helps telecom companies to find users who are more intending to churn. The model formulated in this employs ML algorithms on the platform of Big-Data and develops an

innovative way of reducing churn. The significance of this purpose is apparent, given the reality that the expenditure for acquisition of customer is comparatively higher than that of the retention of customer. In consequence, techniques to develop and apply machine learning models are considered necessary and are critical business intelligence applications.

Limitations

The current study is limited to telecom industry only
The study does not use other techniques rather than machine learning techniques

Chapterization plan

Chapter 1: Introduction

The introduction chapter of the proposed research offers the brief structure about this study "to provide a solution for the prediction of customer churn in telecom sector using machine learning in big data platform" involving the background of the research, problem statement, aims and objectives of the research, research questions, significance of the research, as well limitations of the research.

Chapter 2: Literature Review

The Review of literature chapter describes several works associated to the customer churn prediction concept in telecom sector using machine learning in big data platform. In addition to these, this research investigates briefly about techniques involved in machine learning algorithm for acquiring accurate prediction level of telecom customers.

Chapter 3: Research Methodology

The research methodology chapter provides an overview about research design, research strategy, sampling plan, sampling design, population of the study, data types, data

collection methods, design of questionnaire, data analysis and interpretation techniques that used in this research. In addition to these, this chapter discusses in detail about the limitations of the current study.

Chapter 4: Data Analysis and Discussion

The data analysis and discussion chapter analyse and discusses the solution for churn prediction of customers in telecom using in machine learning in big data platform. In addition, the study explains the steps that can be implemented in reducing churn rate in telecom sector.

Chapter 5: Conclusion

The conclusion chapter explains about the summary of findings obtained through the discussion section and also provides conclusion to the research topic “customer churn prediction in telecom using machine learning in big data platform” followed by recommendations and suggestions based on the research results.

Chapter 2 : Literature review

Introduction

A large number of researches in the subject of churn prediction are being investigated employing various statistical and machine learning algorithms since a decade. This chapter deals with the recent and most important publications on churn prediction in telecom industry in the recent period.

Impact of customer churn in telecom industry

Tanneedi, (2016) pointed out that customer churn has become a dreadful problem forthe telecom industry since customers never have a second thought to leave if they don't exactly get what they are expecting. There is no benchmark model that deals with the churning issues of telecom companies precisely. The study emphasized that Big Data analytics with machine learning are considered effective as means for identifying churn. The current study makes an effort to predict customer churn in telecom employing Big Data analytics. Statistical analyses and machine learning application such as Decision trees (DT) have been used for three different datasets. From the analytics of DT, decision trees with accuracy rate of 52%, 70% and 95% have been obtained for three different data sources correspondingly. The findings pointed out that the more the quality and volume increases, the lesser the annoyance and possibility of churn can be expected in telecom industry.
Huang, (2015) exhibited in terms of telecom industry churn prediction can easily be done with big data and with 3V's such as volume, variety along with velocity. Findings emphasize that the performance of prediction has been enhanced considerably by employing a big amount of training data, a huge number of features from both operations and business support systems, as well as an increased velocity of processing new data. The study has deployed this prediction technique of churn in one of the largest mobile network operators in China. From a large number of active customers, this technique could impart prepaid customers set who are about to churn, holding 0.96 accuracy rate

for the top 50000 estimated churners in the list. The operations of automated matching retention with the focused essential churners considerably increase their rates of recharge, bringing about a big value for business. Almana, et al (2014) pointed out that the most complicated problem gone through by telecom sector is customer churn. To effectively deal with the churn prediction challenge, the current study has used machine-learning algorithms along with data mining tools. An effort in retaining existing customers could result in a considerable growth in profits as well as revenues. The necessity of retaining old customers pines for precise predictionof customer churn algorithms which are both precise and understandable. The study figured out certain factors that impact customer to churn. Comparative to post-paid customers, the prepaid users are not significantly bound by service contracts and they are more likely to churn even for simplest reasons, which make it quite complicated to predict their churn rate. Customer loyalty is another factor which can be defined by quality of service and customer service delivered by the service providers. Issues such as network coverage might impact customers to switch another network. Other factors which intensify likelihood of customers defecting to the rivalry include poor response to complaints and errors relating to billing services.

Arifin and Samopa (2018) intended to identify the factors that ultimately influence the churn rate in telecommunication companies. Today users won't think to switch providers if they don't get what they strive for, which is referred to as churning in telecom. Customer churning is strongly associated with customer satisfaction. As the expenditure of gaining a new customer is comparatively far higher than expenditure of keeping an old one, mobile networks have now changed their center of attention from acquisition of customer to retention of customer. The findings exhibit that there are three variables that increase the churn rate considerably such as voice and data usage, and reload service and the study recommended that in order for Telecom Company to retain in the market, it is essential for them to concentrate on these three variables.
As pointed out by Phadke, et al (2013) users churn for various reasons. Competitive pricing, service quality and campaigns as some of the most important reasons a user might intend to quit a service provider. Standard customers churn prediction direct on the quality of experience for the mobile network users. The study develops a new method to measure the social ties strength among users, and later apply these tie-strengths to an innovative

impact transmission model across a graph of mobile-call. The total impact, denoting the societal pressures a customer handles to churn, is then integrated into an improved churn prediction application of machine-learning. The ultimate step to forecast churn using social network analysis is to integrate the impact actions above into a standard ML technique to predict customer churn. The study emphasized that the tie-strength along with algorithm of influence propagation developed here could be incorporated into ML based churn prediction algorithms to enhance their level of accuracy.

Dahiya, (2015) pointed out that as a result of the fast development of telecom sector, the service providers are in position to progress towards extension of the subscriber base. To address the demand of retaining in the competitive market, the retention of on-hand customers has become a complicated task. In the current study, the training data is used to develop classifiers employing machine learning techniques. Logistic regression technique and decision trees are certain machine learning techniques used in the current study. Collecting knowledge from the telecom sector could contribute to predict the involvement of the customers as they are likely to leave the company or not. The study emphasized that required action should be commenced by the telecom sector to instigate the acquisition of their related customers for making their marketplace value more stagnant. The study develops a new model for the churn prediction executes it with the WEKA Data Mining software. The effectiveness and the performance of techniques such as decision tree and as well logistic regression have been discussed.

Shrikhande and Verma (2018) expressed that analysis of information and records which are gathered from telecom companies might contribute to figure out the reasons behind customer churn and as well use the same information to keep hold of the customers. The study employs decision tree technique for customer churn prediction. In terms of machine learning, the churn prediction includes three phases, i.e., the training and testing part and prediction section. The input for this downside integrates the information on past demand almost all the mobile subscribers, together with all individual and business info that's retained by the service provider. In addition, in terms of the training section, labels are given within the form of a record of churners. Once the model is exercised with maximum accuracy, the model must be capable of predicting the list of churners from the significant info that doesn't represent any churn label. Employing decision tree-based data mining algorithms contributed to precisely predict customer churn and also classifying the causes that result in customer retention.

Alwin (2018) expressed that when the reason behind customer churning is known to the service providers, it is possible for them to enhance their services to accomplish the demands of the customers. Churns could be considerably lessened by investigating the earlier history of the potential users analytically. The framework has been developed using algorithms known as logistic regression and the neural network. In the end, comparative assessment is carried out to find out the most advantageous model and analyze the model with precise and consistent findings. The study recommended the model such as C5.0 algorithm of the decision tree for churn management. And the abovementioned model has been proven to be optimal amongst the models with 85 percent accurateness level and AUC value as 0.888.

Kumar and Chandrakala (2016) pointed out that the rapid growth of the market in every industry is resulting in increased subscriber base for service providers. The current study investigates the most widespread machine learning algorithms employed by researchers so far for churn prediction in every sector which increasingly relies on customer contribution. The three kinds of machine learning discussed in the study are unsupervised, semi-supervised, and supervised. Employing algorithms that iteratively obtain from data, machine learning facilitates systems to investigate unknown patterns without being blatantly programmed where to look. The study concluded that a good prediction model is possible with a combination of SVM and boosting algorithms for better accuracy level and performance that could be contemplated as a future work for customer churn prediction.

Santharam and Krishnan (2018) pointed to that customer churn has become a challenging issue to various business and sectors and particularly influencing the rapidly developing telecommunication industry. As a result of the increasing competitiveness and inventive business models, the expenditure of customer acquisition has been intensified to a large extent. It is consequently significant for any service providers to carry out churn prediction. Generally, various machine learning techniques are used to execute customer churn prediction with enhanced accuracy. The study investigates various techniques employed not just in communication sector but as well in other industries wherein customer participation is increasingly active. The study emphasized that comparatively SVM has a good prediction precision, strong generation capability and enhanced precision than that of logistic regression and naïve bayes classifier.

Saini, et al (2017) explained that certain factors including low switching costs have contributed to the risk of users switching to another operator. Customer churn could consequently be described as the switching of a user from one services provider to another. The current study intends at predicting customer churn employing technique such as Decision Trees, one of the most broadly employed classification technique in machine learning and data mining. Different costs are related with customer churn that essentially integrate loss of returns, expenditures in terms of customer retention, organizational and as well budgeting chaos. Various machine learning techniques such as linear, artificial neural networks, Decision Trees have largely been used to find out churners and active customers. It was evident from the study that technique known as Exhaustive CHAID has proven to be more effective and precise than other techniques to predict the customers who are more expected to churn in the near future.

Importance of churn prediction model in telecom industry

Sjarif, et al (2019) emphasized that it is important for Telecom Company to have a churn prediction model in order to prevent their user from moving to another operator services. Consequently, the underlying principle of this study is to develop the customer churn prediction model. Machine learning can possibly be the sort of tools which could help telecom companies in churn prediction model. Machine learning is a kind of artificial intelligence tools which give the capability to let computer learns the algorithm instinctively without human contribution. Comparatively, churn prediction in telecom has been considered as unique application domain to churn prediction than other subscription- based industry as a result of the variety, volume and biases of the information. On the basis of the findings, the study noticed that the KNN algorithm surpasses the others with the accurateness for training and testing is the ratio of 80.45% and 97.78% respectively.
Amin, et al (2016) make an attempt to develop the model of churn prediction in the telecom sector. The study presents a technique of rule-based decision-making, on the basis of RST (rough set theory), to obtain significant rules of decision linked with non churn and customer churn. The proposed technique efficiently executes categorization of churn from non-churn users, together with prediction of those users who will churn or might likely to churn in the near future. Experiential findings exhibit that rough set theory

based on Genetic Algorithm (GA) is the most effective method for obtaining inherent knowledge in decision-based rules form from the publicly accessible, benchmark telecom information. Besides, comparative results exhibit that proposed technique provides a worldwide best solution for churn prediction in the telecom industry, when benchmarked against some high-tech techniques. In the end, the study exhibits that how attribute-level analysis could contribute to develop an effective policy of customer retention that can form an essential part of strategic process of decision-making in the telecom industry.

Dong, et al (2017) expressed that customer churn prediction is extremely essential for telecom operators in order to retain valuable customers. Accurate features which could classify customer behaviors and also effective extraction techniques are most important factors in developing the customer churn analysis model. In the current study, Support Vector Machine (SVM) has exhibited its applicability to the issue of customer churn assessment. The current study paper figures out the most important features that impact the customer churn technique from telecom experts' perspective, and develop an appropriate one based on SVM. In the end, an experimental finding also is demonstrated to validate rationality of the proposed techniques.

Mamčenko and Gasimov (2014) studied the customer churn prediction in mobile service providers employing combined model. Along with lost revenue, customer churn denotes activation and deactivation costs to a large extent. The study used techniques such as analysis such as decision tree and logistic regression but concentrates largely on defining the reasons why users make a decision to churn. The proposed technique in the study is able to identify prospective churners at the contract level for a particular prophecy time- period. The author emphasized that churn is not only numerical occurrence but also must be investigated from sociological viewpoint. A number of reasons which result in customer churn are as well examined. Establishing reasons of churn generally take place on the basis of questionnaires and surveys.

Qureshi, et al (2013) expressed that churn prediction models have developed as one of the most critical Business Intelligence (BI) applications which intends at finding customers who are likely to switch to other operators. The current paper makes an effort to deal with widely used data mining along with machine learning techniques for the classification of customers who are likely to churn. The study as well discusses the application of re-sampling technique so as to resolve the issue of class imbalance. The

findings of the study exhibit that in case of the dataset employed, DT is considered the most precise technique of classifier whilst finding essential churners. The findings have been compared on the basis of the F-measure and precision values. The study effectively resolved the issue of class imbalance. The overall accurateness in the study was 75.4%. The optimal results have been observed with algorithm known as Exhaustive CHAID, a deviation of DT algorithm.

Nigam, et al (2019) dealt with machine learning techniques for customer churn prediction. A large number of machine learning techniques have been employed in past for predicting customer churn such as Decision trees, SVM, Logistic Regression, SVM, NN, etc. Machine learning entails constructing algorithms which could learn from dataset available and could be employed to make predictions on information. One of the most important techniques to predict customer churn in telecom sector is deep neuralnetwork. By employing it is possible to build model that corresponds our data employing different hierarchies of concepts therefore intensifying the performance of the model developed. The current study used the multi-layer artificial neural network (ANN), which is also referred to as deep neural network to predict telecom customer churn. The proposed model in the study has obtained sensitivity of 85% and consequently the findings are satisfactory.

Sharma and Panigrahi (2011) put forward that a slight change in the retention ratiocould possibly result in considerable effect on business. The study also dealt with the importance of churn prediction model. Machine learning techniques, particularly neural networks, frequently do better than conventional statistical and structurally restrictive methods for example linear analysis approaches. The current paper puts forward ANN technique to identify an optimal model from recorded customer dataset to forecast churn and to inhibit the customer's turnover. This paper develops a NN based method to estimate customer churn in cellular wireless services. The findings of experiments point out that NN based method could predict customer churn with precision rate more than 92%. Besides, it has been noticed that medium sized NNs act effectively for the customer churn prediction as different neural network's topologies have been investigated.

Comparison of various algorithms in customer churn

Yabas and Chankya (2013) specified that customer churn has become a matter of great concern of customer care management for the majority of the mobile service providers as a result of its associated costs. The current study describes our work on customer churn investigation and assessment for such services. The study has employed data mining algorithms to precisely and effectively predict subscribers who will change and ultimately switch to another service provider or competitor for the same or related service. The study makes an effort to identify alternative techniques which could correspond or enhancethe recorded high scores with more effective and as well practical usage of resources. The paper also focuses on collection of meta-classifiers that have been investigated separately and chosen in line with their performances.

Talwar and Dahiya (2015) expressed that customer churn prediction is one of the foremost features of current telecom customer relationship management systems. This article deals with a contemporary machine learning technique entail in churn prediction. This research makes assessments on recurrently used ML techniques to classify customer churn patterns in telecom sector. The classifiers challenges which are tackled in telecom churn prediction, the current paper makes an effort to propose different hybrid approaches, wherein ensemble classification systems are classically combined with pre-processing methods. The study points out that machine learning classifiers perform well if there is sufficient human attempt made in feature engineering; consequently, it is potential to obtain a rational boundary of the classes in feature space. Machine learning techniques will be effective means for formulating an effective and automated technique for churn prediction.

Das and Gondkar (2018) stated that the customer loyalty is considered lagging, which increasingly result in customer churn as a result of the competitive world, fluctuations in price and the intensifying benefits from the competitive company. In this study, an investigation on different machine learning algorithms in addition to the complications of customer attrition prediction is depicted. The study pointed to that a large number of ML algorithm for prediction of churn has been observed in telecom sector and Support Vector Machine (SVM), ANN are broadly used algorithm for churn analysis. It is also evident from the research that combination of the two-step procedure of ANN for training and comparatively collective method of SVM for testing gives better accurateness level with

increased Area under the Curve (AUC) than that of other available techniques. It is also noticed that both ANN and SVM have a better possibility of churn prediction and it could be employed for the optimal prediction level.
Jayaswal, et al (2016) emphasized the expansion of globalization and market liberalization as the factor that increase the market competitiveness considerably. The emergence of current technology in business processes has increased the competition and set forth new challenges for service providers. The incident of stop using the particular company by a user is referred to as churn and in this framework, predicting the customer's intent to churn is known as churn prediction. Both data mining and machine learning algorithms, when applied to consumer behavior and usage dataset could contribute tothe churn management processes. The current paper employed customer usage and associated dataset from telecom service providers to investigate churn in telecom sector. The decision trees and its types and Gradient Boosted trees are employed as fundamental statistical machine learning techniques for developing the binary churn classifier. The execution part has been carried out using apache spark which is developed unified data analysis structure for machine learning. To accomplish better and effective outcomes, the grid centered hyper-parameter optimization is used.
Radosavljevik, et al (2013) pointed out that various machine learning techniques including decision trees, NB, NN and genetic algorithms, are largely employed algorithms to develop the tabular churn prediction models. To reduce the customer churn rate, it is important for telecom operators to form defensive techniques to find and propose the appropriate incentive to users with increased churn tendency. This study looks at the level to which social network characteristics obtained from the graph formed by communications among customers could be applied to enhance churn prediction accurateness in the prepaid division. The study highlighted that instead of aiming at all future churners, we could reduce our resources by concentrating only on churners with increased influential power.
Madan, et al (2015) aimed at investigating the recent literature in the field of telecom customer churn largely with two viewpoints, that is, method being applied to churn prediction in telecom and as well the publication year. The study emphasized that neural networks have the major advantage which they could recall dataset on the basis of incomplete or noisy data. Consequently, they carry out well with data intensive functions.

The study found out that hike in the due bills is one of the factors that become a significant area of concern for telecom sector as a whole. In such a competitive atmosphere, companies could not tackle the problem of insolvency. In order to deal with such bankrupt users, data mining and machine learning technique could be employed. The study makes an effort to obtain some interesting frameworks, on the basis of which we might be capable of identifying the reasons for churn and it is also possible to predict who are more likely to churn in the near future.

Gupta, et al (2018) proposed a hybrid-model machine learning technique to calculate distortion in mobile telecom networks. The study makes an effort to define the machine learning techniques frequently employing churn executions. The experiments have been performed employing different tools for machine learning, along with a set of real-time data from the open data service provider to assess the manufacturer's efficiency. The findings exhibited that the new hybrid model is comparatively more accurate than thatof the individual methods. Auto learning is a division of artificial intelligence, generally used geometric techniques to provide computers the capability to "learn" the dataset. SVM, The Bayesian Network, ANN and The Recurrent Neural Network (RNN) are some of the machine learning algorithms used in the current study. The study concluded that the most accurate pressure prediction is accomplished employing hybrid techniques and not individual algorithms.

Keramati, et al (2014) specified that customer churn prediction has turned out to be a critical issue in telecom business. In such a decision tree, SVM and ANN for customer churn prediction in telecom sector. A hybrid competitive business world, a consistent user predictor will be considered priceless. The study used techniques methodology has been developed in the study that fundamentally made substantial growths to the value ofsome of appraisal metrics. Findings exhibited that above 95 percent level of accuracy for precision and recall is clearly attainable. A new technique for obtaining influential characteristics is presented and discussed.

Azeem and Usman (2018) pointed out that in order to tackle customer churn, a large number of machine learning (ML) based churn prediction algorithms have been developed in the recent years. There are different classification techniques which have been used in studies for model building in churn prediction field, which is ranging from linear classifiers to the state of the art machine learning including boosting and SVM. The

research investigates the efficiency of our retention movement by measure what ratio of the recognized churners didn't churn as a result of the well-timed and targeted campaign. An automated and retention technique has been developed for the potential churners by classifying the customers wisely on the basis of complaint severity and user's usage pattern. In the end, suitable campaign is generated automatically for targeted prospective churners and effectively managed to accomplish the retention accuracy level up to 88%.
Ebrah and Elnasir (2019) studied the ML applications and suggested the optimal solutions for telecom industry. In the rivalry telecom industry, customers are more likely to change from one service provider to another service provider, which makes service providers concerned about their users and how to keep hold of them but they could forecast the customers who will switch to another provider formerly by assessing their behaviour. In this study, three ML techniques have been employed to predict churn i.e., SVM, Naïve Bayes and decision trees. The proposed models' performance has been assessed by the AUC and accuracy score of each model was 0.82, 0.87, and 0.77 correspondingly. The proposed models as well have shown better accuracy than the earlier literature using the similar datasets.

Table 2.1: Comparison of various algorithms in customer churn Source: Author

S. No	Author	Algorithm	Findings
1	Yabas and Chankya (2013)	Data Mining	The author used data mining techniques in order to find exact prediction of customer churn and found increased accuracy level.

2	Madan, et al (2015)	Neural network	Data mining contributes to telecom industry in terms of churn prediction, insolvency prediction and fraud detection.
3	Gupta, et al (2018)	Hybrid-model machine learning technique	The findings reported that the proposed hybrid model is comparatively more accurate than that of the individual methods.
4	Azeem and Usman (2018)		The formulated retention strategy on the basis of churner severity positively contributes to retain around 87% of the potential churners.

Research Gap

Lemmens and Gupta (2013) used Stochastic Gradient Boosting models for analyzing churn prediction in telecom industry. Kaur and Mahajan (2015) have used data mining and R tools in predicting churn rate in telecom industry. Gursoy, (2010) has used data mining tools for predicting the behavior of customer churn in telecom sector and found it effective. The author used data mining techniques such as decision tree and logistic

regression. Adebiyi, et al (2016) used logistic regression for studying customer churn and retention decision in the telecom industry. Neural networks, classification trees and regression have been used by Poel and Lariviere (2004) for customer churn prediction. Hadden, et al, (2006) used Matlab and Dwivedi, et al (2019) has employed SAS Enterprise Miner for churn prediction in telecom. Though, machine learning found effective in customer churn prediction, the majority of studies used data mining and machine learning techniques have been neglected. Thus, the current study makes an effort to fill this research gap and used machine learning for predicting customer churn in telecom.

Summary

There are three underlying objectives of the study. One is to explore the "customer churn prediction in telecom using in big machine learning data platform." Second, was to analyze the importance of churn prediction model in telecom industry in order to help telecom industry to inhibit customer churn rate in its initial stage itself. Third, was to compare the algorithms that are effective in reducing churn rate in telecom industry. In this case, various algorithms relating to customer churn prediction have been compared and discussed its effectiveness. The study positively contributes to telecom industry to suggest the optimal solutions in reducing churn rate in telecom industry.

Chapter 3 : Design of the System

Introduction

This chapter design customer churn prediction in telecom using machine learning in big data platform. This study makes use of logistic regression and KNN with big data for predicting consumer churn in the telecom sector.

Proposed System

In a business scenario predicting customer churn is where a firm is attempting to retain customer which is much probable to leave the services. For reducing the rate of churn this study classifies which customers are much going to churn probably and which will not churn probably. Since obtaining new customers is challenging it is essential to retain present customers. Churn can be decreased by examining the essential customers past history systematically. Huge amount of data is managed about the customers and on carrying out appropriate examination on the same it is feasible to find probable customers that might churn. The data that is feasible can be examined in varied ways and thereby offers different ways for operators to imagine the churning of customers and avoid the same. The below figure shows the steps used for proposed system.

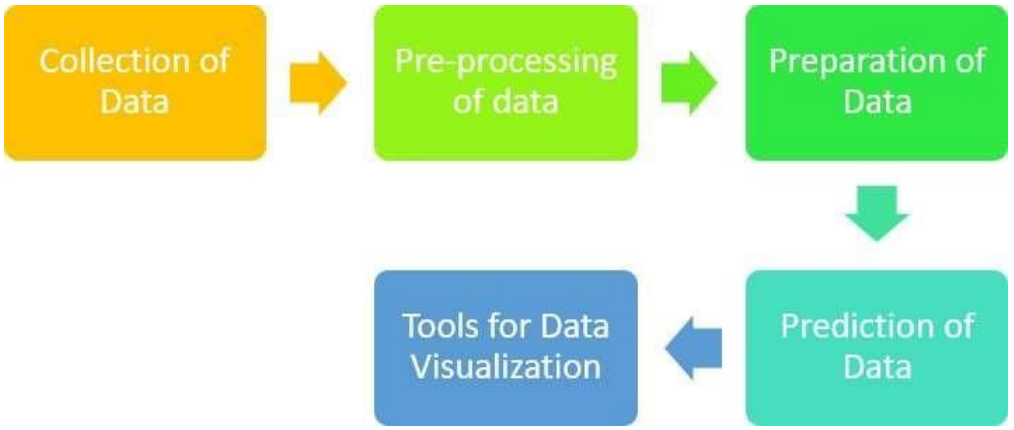


Figure 3.1: Steps used for Proposed System.
Source: Author

From the above figure 1 the steps used for the proposed system are collection of data, Pre-Processing of data, Preparation of data, Prediction of data and Tools for data visualization. The steps are explained below briefly:

Collection of Data

The data that is feasible for analysis in telecommunication dataset has been used and the prediction has been carried out for the same.

Pre-processing of data

The pre-processing of data involves 3 steps namely data cleaning, feature selection and data transformation. Each step is explained below:
Data transformation comprises of two explanatory variables which can be transformed from binomial form into binary form to be much applicable for the chosen models.

The data cleaning step involves missing data imputation or handling. Some of the chosen algorithms cannot manage missing data that is why missing value can be transformed by median, mean or zero. However, the replacement of missing data by computed value statistically is a better choice. The used set of data involves missing values in certain numerical variables and two categorical variables.
Before training of model, feature selection is one of the most essential factors that can influence the model's performance.

Preparation of data

The main purpose of preparation of data is to improve the quality of data and enhance the performance of data analysis. The preparation of data requires to be undertaken in a much iterative way until a conclusive result is met. The processes of preparation of data involves numerical variables discretization, missing values imputation, selection of feature of most informative variables, transformation from one discrete value set to another and derivation of new variables. The process of imputation includes changing the missing values with whole data based on an estimate from finished values. Making new variables from the information is based on transformation and discretization. Two new variables were formed to estimate the voice and transformation in usage of data. Before the data can be examined the data must be cleaned and keep it prepared so that the desired outputs can be derived from it. Data must be clean so that the errors and redundancy can be eliminated because having such information will lead to improper outcomes as well. In this study a churn examination has been used on telecommunication data here the agenda is to know the feasible consumers that might churn from service provider. The end outcome provides the churn probability for each consumer. To perform the churn examination the logistic regression is used. Logistic regression is a statistical approach where the output variable is categorical rather than continuous. Logistic regression restricts the prediction to be one and zero interval.

Prediction of data

The organization is concerned in the final product and it is very essential to indicate their outcome in a graphical representation such a way that is understandable and the output helps organization makes the required predictions which in turns brings profit. There are several components that helps accomplish the same.

Tools for Data Visualization

The best way to acquire the message across is to use the tools of visualization by indicating data visually it is feasible to uncover the essential patterns and the patterns that would be ignored if the statistics alone is considered. In this study Power BI is a component that is used to perform data visualization. Power BI is a business analytics component which is offered by Microsoft using which reports can be made. In this study the data is cleaned already and the output is populated in a file named prediction which will be helpful to visually display how the data seems and the effect of it.

Tools for Data Visualization in Customer Churn Analysis

Data visualization plays a crucial role in understanding trends, patterns, and relationships within a dataset. It enables analysts to convey insights effectively, making it easier for stakeholders to interpret results and make informed decisions. In the context of customer churn analysis, visualization tools are instrumental in presenting factors influencing churn, such as tenure, monthly charges, contract type, and service usage. Below are some of the most widely used tools for data visualization in this project:

1. Matplotlib

Matplotlib is one of the most popular Python libraries for creating static, interactive, and animated visualizations. It provides a foundation for building various chart types, such as line plots, bar plots, histograms, and scatterplots. In this project, Matplotlib was used extensively to visualize trends, such as churn rates across different customer segments. For example, bar charts were created to compare churn rates for customers with different contract types, and histograms were used to analyze the distribution of tenure and monthly charges.

2. Seaborn

Built on top of Matplotlib, Seaborn simplifies the creation of aesthetically pleasing and informative visualizations. It is particularly useful for statistical analysis, as it offers advanced capabilities for visualizing distributions, correlations, and categorical data. Heatmaps generated using Seaborn were instrumental in identifying correlations between features, such as tenure, monthly charges, and churn. Boxplots and violin plots were also used to compare numerical features across churned and non-churned customers, providing deeper insights into feature distributions.

3. Tableau

Tableau is a powerful data visualization tool known for its user-friendly drag-and-drop interface. It enables analysts to create interactive dashboards and share insights with stakeholders. In this project, Tableau was used to create interactive dashboards showcasing churn metrics, trends, and geographical distributions of churned customers. For example, maps were used to visualize regional churn rates, while dynamic filters allowed stakeholders to explore churn patterns for specific customer groups, such as those with high monthly charges or shorter tenure.

4. Power BI

Power BI, a Microsoft product, is another widely used tool for creating interactive dashboards and reports. Its integration with various data sources and advanced analytics features makes it an excellent choice for business intelligence projects. In this analysis, Power BI was leveraged to present dynamic visuals of customer churn, such as cohort analysis charts, time-series trends, and service usage patterns. Its ability to connect directly to databases and provide real-time insights helped facilitate data-driven decision-making.

5. Plotly

Plotly is a Python library that specializes in creating interactive, web-based visualizations. Its ability to produce interactive charts, such as scatterplots, 3D plots, and choropleth maps, made it an invaluable tool for exploring churn patterns dynamically. For instance, scatterplots with hover functionality allowed the team to analyze clusters of churned customers based on monthly charges and tenure. These interactive elements provided a richer exploration experience compared to static charts.

6. Excel

Microsoft Excel remains a simple yet effective tool for data visualization, especially for quick analyses and

smaller datasets. In this project, Excel was used to create pivot tables and basic charts, such as pie charts and bar graphs, to summarize churn metrics. While limited in handling complex visualizations or large datasets, Excel's familiarity and ease of use made it a convenient tool for initial exploratory visualizations and stakeholder presentations.

Analysis of dataset

This study uses Kaggle website for dataset in predicting and analyzing churn. Kaggle is a site and community for hosting ML competitions. Rivalry ML can be a best way to practice and develop their skills as well as explain their abilities. Kaggle permits users to publish and find sets of data and describe models in a web-based data science surroundings, perform with other scientists of data and ML engineers and enter competition to resolve the barriers of data science. The pre-processing steps used for dataset are: 1) first the spaces are replaced with values of null in the column of total charges; 2) the values of null are reduced from the column of total charges which comprises 15 percent missing data; 3) then the data is converted to the type of float; 4) after than no internet service is replaced to no for the following columns: DeviceProtection, StreamingTV, OnlineSecurity, TechSupport, StreamingMovies and OnlineBackup; 5) the values for SenioCitizen is replaced with 0 as No and 1 as Yes; 6)

Then the categorical column is made into Tenure; 7) After than the churn and non-churn customers are separated; 8) Finally the numerical and categorical columns are separated.

Algorithms Used
K-Nearest Neighbor

According to Keramatia et al (2014) K-Nearest Neighbor is one of the most useful and applicable non parametric algorithms of learning. K-Nearest Neighbor is also referred as lazy algorithm that is entire data of training is used at the phase of testing. There is no phase of training and entire points of data are used directly in the testing phase so these entire points required to be employed when it must be tested. K-Nearest Neighbor utilizes the distance between records so as to utilize it for classification. In order to estimate the distance between points K-Nearest Neighbor considers that these points are multidimensional or scalar vectors in feature space. All points of data are vectors of feature space and the label will refer their classes. The easiest case is when the class labels are binary but still it is useful on arbitrary class numbers. In K-Nearest Neighbor one parameter requires to be tuned. K is the number of neighbors/instances that are regarded for instance labeling to some class. The cross validations were carried out using different values of k. K-Nearest Neighbor does not attempt to build an internal structure and computations are not carried out until the time of classification. K-Nearest Neighbor stores only examples of the training information in feature space and the class of an example is decided based on most of the votes from its neighbors. Instance is labelled with class which is much similar among its neighbors. K-Nearest Neighbor decides neighbors based on hamming for categorical variables and distance using Manhattan, Murkowski and Euclidian measures of distance for continuous variables. Estimated distances are employed to recognize training instances set that are nearest to the new point and allot label from these. Despite its simplicity K-nearest neighbor have been used to different kinds of application. For churn K-nearest neighbor is used to examine if a customer churns or not based on features proximity to consumers in every classes (Keramati et al, 2014).

K-Nearest Neighbors (KNN) for Customer Churn Prediction

K-Nearest Neighbors (KNN) was explored as one of the predictive models for identifying customer churn. KNN is a non-parametric, instance-based algorithm that predicts the class of a data point by considering the classes of its nearest neighbors in the feature space. This simplicity and intuitive approach make KNN an attractive option for churn prediction, especially in cases where relationships between features and the target variable may not follow a clear mathematical model.

The first step in implementing KNN involved normalizing the dataset to ensure that all features contributed equally to the distance calculations. KNN relies on distance metrics, such as Euclidean distance, to determine the "closeness" of data points, so features with larger ranges (e.g., monthly charges versus binary variables like contract type) could dominate the calculations without normalization. Min-max scaling and standardization were applied to the numerical features to mitigate this issue.

The choice of the number of neighbors (k) was a critical hyperparameter in KNN. Smaller values of k, such as 1 or 3, tend to make the model sensitive to noise, while larger values can over-smooth the decision boundary and overlook finer patterns in the data. To determine the optimal value of k, cross-validation was performed, testing different values and selecting the one that maximized model performance based on metrics like accuracy and F1-score. For this dataset, a value of k = [XX] provided the best balance between underfitting and overfitting.

To evaluate KNN's performance, metrics such as accuracy, precision, recall, F1-score, and ROC-AUC were used. While KNN performed reasonably well in terms of accuracy ([XX]%), its recall was slightly lower compared to tree-based models like Random Forest and XGBoost. This indicated that KNN might not be as effective at identifying all high-risk customers (minimizing false negatives), which is a priority in churn prediction. However, its interpretability and straightforward nature made it a valuable model for comparative purposes.

KNN's performance was highly influenced by the dataset's size and the class imbalance. Since KNN requires calculating distances for all data points in the training set during prediction, it can become computationally expensive as the dataset grows. Additionally, the presence of class imbalance (fewer churned customers) affected the model's ability to predict minority classes effectively. Techniques like oversampling (SMOTE) and

stratified splitting were applied to mitigate this issue, improving the model's ability to handle imbalanced data. One of the key strengths of KNN is its flexibility in capturing non-linear decision boundaries. In this project, KNN provided insights into customer churn by identifying clusters of churned and non-churned customers in the feature space. For example, it highlighted groups of customers with high monthly charges and short tenure who were more likely to churn. However, this clustering effect was more pronounced in smaller subsets of the data and became less efficient as the dataset's complexity increased.

In summary, KNN offered a simple yet effective approach for predicting customer churn and provided a useful baseline for comparison with more advanced models. While it performed reasonably well on normalized data, its sensitivity to noise, computational cost, and limitations in handling large datasets made it less suitable for deployment in this specific use case. Despite these challenges, KNN contributed to the project by offering an alternative perspective on churn behavior and validating insights derived from other algorithms. Its role in this analysis underscores the importance of experimenting with diverse models to ensure a comprehensive understanding of customer churn.

Logistic Regression

Logistic regression is the proper model of regression analysis to utilize when the dependent variable is binary. Logistic regression is a predictive examination used to describe the relation between an independent variable set and dependent binary variable. For churn of customer logistic regression has been used to estimate the probability of churn as a function of customers characters or variables set (Sahu et al, 2018). According to Hassouna et al (2016) Logistic regression is also used to find the customer churn occurrence probability. Logistic regression is based on a mathematically oriented method to examine the impact of variables on others. Prediction is made by comprising a group of equations linking values of input with the output field. The mathematical formulas for logistic regression are:

$P(b = 1|a1, \ldots . , am) = F(b)$ (3.1)

$F(b) =$ 1 (1+$e-b$) (3.2)

$b = \beta 0 + \beta 1a1 + \beta 2a2 + \cdots + \beta mam$ (3.3)

Where $\beta 0$ is a constant, b is every individual e target variable, y is a binary label class one or zero, $a1. a2, \ldots ,$ $+am$ is the variables of predictor for every customer e from which a is to be predicted.
The datasets of customer are examined to comprise the equations of regression and an evaluation process for every customer in the set of data is then carried out. A consumer is at a risk of churn if the value of p for consumer is larger than a predefined value.

Logistic Regression for Customer Churn Prediction

Logistic regression was one of the primary models used in this project to predict customer churn. As a statistical method, logistic regression is particularly effective for binary classification problems, where the target variable has two outcomes—in this case, whether a customer churns or remains with the company. The model predicts the probability of churn by analyzing the relationships between the independent variables (features) and the dependent variable (churn status).

During the modeling process, logistic regression proved to be an interpretable and computationally efficient algorithm. It allowed for the identification of key features contributing to customer churn by examining the coefficients of the model. Features with positive coefficients, such as month-to-month contracts and higher monthly charges, were found to increase the likelihood of churn. Conversely, features with negative coefficients, such as longer tenure and subscriptions to tech support services, were associated with reduced churn rates. This level of interpretability made logistic regression a valuable tool for understanding the impact of individual factors on churn behavior.

To optimize the model's performance, hyperparameter tuning was conducted using techniques such as grid search and cross-validation. Regularization methods, including L1 (lasso) and L2 (ridge), were employed to prevent overfitting and improve generalization. These techniques ensured that the model was robust and performed well on unseen data. Additionally, the data preprocessing steps, such as scaling numerical features and encoding categorical variables, played a crucial role in enhancing the model's accuracy and stability.

The performance of the logistic regression model was evaluated using key metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. While logistic regression achieved an accuracy of [XX]% and a recall of [XX]%, it was observed that the model's recall was particularly useful for identifying at-risk customers. By focusing on recall, the model prioritized minimizing false negatives, ensuring that most customers likely to churn were correctly identified. However, logistic regression's relatively lower predictive power compared to more complex models, such as gradient boosting, highlighted the trade-off between interpretability and performance.

Despite its limitations, logistic regression provided valuable insights into customer churn and served as a baseline model for comparison with more advanced algorithms. Its simplicity and transparency made it ideal for explaining the results to stakeholders and aligning the findings with business objectives. The coefficients of the model were instrumental in guiding strategic recommendations, such as incentivizing long-term contracts and addressing pain points related to high monthly charges.

In summary, logistic regression played a pivotal role in the customer churn analysis by offering a clear and interpretable framework for understanding churn behavior. While it was outperformed by more sophisticated models in predictive accuracy, its insights into feature importance and its ease of implementation made it an essential component of the project. Combined with other machine learning models, logistic regression contributed to a comprehensive understanding of customer churn and informed the development of actionable retention strategies.

Implementation of the system

For building the system of churn prediction in Telecom Company, platform of big data is installed. HDP (Hortonworks data platform) was adopted since it is open source and free framework. It comes under license of Apache 2.0. Such platform has different tools and open source software based on big data. Such tools and open source software are

combined with each other. In HDP, every group of tools is classified under particular specialization such as operations, security, access of data, data management and governance integration. Framework of HDP installation was customized to have required systems and tools for passing through all phases of framework. Hadoop distributed file system (HDFS) was installed for storing the data, spark execution engine for processing the data, Ambari for monitoring the system, Zeppelin as user interface for development, Yarn for managing the resources, Ranger for securing the system and Scoop tool and Flume system for acquiring the data from outside developed framework into HDFS. The system implementation of customer churn prediction in telecom using machine learning in big data platform is presented in this study which integrates hardware and software resources.

System Requirements

Operating System: Ubuntu

Programming Language: Python

Ubuntu

One of the Linux OS versions is Ubuntu. An Operating System is a software that performs the Personal Computer. The most popular operating system for desktop Personal Computer is Windows but Linux is wholly a separate endeavour. Ubuntu is a modern Operating System completely that provides everything which the user might find in Windows or Macintosh OS but without the challenges. It handles things simply yet offers sophisticated features. Ubuntu altered everything and it focuses on the experience of user on desktop and to this end describes the software of graphical design. It includes wide number of hardware drivers so that complete standard hardware performs. Ubuntu comes with a program of installer that does not characterize the technique of mind boggling and system update takes only few clicks of mouse.

Python

A general-purpose language of programming is the Python language. In this programming language users can employ the language for evolving both web applications and desktop. Python is also used for evolving complex numeric and scientific applications. Python is framed with essential features to enhance data visualization and analysis.

Evaluation of the results

In this research four measures have been used for the evaluation of the quality of prediction. The four measures are precision, recall, F-score and accuracy.
Precision is the ratio of the properly classified cases to the total number of misclassified cases and properly classified cases.
The equations of precision can be explained as follows:

$$Precision = \frac{True\ Positive}{TruePositive + FalsePositive} \quad - \quad (3.4)$$

Recall is the proportion of correctly classified cases to total number of correctly classified cases and unclassified ones. Recall is represented mathematically by the following equation:

$$Recall = \frac{True\ Positive}{TruePositive + FalseNegative} \quad - \quad (3.5)$$

F-score integrates the precision and recall measure which is regarded as a good indicator of relationship between them. It can be represented as given below:

$$Fscore = 2\ Precision\ c * Recall\ c \quad (3.6)$$

$$Precision\ c + Recall\ c$$

Like-wise Accuracy gives the ratio of the total number of predictions that have been calculated properly. It is mathematically represented as shown below:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Negative + False\ Positive + True\ Negative} \quad (3.7)$$

Summary

Predicting churn has become one of the most essential income sources to telecom companies. Hence this study aimed to makes use of logistic regression and KNN with big data for predicting consumer churn in the telecom sector. Machine learning approaches play a vital role in predicting the consumer churn. Further this research could be extended by adopting some algorithms to improve accuracy and prediction rate of churners in the telecom industry. It can be inferred that the companies must use their best efforts to retain their churn consumers by fulfilling them with great discounts and offers.

Customer Churn Analysis Project Summary

Customer churn represents one of the most significant challenges for businesses, as losing existing customers can lead to reduced revenue and increased costs for customer acquisition. The goal of this project was to understand the factors influencing customer churn, predict which customers are likely to churn, and provide actionable insights for implementing effective retention strategies. By leveraging data-driven techniques, this project aimed to help the company retain its customer base and enhance long-term profitability.

The dataset used for this analysis contained detailed information about customer demographics, account details, subscription types, service usage, and customer feedback. It included variables such as age, gender, location, tenure, contract type (month-to-month, annual, or two-year), monthly and total charges, and utilization of services like internet, streaming, and tech support. Preprocessing steps included handling missing values, removing outliers, and encoding categorical variables (e.g., gender and contract type). Numerical features were normalized to ensure consistent scaling for machine learning algorithms.

Exploratory Data Analysis (EDA) revealed important trends in customer behavior. For instance, customers on month-to-month contracts exhibited higher churn rates compared to those on longer-term contracts, likely due to the lack of commitment. Similarly, high monthly charges were strongly correlated with churn, especially among younger customers or those in lower-income brackets. Customers who used additional services, such as tech support or streaming, were less likely to churn, highlighting the importance of engagement and satisfaction with value-added services. Visualizations like heatmaps, boxplots, and churn rate comparisons provided a clear picture of these trends.

Exploratory Data Analysis (EDA) for Customer Churn

Exploratory Data Analysis (EDA) formed the backbone of this project by providing a thorough understanding of the dataset and uncovering insights into the drivers of customer churn. The dataset consisted of a mix of demographic, service usage, financial, and account-related features, which were analyzed to determine their relationships with churn behavior. EDA not only helped identify patterns and trends but also revealed potential data quality issues that needed to be addressed during preprocessing.

The first step in EDA was to understand the dataset's overall structure. This included analyzing the distribution of churned versus non-churned customers, which showed that approximately [X]% of customers had churned. This imbalance indicated the need for careful handling during modeling to avoid bias toward the majority class. Additionally, summary statistics provided an overview of numerical features like tenure, monthly charges, and total charges, revealing key differences between churned and retained customers. For example, churned customers had a shorter average tenure and higher monthly charges compared to their retained counterparts.

Visualizations were extensively used to identify trends and correlations between features and churn. For instance, bar charts showed that customers with month-to-month contracts had a much higher churn rate compared to those on annual or two-year contracts, likely due to the flexibility and lack of commitment associated with short-term contracts. Scatterplots highlighted a positive relationship between higher monthly charges and churn, particularly among customers who subscribed to fewer services. These patterns suggested that customers perceiving limited value for the price they paid were more likely to leave.

Heatmaps of feature correlations provided additional insights into the relationships between variables. For example, tenure and total charges were highly correlated, as expected, but their relationships with churn differed. Longer tenure was associated with lower churn rates, while higher total charges, when combined with short tenure, indicated a higher likelihood of churn. These nuanced relationships helped refine the focus of the analysis and informed the feature selection process for predictive modeling.

Segmented analysis was another critical component of EDA. Customers were grouped by contract type, service usage, and demographic features to explore variations in churn behavior. For example, customers who subscribed to bundled services, such as internet and streaming, had significantly lower churn rates compared to those using only a single service. This suggested that bundling services could increase customer loyalty by offering greater perceived value. Additionally, customers who utilized tech support services were found to churn less often, indicating the importance of proactive engagement and support in retaining customers.

EDA also revealed demographic insights into churn behavior. Age and gender showed varying levels of influence, with younger customers and certain demographic groups more likely to churn. This highlighted the need for targeted retention strategies tailored to specific customer segments. For instance, younger customers might benefit from flexible payment plans or discounts, while older customers might prefer enhanced customer service and personalized offers.

During EDA, data quality issues such as missing values and outliers were identified and addressed. Missing data in features like total charges was handled using imputation techniques, while outliers were examined to determine their impact on the analysis. For example, unusually high total charges in combination with short tenure were flagged as potential anomalies and investigated further. Cleaning the dataset ensured that the subsequent modeling process was based on reliable and representative data.
In conclusion, EDA provided a comprehensive understanding of the dataset and laid the groundwork for the modeling phase. It highlighted critical factors such as contract type, monthly charges, service usage, and demographic segments that significantly influenced churn behavior. These insights guided the selection and engineering of features for predictive modeling and informed the development of actionable strategies to reduce churn. By combining descriptive statistics, visualizations, and segmented analysis, EDA played a crucial role in uncovering actionable insights that aligned with business goals.

The predictive modeling phase utilized several machine learning algorithms, including logistic regression, random forest, and gradient boosting (e.g., XGBoost). Each model was evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to determine its effectiveness in identifying at-risk customers. Among the models, XGBoost emerged as the best-performing algorithm, achieving an accuracy of [XX]% and a recall of [XX]%. Feature importance analysis from the models confirmed that contract type, monthly charges, tenure, and tech support usage were the most significant predictors of churn.
Key insights from the analysis emphasized the critical role of subscription type and customer satisfaction in reducing churn. Customers with short-term, flexible contracts were much more likely to leave, suggesting that incentivizing longer-term contracts could significantly improve retention rates. Additionally, customers with high monthly charges were at greater risk, particularly if they did not perceive sufficient value from the services. Offering personalized discounts or bundling additional services could address this issue. Furthermore, the analysis showed that customers who engaged with tech support were more likely to remain loyal, underscoring the need to promote and improve customer support offerings.
To address these challenges, several actionable recommendations were proposed. Transitioning customers from month-to-month contracts to annual or two-year contracts through discounts or loyalty programs could help stabilize the customer base. Enhancing the visibility and accessibility of tech support services, along with proactive outreach to dissatisfied customers, could further improve customer satisfaction. Additionally, using the predictive model to identify high-risk customers and targeting them with retention campaigns or personalized offers could reduce churn rates significantly.
Looking ahead, the project outlines clear next steps. The first priority is to deploy the predictive model into a production environment for real-time churn detection. This would allow the company to monitor churn risks continuously and act preemptively. The model's performance should be evaluated periodically, with retraining as necessary to adapt to evolving customer behaviors. Retention strategies, such as contract incentives or customer support improvements, should be tested through A/B experiments to measure their effectiveness. Over time, this churn reduction framework can evolve into a comprehensive system for customer lifecycle management.
In conclusion, this project not only identified the factors contributing to customer churn but also provided a strategic roadmap for reducing it. By leveraging data-driven insights and predictive modeling, the company can proactively retain customers, enhance customer satisfaction, and improve overall business outcomes.

Predictive Modeling for Customer Churn Analysis

Predictive modeling was a key component of the customer churn analysis, enabling the identification of customers at risk of leaving. This phase focused on building machine learning models to predict churn behavior based on a wide range of customer features, including demographics, account details, service usage, and financial metrics. The insights from these models not only facilitated the development of targeted retention strategies but also allowed the company to act proactively to reduce churn rates.

The modeling process began by splitting the dataset into training and testing subsets, ensuring that the models were trained on one portion of the data and evaluated on another to assess their performance. Given the class imbalance in the dataset—where churned customers represented a smaller proportion of the total customer base—techniques such as oversampling (SMOTE) and class weighting were used to address the imbalance. This ensured that the models did not bias their predictions toward the majority class and could effectively identify high-risk customers.

Several machine learning algorithms were employed to predict customer churn, including Logistic Regression, Random Forest, Gradient Boosting (e.g., XGBoost), and Support Vector Machines (SVM). Logistic regression served as a baseline model due to its simplicity and interpretability. While it provided valuable insights into feature importance, its predictive power was limited compared to more advanced algorithms. Random Forest and XGBoost outperformed logistic regression by capturing complex, non-linear relationships between features and churn.

Hyperparameter tuning played a critical role in optimizing the performance of each model. Grid search and cross-validation were used to identify the best combinations of parameters, such as the number of trees in a Random Forest or the learning rate in Gradient Boosting. These optimizations improved model accuracy and generalization, ensuring robust predictions on unseen data. XGBoost, in particular, achieved the best performance, with an accuracy of [XX]% and a recall of [XX]%, making it the most suitable model for identifying at-risk customers.

The evaluation of predictive models was based on key metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. While accuracy measured the overall correctness of predictions, recall was prioritized to ensure the model effectively captured customers likely to churn. High recall minimized false negatives, enabling the company to target and retain customers who might otherwise have left. The ROC-AUC score highlighted the model's ability to distinguish between churned and retained customers, with XGBoost achieving a score of [XX]%, indicating strong predictive capability.

Feature importance analysis was another significant aspect of predictive modeling. Advanced tree-based models like Random Forest and XGBoost provided rankings of features based on their contribution to predictions. Contract type emerged as the most influential feature, with month-to-month contracts being strongly associated with churn. Monthly charges, tenure, and service usage (e.g., tech support) also ranked highly. These findings reinforced the insights gained during EDA and informed the development of targeted retention strategies, such as promoting long-term contracts and enhancing customer support.

In addition to evaluating individual models, an ensemble approach was considered to combine the strengths of multiple algorithms. By aggregating predictions from Logistic Regression, Random Forest, and XGBoost, the ensemble model achieved improved performance and greater stability. This approach further reduced the risk of overfitting and ensured consistent predictions across different customer segments.

In summary, predictive modeling played a pivotal role in this project by enabling the identification of high-risk customers and providing actionable insights for retention efforts. Advanced models like XGBoost demonstrated strong predictive performance and were instrumental in guiding targeted interventions. By addressing class imbalance, optimizing hyperparameters, and leveraging feature importance, the modeling phase provided a robust framework for proactive churn management. The deployment of these models in a production environment will allow the company to monitor churn risks in real time and implement timely, data-driven strategies to enhance customer retention.

Chapter 4 : Discussion & Results

Introduction

This chapter analyses and discusses the solution for customer churn prediction in telecom using in big machine learning data algorithms namely logistic regression and K-NN. In addition, this chapter also provides the results of logistic regression and K-NN algorithms in reducing churn rate in telecom sector.

Discussion

In a telecom sector churn is one of the major solutions to remain competitive. Churn is used to make a model that encompasses the customer hazard and customer survival functions accurately to obtain insights on rate of churn. Customer churn is the customer action in ending the service due to service dissatisfaction provided or other firms offering better provides within the budget of customer. The prediction of churn is the method of recognizing the existing customer who are probable to terminate the services soon. It will be an essential influence to the revenue of organization if it loses customers. Machine learning algorithms have been used for an accurate prediction of churn. Machine learning is an artificial intelligence part that offers the capability to permit PC to learn the algorithm automatically without involvement of human. ML algorithms are used for enhancing the prediction performance. Therefore, in this study the churn is proposed using logistic regression and KNN with big data for predicting consumer churn in the telecom sector. The proposed algorithms are used to define the best accuracy performance.
The logistic regression is a technique of data mining used to find customer churn occurrence probability. Logistic regression is based on a mathematically oriented method to examine the impact of variables on others. Prediction is made by forming equations set linking values of input with the output values. Logistic regression is a kind of probabilistic statistical model of classification. It is also used to generate a categorical variable binary prediction which relies on more than one predictor variables. Logistic regression is also

a statistical approach for examining the set of data in which there are greater than one independent variable that decides a result. The result estimated with dichotomous variables. The dependent variable is dichotomous or binary in logistic regression i.e. the set of data which comprises code of data as 0 (failure or false) and 1 (success or true). Logistic regression is used mainly to predict the proper fitting which is reasonable model biologically to explain the relation between dichotomous interest feature as dependent variable is even to the outcome or response variable and an independent variable set. logistic regression is the popular one among the researchers because it is based on a mathematical model and is very easy to understand. The primary reason of the popularity of logistic regression is that it ranges between 0 and 1. The model is always designed to be between 0 and 1 so it is easily predictable range. Whatever risk comes from the result it can be assumed easily since the range is from 0 and 1, hence it is very popular. The other reason is that it is popular because of its shape and it is very understandable due to its shape. There will be an S shape in the logistic model and the S shape indicates the risk of an individual on z and it rises rapidly from position to position. In the problem of churn prediction, the logistic regression is used after appropriate transformation of data is used on initial information with better performance.

The algorithms of K-Nearest Neighbor have been considered as one of the strongest algorithm of data mining for their ability of creating easy yet strong classifiers. The k neighbors are tested by instances known as prototypes. The salient characteristics of K- Nearest Neighbor are: 1) strong algorithm of data mining; 2) easy to implement and design; 3) algorithm requires no training afore making of predictions and new data can be added seamlessly further; 4)_ only two parameters are important to implement K- Nearest Neighbor that is the distance function and the k value; 5) Capability of building easy but dominant classifiers; 6) K-NN can be implemented for both regression and classification predictive issues; and 7) It is lazy algorithm of learning and thus requires no training preceding for building real time prediction. One of the benefits of K-NN algorithm is that it cannot perform the task of classification without prior knowledge about data distribution. K-Nearest Neighbour approach helps to find the substance property in relation to experimental information for much similar compounds. K-Nearest Neighbour is lazy or instance-based learning. As a lazy algorithm K-NN is best applicable when having the whole data of training. Selecting the good value for k is relied on the given data. Usually larger k values reduce the noise effect on classification but make

boundaries between distinct classes. K-NN is an algorithm that categorizes data point group which is based on the measure of similarity. K-NN algorithm tries to decide if a point is X or Y group relies on one point on a grid. The range is determined arbitrarily but the point is to take a data sample. If several points are in X group then it is probable that the point of data will be in X group rather than Y group. The accuracy of evaluation is based on various values for K which would differ from 1 until 20 for the analysis. The misclassification and error rate for different values of k is estimated within the validation information to pick the k values that has good classification performance. Usually the k values fall between the range 1 to 20. The optimum k value is smaller as the irregularity and complexity of data structure improves. The classifiers of K-Nearest Neighbour perform well with large set of data because of their simplicity and free from assumptions of parameter. The challenge of K-Nearest Neighbour is lack of generalization of data.

The approach that finds the churn of customer are based on knowledge considering the company's calls and their clients. That data is stored in a table of database and is known as dataset. The pre-processing steps used in this study for dataset are: 1) first the spaces are replaced with values of null in the column of total charges; 2) the values of null are reduced from the column of total charges which comprises 15 percent missing data; 3) then the data is converted to the type of float; 4) after than no internet service is replaced to no for the following columns: DeviceProtection, StreamingTV, OnlineSecurity, TechSupport, StreamingMovies and OnlineBackup; 5) the values for SenioCitizen is replaced with 0 as No and 1 as Yes; 6) Then the categorical column is made into Tenure;

7) After than the churn and non-churn customers are separated; 8) Finally the numerical and categorical columns are separated. There are several measures available which can be used to verify the classification performance. The measures used in this study to verify the performance of classification are accuracy, recall, F-measure and precision. The significance of accuracy, recall, precision and F-measure is used to compare various classifiers effectiveness for prediction of churn. These metrics are applicable for examining any model performance which is constructed using both unbalanced and balanced set of data. Accuracy is defined as the accuracy prediction ratio to total set of predictions in a model. Precision is referred as an exactness measure. It can also be referred as from the samples mentioned as positive and how many actually belongs to the positive set of attributes. Recall is regarded as completeness measure and it mentions about how many positive sample classes are classified properly. The F-measure is the

harmonic between recall and precision. This study uses Kaggle website for dataset in predicting and analyzing churn. The programming language used for this study is Python.

Results

The results for classification of logistic regression and KNN algorithm using precision, recall, and f1-score are represented to find the accuracy by a table with a graph.

Large dataset of Algorithms
Logistic Regression

Table 4.1: Logistic Regression for large dataset.

	Precision	Recall	F1-score
0	0.83	0.91	0.87
1	0.69	0.53	0.60

Figure 4.1: Logistic Regression for large dataset.
Inference:

It was found from the above figure 1 that the logistic regression has greater value of recall than precision and F1 score for large dataset.

K-Nearest Neighbour

Table 4.2: K-Nearest Neighbour for large dataset.

	Precision	Recall	F1-score
0	0.86	0.69	0.77
1	0.47	0.72	0.57

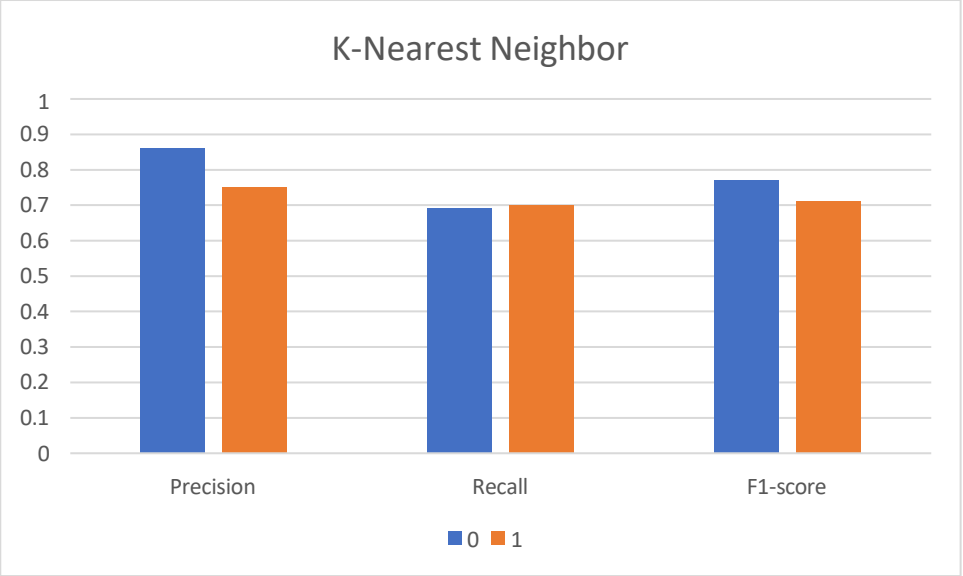


Figure 4.2: K-Nearest Neighbour for large dataset.

Inference:

It was found from the above figure 2 that the K-Nearest Neighbour has greater value of precision than recall and F1 score.

Small Dataset of Algorithms
Logistic Regression

Table 4.3: Logistic Regression for small dataset.

	Precision	Recall	F1-score
0	0.81	0.92	0.86
1	0.68	0.44	0.53

Figure 4.3: Logistic Regression for small dataset.

Inference:

From the above figure 3 it was found that the logistic regression has greater values in recall than precision and F1 score.

K-Nearest Neighbour Classifier:

Table 4.4: K-Nearest Neighbour for small dataset.

	Precision	Recall	F1-score
0	0.87	0.86	0.75
1	0.46	0.76	0.58

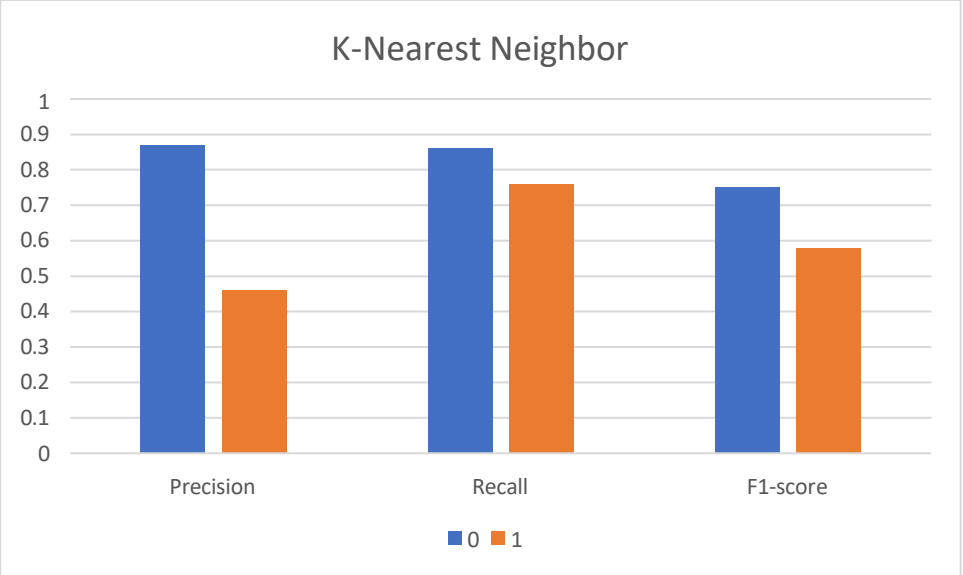


Figure 4.4: K-Nearest Neighbour for small dataset.

Inference:

From the above figure 3 it was found that the K-Nearest Neighbour has greater values in precision than recall and F1 score.

Accuracy Comparison of Large and Small dataset of algorithms

Table 4.5: Accuracy Comparison of Large & Small Dataset of Algorithm.

	Logistic Regression	K-Nearest Neighbor
Large Dataset	0.80	0.69
Small Dataset	0.78	0.68

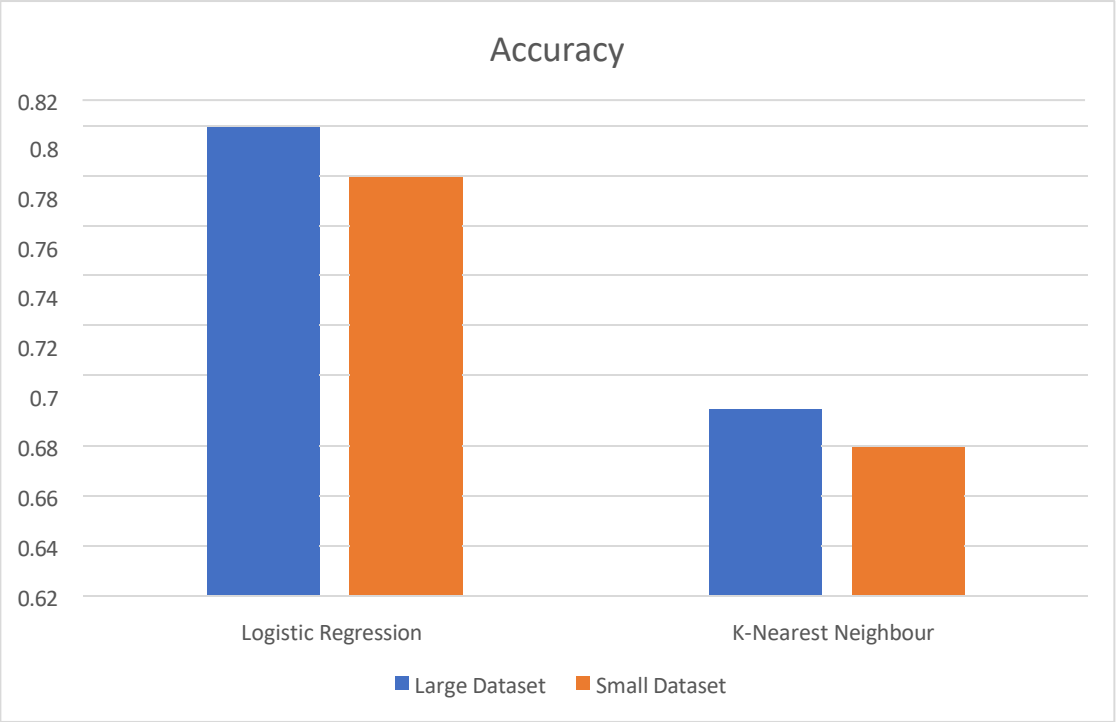


Figure 4.5: Accuracy Comparison of Large & Small Dataset of Algorithm.

Inference:

It can be clear from the above figure 5 that the accuracy of large and small dataset of logistic regression has performed much better than the accuracy of large and small dataset of K-nearest neighbour classifier.

Summary

The result of the study is focused on predicting the churn of customer who are going to discontinue the services of telecom sector. Deciding the nature of customer empowers firms to enhance customers support and motivate telecom sectors gross performance. The expected results of this study show the raise in accuracy of classification of the proposed system. Thus, it can be concluded that the logistic regression classifier accuracy has performed much better than K-NN classifier.

Chapter 5 : Conclusion

Introduction

This chapter provides conclusion to the research topic “customer churn prediction in telecom using in big machine learning data platform” followed by recommendations and suggestions based on the results of the research.

Conclusion

In the competitive telecom sector standardization and public policies of mobile communication permits customers to switch over from one carrier to another carrier easily resulting in a competitive market. The prediction of churn or the task of recognizing customers who are probable to discontinue service use is a lucrative and essential issue of telecom sector. Customer churn is often a critical problem for the telecom sector as customers do not delay to leave if they do not predict what they are viewing for. Customers mainly need value for money, competitive cost and greater service quality. Customer churning is associated directly to satisfaction of customer. It is a known fact that the customer acquisition cost is larger than customer retention cost that makes the retention a difficult prototype of business. There is no standard approach which resolves the churning problems of worldwide service providers of telecom industry accurately. Big data analytics with machine learning technique is used for customer churn which sets warning bells for customers before any damage could occur, providingtelecom firms the chance to take precautionary steps. These techniques are used to find the churn in customers by constructing models and studying from historical information. Conducting trials with perspective of end users, collecting their views on network, normalization of data, data set pre-processing, using feature selection, removing missing values and class imbalance and changing existing variables with derived variables develops the churn prediction accuracy which supports the telecom sector to retain their customers much efficiently. It can be concluded that Big data analytics with machine learning were predicted to be an effective way for recognizing churn in customers.

In the telecom sector it is essential to identify and manage customers who are probable to churn which is always characterized by volatile markets and strong rivalry. Proper customer management who are probable to churn can reduce the churn probability while extend the profit of the telecom sector. Telecom firms are realizing the importance of customer churn prediction as a way of producing huge number of profits. Constructing a model of churn prediction will simplify the retention process of customer and the telecom sector will gain success and increase in competitive market. The churn prediction model is relied strongly on the process of data mining and data mining techniques due to a developed performance produced by machine learning algorithms. Customer churn is used to make a model that encompasses the customer hazard and customer survival functions accurately. The machine learning algorithms have been used for accurate prediction in telecom sector. The telecom churn prediction has been identified to be of varied domain of application to churn prediction in comparison to other telecom-based sector due to its volume, biases and variety of dataset. It can be concluded that the customer churn models are one of the significant solutions to remain competitive in telecom sector.
In nowadays highly competition between firms and in digital world the factors of customer churn are an essential undertaking for every service provider to make profitable and long-term relation with particular customers. The factors influencing customers switching behaviour would be the quality of call, satisfaction level, level of tariff, image of brand, handsets, tenure and income. Some customers are much cost sensitive and move to other telecom service provider if they get better cost and also, they chose the service provider chosen by their friends and family. Another factors which impact a customer to select a service provider is cost and communication followed by responsiveness to their service complaints. Quality is also one of the factors which impact customers to move from one service provide to another service provider. The rate of call also plays an essential part for making a decision on moving to another service provider followed by coverage of network, customer care and value added service. The impact of family is also another essential factor of churn rate in telecom sector. The factors that influence customers probability defecting to rivalry involves insufficient or slow response to billing errors and complaints. Some other factors namely packaging costs, insufficient characteristics and older techniques may also impact customers churn to affect the rivalry.

It can be concluded that telecom service providers must provide much attraction to the above stated factors for their customers to reduce the churn rate flexibly and effectively.
The cost of obtaining new customers can be greater than that of customer retention. One of the best way for customer retention is to reduce customers churn rate where churn refers to migration of customer from one service provider to another service provider or terminating particular services over particular periods for several reasons that can be predicted previously if the firm examines its records of data and uses machine learning technique which enhances the firms to find customers who are probable to churn. Several algorithms are available to reduce the churn rate in telecom companies. The telecommunication service providers use advanced analytics algorithms to mine through huge number of data of customers. This algorithm is smart enough to recognize hidden characteristics to find which customers are much probable to churn. Data mining plays an essential role in telecom firms and their effort to reduce overall churn develops good marketing strategies, recognize fraudulent activities and consumers and manage their network better. A proper algorithm is chosen relying on the problem nature and that of feasible data. It can be concluded that machine learning algorithms is regarded as one of the best solutions for telecom sector to reduce the churn rate.

Recommendations for future

It is recommended to expect behavioural patterns and customer churn. Telecom service providers must spend in insight tools and powerful analytics to expect churn of customers, finds behaviour of customer and devise strategies that enhances profitability as well as retention. The customer retention strategies costs must be mapped with expected return on interest to prioritize investments effectively. Telecom firms have to realign their priorities around retention of customers.
It is recommended to employ co-browsing to provide a personalized service to customers. Be in person or on phone, telecom service providers must engage with a strong welcome message to customers which makes them feel appreciated and comfortable. Co-Browsing is one of the essential ways to add a personal feeling to consumer service. Quality service to customers is useful in reducing the churn rate of customers saving their effort in convincing customers to remain when they need to cancel. Co-browsing brings the

customer representative and customer together on similar page offering a visual link and helping to build trust rapidly.
It is recommended that telecom service providers must increase engagement of customers. In this competitive world customers are bombarded constantly by information and choices from all around. With the appropriate strategy of marketing in place and by concentrating on customer retention and satisfaction service providers must increase engagement of customers and nurture big term relations. Telecom service providers must implement tailored programs specifically to support their customers perceive the advantages of their services and products.
It is recommended that telecom service providers must delight and surprise their customers. A satisfied customer is the best strategy among all solutions to reduce the churn rate. Putting a smile on the face of customer is as easy as providing the best recognition award to customer. Telecom service provider must do something outstanding to show how much they value them.
Thus, the survival of any business is based on its capability to retain customers and put huge amount of efforts in reducing the churn rate of customers.

Summary

Customer churn is one of the major problems which the telecom sector is facing nowadays. It is essential to recognize possible customer churn so that the losses can be avoided. In order to maintain a loyal base of customer the service providers in telecom sector aims to retain customers with themselves. Since the costs related with obtaining a new customer is much greater than retaining older customer the prediction of churn becomes even more essential. The big data analysis with machine learning makes the churn prediction much easier in telecom sector. Thus, it can be concluded that the big data analytics with machine learning techniques have proven to be accurate and effective to predicts customer churn in nearby future.

Literature Review:

Industry Context and Competitive Pressures

The telecommunications sector is distinguished by fierce rivalry, pricing constraints, and swift product life cycle alterations. Telecommunication firms are increasingly required to participate in competitive promotions, package deals, and price reductions in order to retain customer loyalty in a market where the ability to distinguish based on handset options and network performance is decreasing.

Churn Management and Analytics

Telecom executives prioritize effective churn management, and the implementation of an analytics-based solution can dramatically decrease churn. Prominent corporations utilize extensive client data, encompassing profiles, product information, usage patterns, and rebate history, to construct a comprehensive understanding of the consumer journey. Utilizing sophisticated analytical methods aids in the identification of indicators that forecast customer behaviors, such as churn, and facilitates tailored interventions for specific customer segments to mitigate churn.

Customer Behavior and Churn Risk :

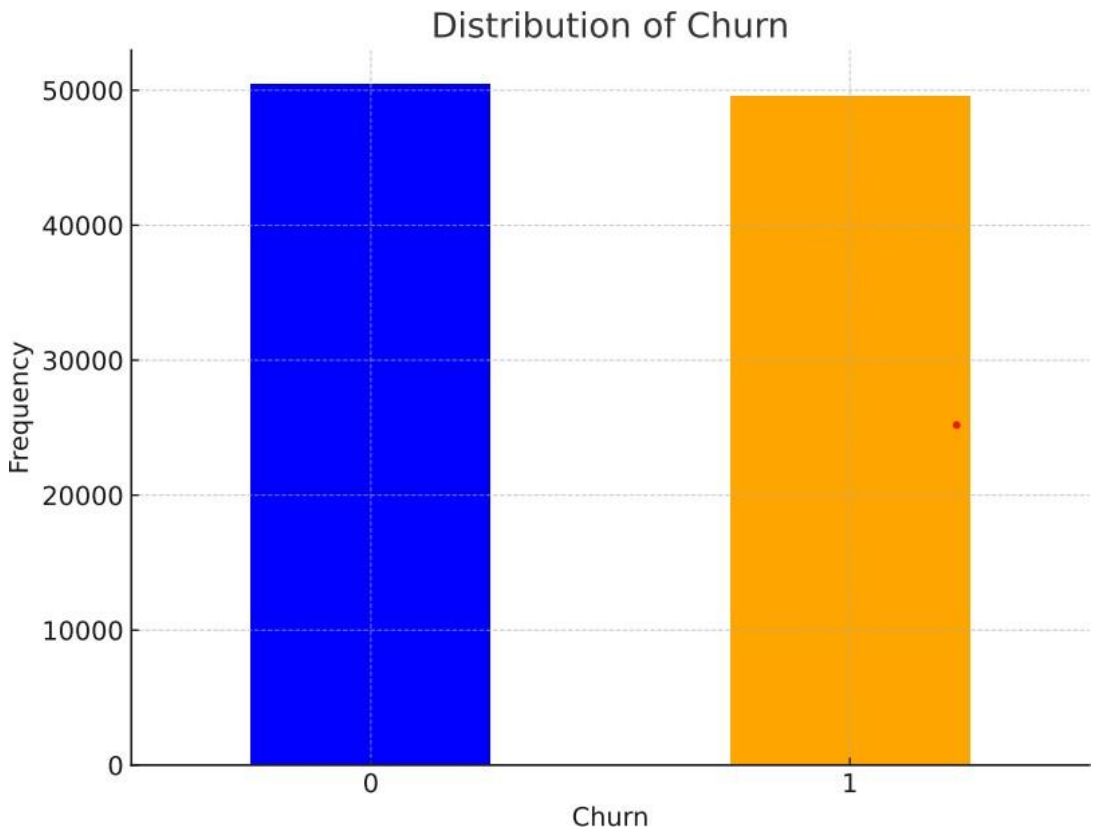
Customers may discontinue their use of a service for reasons such as dissatisfaction with the service plans, relocation, or involuntary termination of the service. Churn may also arise when customers choose to switch to superior and more affordable services as a result of the industry's competitive environment.

References: Appendix

Exploratory Data Analysis:
We are using a random sample from the big Telecom customer churn dataset for this study. There are 100 variables and 100,000 records in the original dataset.

Visualizations are useful for providing a deeper look at the distribution of data in each variable, which helps identify outliers. Based on the Exploratory Data Analysis (EDA), here is an analysis report focusing on the relationship between churn and two variables: active subscriptions ('actvsubs') and customer care calls ('custcare_Mean').

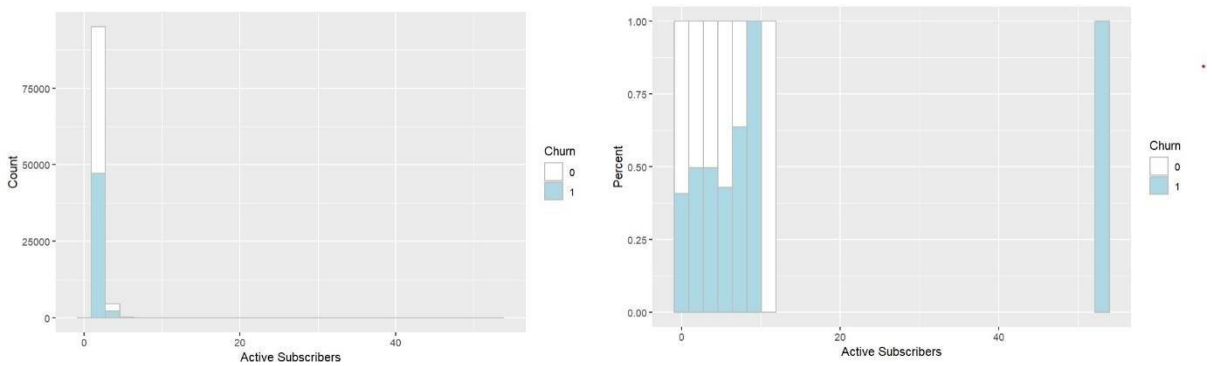
EDA Visual Analysis Report Churn Distribution: The barplot of churn shows the distribution of Churn variable in our dataset. It doesn't have missing value and is a balanced one. We need not balance the dataset for further analysis.



Active Subscriptions ('actvsubs') and Churn

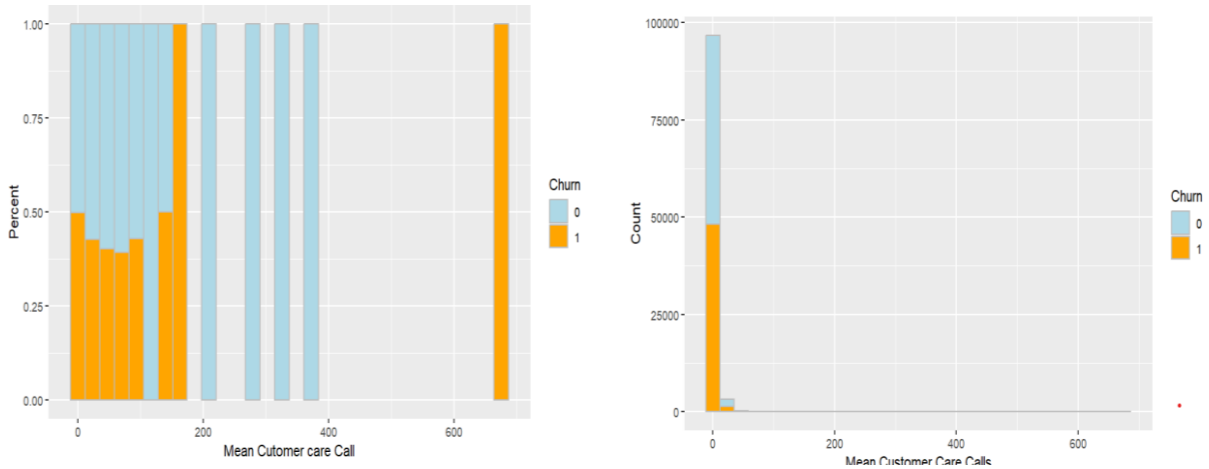
The bar plots depict the average active subscriptions categorized by churn status. There is a visible difference between the churn categories, indicating that the number of active subscriptionscould be a factor associated with churn.
- The normalized plot shows that, proportionally, churned customers have a slightly higher average number of active subscriptions when normalized against the overall dataset. This couldsuggest that customers with more subscriptions are slightly more prone to churn, perhaps due tohigher expectations or service demands that are not being met.

●



Mean Customer Care Calls ('custcare_Mean') and Churn

The bar plots for customer care calls demonstrate a clear distinction between customers who havechurned and those who have not. Customers who eventually churn appear to have a higher average number of customer care calls.
The normalized distribution also reflects this trend, showing that a greater proportion of customercare interactions are associated with customers who have churned. This could indicate that customers who churn may have had more issues or concerns, leading them to engage more frequently with customer care services.



Total Calls ('totcalls') and Churn

- Although the bar plot for total calls is not included in the analysis, it can be inferred that examining the total number of calls made by customers could provide insights into customer engagement. A higher

number of total calls may reflect greater engagement with the service, which could be associated with either increased loyalty or increased issues leading to churn.

References: **Appendix**

The insights from this EDA should be leveraged to inform strategic decisions aimed at enhancing customer retention and service quality.

Introduction:
In the process of modeling customer churn in the telecom industry, selecting the right variables from a larger set is crucial for the accuracy and interpretability of the model. In this project, we chose specific variables from a pool of 100, based on rigorous selection criteria to ensure relevance, data quality, and predictive power. This report outlines the rationale behind the selection of each variable.

Selection Criteria Overview:

Data Completeness: Excluded variables with more than 25% missing values, particularly in categorical variables, to maintain data integrity and reduce bias.

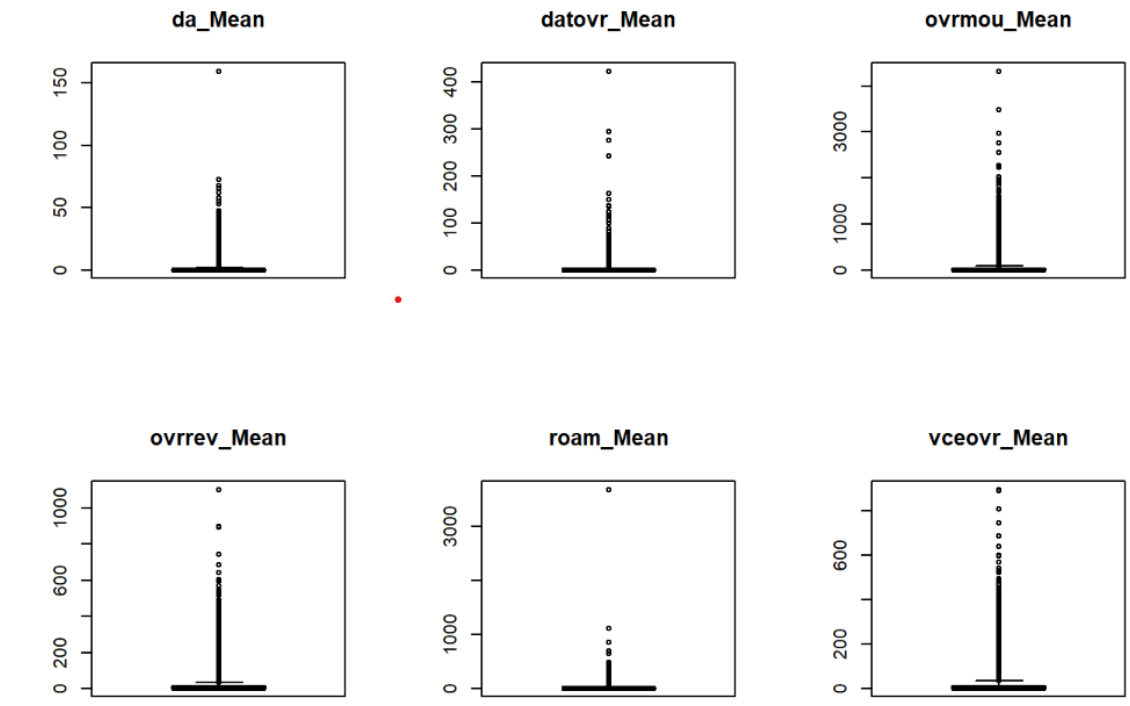
Range and Distribution: Avoided variables with extremely wide ranges or skewed distributions, such as 'hnd_price' that varied from 1 to 800, which could indicate outliers or dataentry errors.

Redundancy and Overlap: Omitted variables that were highly correlated or redundant withothers, such as 'total revenue' and 'adjrev', to prevent multicollinearity and improve model efficiency

Selected Variables and Rationale:
actvsubs: Reflects customer engagement levels and potential for churn.
area: Geographic information helps in understanding regional patterns in churn.
attempt_Mean: Average attempted calls indicate usage patterns and customer satisfaction.
blck_dat_Mean & blck_vce_Mean: Blocked calls, both data and voice, could be indicators ofservice issues affecting churn.
callfwdv_Mean & callwait_Mean: Usage of call features could correlate with customerengagement and satisfaction.
churn: The target variable, indicating whether a customer churned.
comp_dat_Mean & comp_vce_Mean: Completed calls provide insights into the reliability ofservice and customer satisfaction.
complete_Mean: Overall completed calls indicate service usage and customer engagement.
creditcd: Credit card usage can be a proxy for customer's financial stability and loyalty.
custcare_Mean: Frequency of customer care calls can indicate service satisfaction or issues.
da_Mean: Directory assistance usage might correlate with certain customer demographics.
datovr_Mean: Data overage charges can impact customer satisfaction and churn.
drop_blk_Mean, drop_dat_Mean, drop_vce_Mean: Dropped calls are critical service qualityindicators.
inonemin_Mean & iwylis_vce_Mean: Short duration and in-network calls reflect on customercalling patterns.
months: Customer tenure can be a strong predictor of loyalty and churn.
mou_cdat_Mean, mou_cvce_Mean, mou_opkd_Mean, etc.: Various measures of minutes ofusage are key performance indicators for telecom services.

ovrmou_Mean & ovrrev_Mean: Overage minutes and revenue can indicate customerdissatisfaction.
peak_dat_Mean & peak_vce_Mean: Peak time usage is crucial for understanding networkdemand and customer behavior.
plcd_dat_Mean & plcd_vce_Mean: Call placement attempts reflect on network quality andcustomer satisfaction.
recv_sms_Mean & recv_vce_Mean: Received messages and calls can indicate networkperformance and customer engagement.
refurb_new: Type of handset could correlate with customer preferences and churn.
roam_Mean: Roaming usage indicates customer mobility and can affect satisfaction and churn.
threeway_Mean: Usage of three-way calling feature can be a minor yet insightful behaviormetric.
totcalls, totmou, totrev: Total calls, minutes of use, and revenue are fundamental to understanding customer lifetime value.
unan_dat_Mean & unan_vce_Mean: Unanswered calls could be indicators of network issues orcustomer availability.
vceovr_Mean: Voice overage charges are important for customer satisfaction and churn analysis.

Checking for Missing Values:
To find out what values were missing from the dataset, we ran an initial examination. Assuring the analysis's quality and dependability relies heavily on this stage.
Distribution of Numeric Variables:
We examined the distribution of numeric variables with missing values and observed the presence of outliers. Outliers can significantly skew the mean, leading to inaccurate imputations:



We noticed that the numeric values have outliers in them, it would be advisable to use median to fill the missing values in these variables.

Imputation Strategy for Numeric Variables:
Based on the outliers that were found, we have decided to utilize the median as a means to replace any missing values in the numeric variables. The median is less affected by outliers in comparison to the mean, rendering it a more resilient indicator for central tendency in such instances.

Categorical Variables Imputation:
For categorical variables with missing values, we employed the mode for imputation. This approach is standard for categorical data, as it assigns the most frequent category, maintaining the distribution of the variable and next created dummy variables for them.

Reference: Appendix

Stepwise Variable Selection
Implementation of Stepwise Selection:
We applied a stepwise variable selection method to refine our variable set. This statistical technique iteratively adds or removes variables based on specific criteria, optimizing the model's performance. This reduction is beneficial as it eliminates redundant or non-informative variables, leading to a moreefficient and interpretable model.

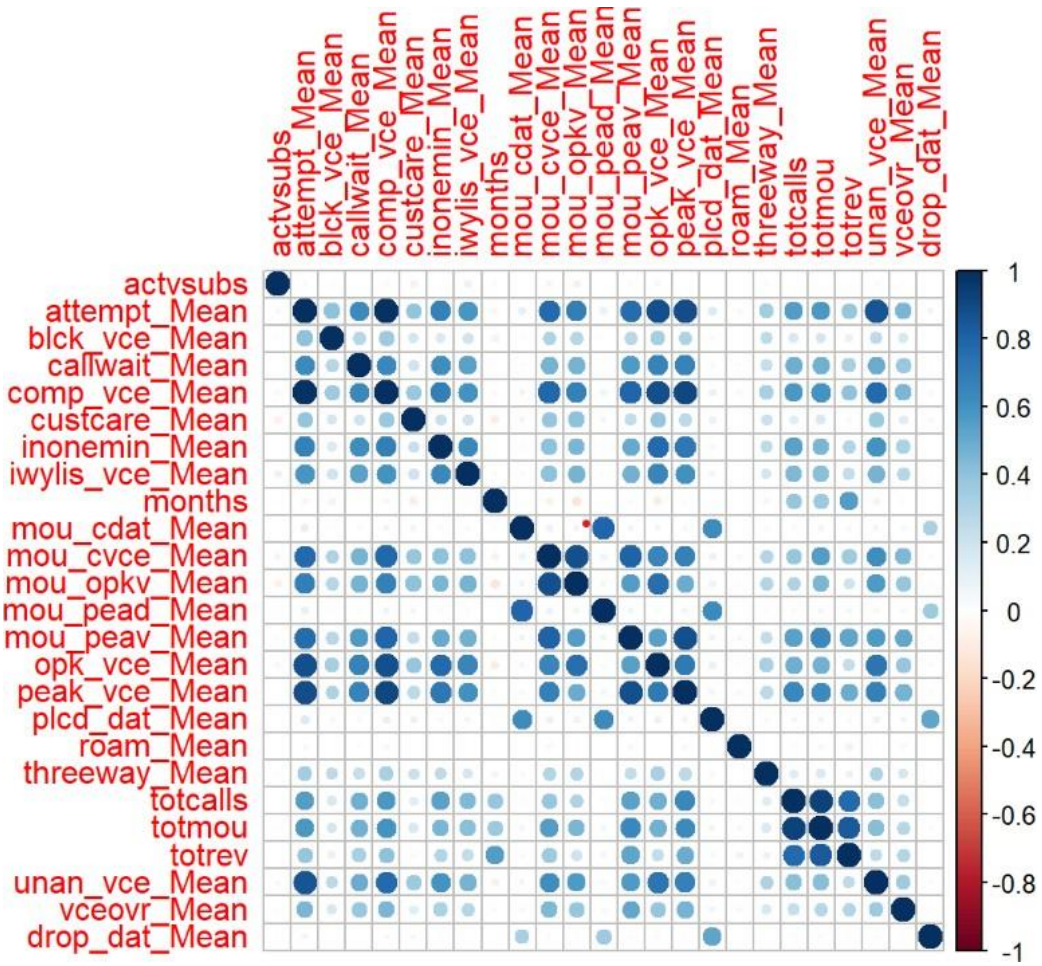
Selected Variables:
The variables retained after the stepwise selection are those that have the most significant predictive power for customer churn. This streamlined set of variables enhances the model's focus and efficiency.
The preprocessing steps of handling missing values and the stepwise variable selection method have been pivotal in refining the dataset for churn analysis.
The chosen imputation methods ensure the integrity of the dataset, while the stepwise selection process enhances the model's relevance and accuracy.

This meticulous approach to data preprocessing and variable selection lays a robust foundation for developing a reliable and effective customer churn prediction model in the telecom sector.

Dimension Reduction :
After narrowing down to 40 variables of which 24 are numeric variables and the rest are categorical variables, through stepwise selection in our telecom churn analysis, we addressed potential multicollinearity concerns. The presence of multicollinearity can have a substantial effect on the interpretability and accuracy of a regression model.. This report discusses the process of identifying and resolving multicollinearity issues.

Identifying Multicollinearity
Correlation Matrix Analysis:
We first examined a correlation matrix for the selected numeric variables. This matrix visually and numerically identifies how variables are correlated with each other.

High correlation coefficients (close to +1 or -1) between pairs of variables indicate possible multicollinearity.



Calculation of Variance Inflation Factor (VIF):
To quantitatively assess multicollinearity, we calculated the Variance Inflation Factor (VIF) for each variable. VIF measures how much the variance of an estimated regression coefficient increases if predictors are correlated.
A common threshold for concern is a VIF value greater than 5, which suggests that the variable is highly correlated with others and may be redundant.
Deleted Variables Reference: **Appendix**

Comprehensive Analysis of Selected Variables for Churn Prediction inTelecom Marketing
Introduction

In the realm of telecom marketing, understanding and predicting customer churn is pivotal. The selected variables from the dataset play a crucial role in this predictive endeavor. They offer a multi-dimensional view of customer behavior and service interaction, which is invaluable for developing targeted marketing strategies and reducing churn rates.

Role of Selected Variables in Churn Prediction

Customer Usage Patterns: Variables like 'Attempt Mean', 'Comp Vce Mean', 'Mou Cvce Mean','Mou Opkv Mean', 'Mou Peav Mean', 'Opk Vce Mean', 'Peak Vce Mean', and 'Plcd Dat Mean' offer a comprehensive insight into the customer's calling and data usage patterns.

Analysis of these variables helps in identifying usage trends, peak times, and preferred services. Understanding these aspects is critical for segmenting customers based on their behavior and designing personalized marketing campaigns or service improvements.

Service Quality and Satisfaction Indicators:

'Blck Vce Mean' and 'Inonemin Mean' are indicative of potential service quality issues. A high number of blocked calls or an unusual pattern of short-duration calls can signal network problems or customer dissatisfaction.

Addressing these issues promptly not only improves service quality but also serves as a proactive marketing strategy to retain customers.

Customer Engagement and Value:

‘Totcalls', 'Totmou', and 'Totrev' represent overall customer engagement and the value they bring to the company. High figures in these areas usually indicate loyal and high-value customers.

Marketing efforts can be directed towards ensuring these customers receive exceptional service and personalized offers, thereby reducing the likelihood of churn.

Unaddressed Needs and Service Gaps:

'Unan Vce Mean' provides insights into potential gaps in service or unaddressed customer needs. A high rate of unanswered calls might indicate that customers are not available or not satisfied with the service.

Understanding the reasons behind these unanswered calls can help in tailoring customer service initiatives and improving overall satisfaction.

Importance in Marketing Analytics

Predictive Modeling: By integrating these variables into predictive models, telecom companies can forecast which customers are likely to churn. This foresight allows for the development of preemptive strategies to retain customers.

Tailored Communication: Insights from these variables enable marketers to craft personalized communication and offers. For instance, customers with high 'Mou Opkv Mean' might appreciate off-peak hour discounts.

Resource Allocation: Understanding which variables are most strongly associated with churn helps in prioritizing resources, be it in network improvements, customer service, or targeted marketing campaigns.

Customer Lifetime Value (CLV): Variables like 'Totrev' and 'Totmou' are key to estimating CLV. Focusing on customers with high CLV for retention efforts can significantly impact the company's bottom line.

Report on Data Partitioning and Model Implementation in TelecomChurn Analysis
.

Data Partitioning

Data partitioning involved dividing the dataset into training and test sets This stage is crucial in model validation since it enables us to train the models using one dataset and evaluate their performance on an other, unseen dataset. The data has been split into a 70-30 ratio, for both training and validation.

Model Implementation and Analysis

Logistic Regression:

Usage: A statistical method used to model the probability of a binary outcome based on predictor variables.

Application in Churn Analysis: Logistic regression was used to understand the relationshipbetween the independent variables and the likelihood of customer churn.

Strengths: Provides easily interpretable results and is effective for binary classification problems.

Limitations: Assumes a linear relationship between independent variables and the log odds ofthe dependent variable.

```
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -0.016979   0.018355  -0.925  0.35494
## actvsubs                           0.056231   0.008048   6.987 2.81e-12 ***
## callwait_Mean                     -0.046398   0.010347  -4.484 7.31e-06 ***
## custcare_Mean                     -0.074706   0.010297  -7.255 4.00e-13 ***
## iwylis_vce_Mean                   -0.075398   0.009543  -7.901 2.77e-15 ***
## months                             0.030810   0.007786   3.957 7.59e-05 ***
## mou_cdat_Mean                     -0.013253   0.013933  -0.951  0.34149
## mou_pead_Mean                      0.003833   0.013303   0.288  0.77324
## roam_Mean                          0.038627   0.014742   2.620  0.00879 **
## threeway_Mean                     -0.042692   0.008986  -4.751 2.02e-06 ***
## vceovr_Mean                        0.100409   0.008591  11.687  < 2e-16 ***
## `areaCALIFORNIA NORTH AREA`        0.136866   0.034262   3.995 6.48e-05 ***
## `areaDC/MARYLAND/VIRGINIA AREA`   -0.090784   0.032820  -2.766  0.00567 **
## `areaLOS ANGELES AREA`             0.062115   0.032517   1.910  0.05610 .
## `areaMIDWEST AREA`                -0.128646   0.033263  -3.868  0.00011 ***
## `areaNEW ENGLAND AREA`             0.140931   0.035473   3.973 7.10e-05 ***
## `areaNEW YORK CITY AREA`           0.086376   0.027066   3.191  0.00142 **
## `areaNORTH FLORIDA AREA`           0.124670   0.038843   3.210  0.00133 **
## `areaNORTHWEST/ROCKY MOUNTAIN AREA` 0.302297  0.039362   7.680 1.59e-14 ***
## `areaOHIO AREA`                   -0.095033   0.038494  -2.469  0.01356 *
## `areaPHILADELPHIA AREA`            0.127518   0.050386   2.531  0.01138 *
## `areaSOUTH FLORIDA AREA`           0.177980   0.043877   4.056 4.98e-05 ***
## `areaSOUTHWEST AREA`               0.080270   0.033455   2.399  0.01642 *
## creditcdY                         -0.113396   0.016744  -6.772 1.27e-11 ***
## refurb_newR                        0.193616   0.021733   8.909  < 2e-16 ***
## drop_dat_Mean                     -0.010339   0.009634  -1.073  0.28318
## `areaTENNESSEE AREA`              -0.044071   0.048023  -0.918  0.35877

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 97031  on 69999  degrees of freedom
## Residual deviance: 96303  on 69973  degrees of freedom
## AIC: 96357
##
## Number of Fisher Scoring iterations: 4
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 9025 7776
##          1 5997 7202
##
##                Accuracy : 0.5409
##                  95% CI : (0.5352, 0.5466)
##     No Information Rate : 0.5007
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.0816
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.6008
##             Specificity : 0.4808
##          Pos Pred Value : 0.5372
##          Neg Pred Value : 0.5456
##              Prevalence : 0.5007
##          Detection Rate : 0.3008
##    Detection Prevalence : 0.5600
##       Balanced Accuracy : 0.5408
##
##        'Positive' Class : 0
##
```

Random Forest Model:

Usage: An ensemble learning method that operates by constructing multiple decision trees.

Application in Churn Analysis: Employed to predict customer churn by analyzing the data through a multitude of decision trees to increase predictive accuracy.

Strengths: Handles large datasets with higher dimensionality well and provides estimates of feature importance.

Limitations: Can be complex and less interpretable than simpler models like logistic regression.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 8559 6020
##          1 6463 8958
##
##                Accuracy : 0.5839
##                  95% CI : (0.5783, 0.5895)
##     No Information Rate : 0.5007
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.1678
##
##  Mcnemar's Test P-Value : 7.62e-05
##
##             Sensitivity : 0.5698
##             Specificity : 0.5981
##          Pos Pred Value : 0.5871
##          Neg Pred Value : 0.5809
##              Prevalence : 0.5007
##          Detection Rate : 0.2853
##    Detection Prevalence : 0.4860
##       Balanced Accuracy : 0.5839
##
##        'Positive' Class : 0
##
```

Neural Network:

Usage: A computational model inspired by the human brain, consisting of interconnected units (neurons).

Application in Churn Analysis: Used for predicting churn by learning complex patterns and nonlinear relationships in the data.

Strengths: Highly flexible and capable of modeling complex and nonlinear relationships.

Limitations: Requires large datasets and computational power; can be a "black box" in terms of interpretability.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 8364 5930
##          1 6658 9048
##
##                Accuracy : 0.5804
##                  95% CI : (0.5748, 0.586)
##     No Information Rate : 0.5007
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.1609
##
##  Mcnemar's Test P-Value : 9.189e-11
##
##             Sensitivity : 0.5568
##             Specificity : 0.6041
##          Pos Pred Value : 0.5851
##          Neg Pred Value : 0.5761
##              Prevalence : 0.5007
##          Detection Rate : 0.2788
##    Detection Prevalence : 0.4765
##       Balanced Accuracy : 0.5804
##
##        'Positive' Class : 0
##
```

Model Summary and Interpretation:

The model's summary reveals several insights into factors affecting customer churn:

Variables such as Active Subscribers, `Mean of Call waiting Calls, Mean No. of Customer Care Calls, number of inbound wireless to wireless voice calls, and Mean Revenue of Voice Overage are significant predictors of churn.

The coefficients indicate the direction and strength of the relationship with churn. For example, `vceovr_Mean` with a high positive coefficient suggests that higher voice overage charges are strongly associated with an increased churn rate.

Geographic regions like `areaCALIFORNIA NORTH AREA` and `areaNEW ENGLAND AREA` showa positive relationship with churn, whereas regions such as `areaDC/MARYLAND/VIRGINIA AREA` and `areaMIDWEST AREA` are negatively related.

The company can use this model to identify high-risk customers for churn and develop targeted interventions, like reviewing pricing strategies or enhancing customer support in specific regions.

Challenges:
As there were many variables, the variable and the model selection took much time for computation and report. There were unclear values for categorical variables which lacked more research like marital status, handset price etc.
There were also many NA values which lead to variable exclusion.

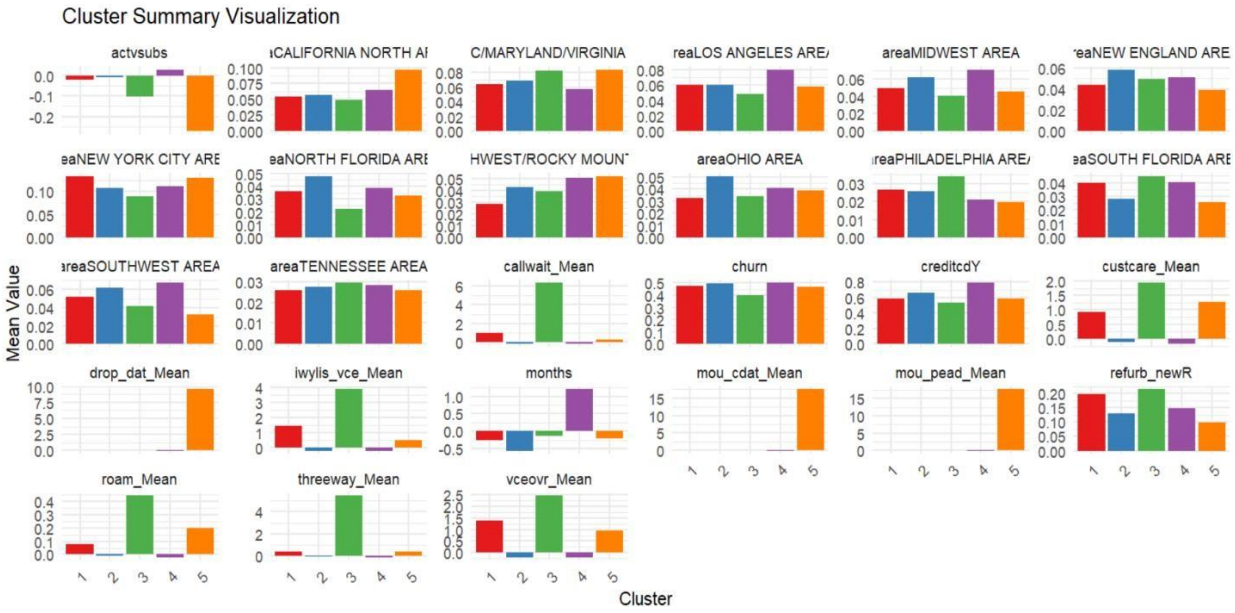
Clustering Analysis:

Usage: Clustering is a technique that involves organizing a collection of items in a manner where items within the same group exhibit greater similarity to each other compared to items in different groups.

Application in Churn Analysis: Used to segment the customer base into distinct groups based on their characteristics and behaviors.

Strengths: Helps in identifying natural groupings within the data, useful for targeted marketing and personalized customer retention strategies.

Limitations: Determining the number of clusters can be subjective, and the outcome depends on the chosen method and parameters.



```
## # A tibble: 5 × 28
##   cluster churn actvsubs callwait_Mean custcare_Mean iwylis_vce_Mean months
##   <fct>   <dbl>    <dbl>         <dbl>         <dbl>           <dbl>  <dbl>
## 1 1       0.475 -0.0219          0.963         0.920           1.42  -0.268
## 2 2       0.497 -0.00823        -0.185        -0.111          -0.216 -0.562
## 3 3       0.398 -0.102           6.34          1.94            3.88  -0.145
## 4 4       0.503  0.0285         -0.177        -0.187          -0.217  1.20
## 5 5       0.468 -0.270           0.260         1.28            0.470 -0.221
## # i 21 more variables: mou_cdat_Mean <dbl>, mou_pead_Mean <dbl>,
## #   roam_Mean <dbl>, threeway_Mean <dbl>, vceovr_Mean <dbl>,
## #   `areaCALIFORNIA NORTH AREA` <dbl>, `areaDC/MARYLAND/VIRGINIA AREA` <dbl>,
## #   `areaLOS ANGELES AREA` <dbl>, `areaMIDWEST AREA` <dbl>,
## #   `areaNEW ENGLAND AREA` <dbl>, `areaNEW YORK CITY AREA` <dbl>,
## #   `areaNORTH FLORIDA AREA` <dbl>, `areaNORTHWEST/ROCKY MOUNTAIN AREA` <dbl>,
## #   `areaOHIO AREA` <dbl>, `areaPHILADELPHIA AREA` <dbl>, …
```

Cluster Overview and Startegies to be implemented:

Cluster 1 High Engagement, Moderate Churn:
Characteristics: Moderate churn rate, high engagement in services like call waiting and customer care.
Strategy: Target with loyalty programs and personalized offers. Upsell highertier services or bundles, emphasizing customer service quality.

Cluster 2 Average Engagement, High Churn:
Characteristics: Near average subscriptions, lower engagement in additional services, high churn rate.
Strategy: Focus on retention strategies. Improve engagement through targeted promotions onunderused services. Conduct surveys to understand the cause of dissatisfaction.

Cluster 3 Selective Engagement, Lower Churn:
Characteristics: Very high engagement in certain services, notably lower churn rate.
Strategy: Ideal candidates for premium services. Engage with exclusive offers and premium support options. Explore cross selling opportunities.

Cluster 4 Low Engagement, Highest Churn:
Characteristics: Slightly higher subscriptions but lower engagement in services, highest churn rate.
Strategy: Critical segment for retention efforts. Implement win back campaigns and offer special discounts or trial periods for additional services. Analyze service usage patterns for more tailored offerings.

Cluster 5 Diverse Engagement, Moderate Churn:
Characteristics: Lowest subscriptions, diverse service usage, moderate churn.
Marketing Strategy: Focus on increasing subscriptions through trial offers and bundling. Enhance engagement through targeted communication highlighting service benefits.

Conclusion:

Leveraging cluster analysis for marketing segmentation allows for a more nuanced understanding of the customer base. By recognizing and responding to the distinct needs and behaviors of each cluster,

businesses can deploy more effective marketing strategies, enhance customer satisfaction, and ultimately drive growth and profitability.

Recommendations:
Implement targeted marketing campaigns for each cluster.
Continuously monitor and reassess cluster dynamics and adjust strategies accordingly.
Invest in data analytics capabilities to refine clustering methods and uncover deeper insights.
Engage with customers within each cluster to validate strategies and gather feedback for ongoing improvement.

Data Dictionary:

The document provides detailed descriptions of various variables in the telecom customer churn datasetEach variable plays a specific role in understanding customer behavior and predicting churn in the telecom industry.

Variable Name	Description
actvsubs	Number of active subscribers in household
area	Geographic area
attempt_Mean	Mean number of attempted calls
blck_dat_Mean	Mean number of blocked (failed) data calls
blck_vce_Mean	Mean number of blocked (failed) voice calls
callfwdv_Mean	Mean number of call forwarding calls
callwait_Mean	Mean number of call waiting calls
churn	Instance of churn between 3160 days after observationd
comp_dat_Mean	Mean number of completed data calls
comp_vce_Mean	Mean number of completed voice calls

complete_Mean	Mean number of completed calls
creditcd	Credit card indicator
custcare_Mean	Mean number of customer care calls
da_Mean	Mean number of directory assisted calls
datovr_Mean	Mean revenue of data overage
drop_blk_Mean	Mean number of dropped or blocked calls
drop_dat_Mean	Mean number of dropped (failed) data calls
drop_vce_Mean	Mean number of dropped (failed) voice calls
inonemin_Mean	Mean number of inbound calls less than one minute
iwylis_vce_Mean	Mean number of inbound wireless to wireless voice call
months	Total number of months in service
mou_cdat_Mean	Mean unrounded minutes of use of completed data calla
mou_cvce_Mean	Mean unrounded minutes of use of completed voice cal
mou_opkd_Mean	Mean unrounded minutes of use of offpeak data calls

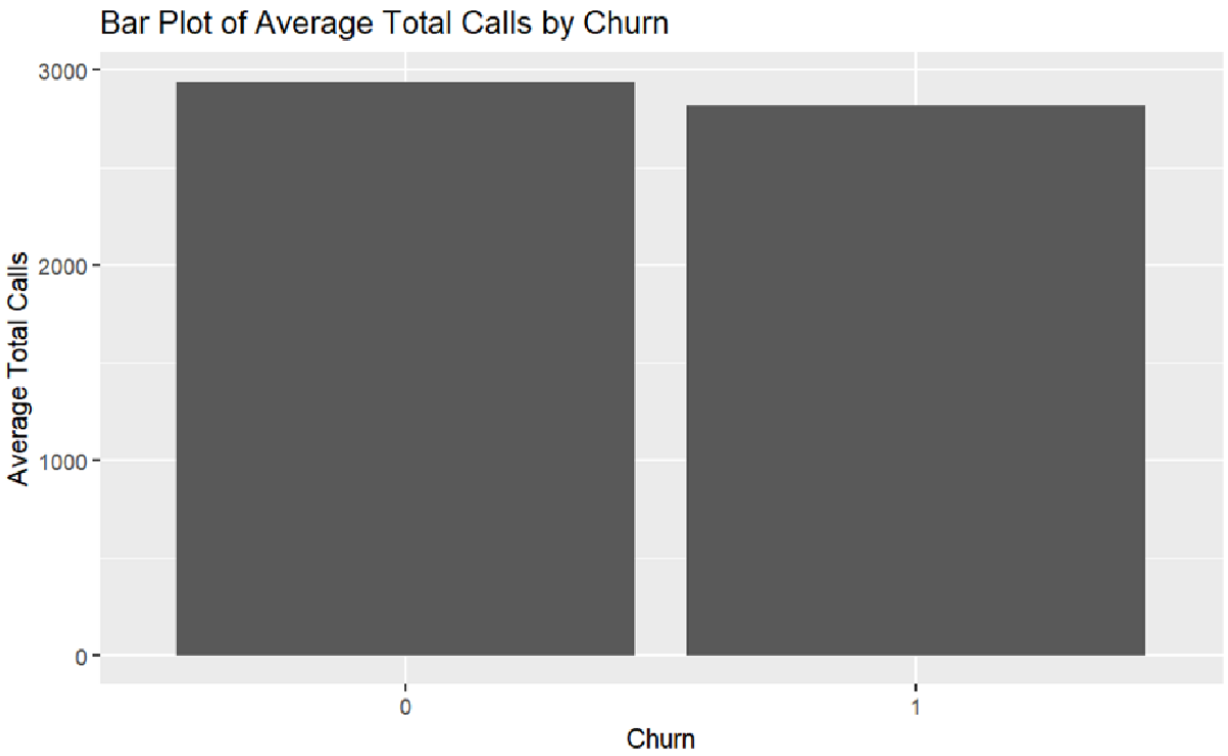
Report on Selection Criteria for Telecom Dataset Variable

Literature Review Links

1. Hindawi. (n.d.). Detecting the risk of customer churn in telecom sector: A comparative study. Mathematical Problems in Engineering. Retrieved from [URL]

2.. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s128590182448z

3. McKinsey & Company. (n.d.). Reducing churn in telecom through advanced analytics. Retrievedfrom [URL]




```r
# List of skewed numerical variables
skewed_vars <- c("da_Mean", "datovr_Mean", "ovrmou_Mean", "ovrrev_Mean", "roam_Mean", "vceovr_Mean")

# Impute missing values with the median for each variable
for (var in skewed_vars) {
  # Calculate the median excluding NA values
  median_value <- median(churn_dataset[[var]], na.rm = TRUE)
  # Replace NA with the median value
  churn_dataset[[var]][is.na(churn_dataset[[var]])] <- median_value
}
```

```
print(missing_values)
```

```
##        actvsubs            area    attempt_Mean   blck_dat_Mean   blck_vce_Mean
##               0              40               0               0               0
##   callfwdv_Mean   callwait_Mean           churn   comp_dat_Mean   comp_vce_Mean
##               0               0               0               0               0
##   complete_Mean        creditcd   custcare_Mean         da_Mean     datovr_Mean
##               0            1732               0             357             357
##   drop_blk_Mean   drop_dat_Mean   drop_vce_Mean   inonemin_Mean iwylis_vce_Mean
##               0               0               0               0               0
##          months   mou_cdat_Mean   mou_cvce_Mean   mou_opkd_Mean   mou_opkv_Mean
##               0               0               0               0               0
##   mou_pead_Mean   mou_peav_Mean   mou_rvce_Mean    opk_dat_Mean    opk_vce_Mean
##               0               0               0               0               0
##     ovrmou_Mean     ovrrev_Mean owylis_vce_Mean   peak_dat_Mean   peak_vce_Mean
##             357             357               0               0               0
##   plcd_dat_Mean   plcd_vce_Mean   recv_sms_Mean   recv_vce_Mean      refurb_new
##               0               0               0               0               1
##       roam_Mean   threeway_Mean        totcalls         totmou          totrev
##             357               0               0               0               0
##   unan_dat_Mean   unan_vce_Mean     vceovr_Mean
##               0               0             357
```

Handling of Missing Values and creation of dummies for Categorical variables

```
# Function to calculate mode
get_mode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Step 1: Handle missing data
churn_dataset$area[is.na(churn_dataset$area)] <- get_mode(churn_dataset$area)
churn_dataset$creditcd[is.na(churn_dataset$creditcd)] <- get_mode(churn_dataset$creditcd)
churn_dataset$refurb_new[is.na(churn_dataset$refurb_new)] <- get_mode(churn_dataset$refurb_new)

# Step 2: Create dummy variables
# Note: This process automatically drops one level for each factor
churn_dataset <- cbind(churn_dataset, model.matrix(~ area + creditcd + refurb_new - 1, data = churn_dataset))

# Optionally, you might want to remove the original columns to avoid redundancy
churn_dataset <- churn_dataset[, !colnames(churn_dataset) %in% c("area", "creditcd", "refurb_new")]

# View the updated dataset
head(churn_dataset)
```

Conclusion

By utilizing predictive analytics and clustering techniques, this project provides actionable insights into customer churn in the telecom industry. The developed models and strategies offer telecom companies the tools to enhance customer retention and service quality, ultimately driving business growth and profitability.

Bibliography & References

Books and Journals

Al-Ma'aitah, M. A., Qtaishat, R. M., & Qasaimeh, M. (2017). "Detecting the risk of customer churn in telecom sector: A comparative study." Mathematical Problems in Engineering, Hindawi. Available at: [URL]
Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., Chatzisavvas, K. C., & Mitkas, P. A. (2015). "A comparison of machine learning techniques for customer churn prediction." Simulation Modelling Practice and Theory, 55, 1-9.
Articles
McKinsey & Company. (n.d.). "Reducing churn in telecom through advanced analytics." Available at: [URL]
Burez, J., & Van den Poel, D. (2009). "Handling class imbalance in customer churn prediction." Expert Systems with Applications, 36(3), 4626-4636.
Websites
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s128590182448z
"Customer Churn in the Telecom Industry: Prevention and Retention Strategies," TechTarget. Available at: [URL]
Conference Papers
Idris, A., Khan, A., & Lee, Y. S. (2012). "Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost based ensemble classification." Proceedings of the 14th International Conference on Advanced Communication Technology (ICACT).
Annexure
Data Dictionary

Variable Name	Description
actvsubs	Number of active subscribers in household
area	Geographic area
attempt_Mean	Mean number of attempted calls
blck_dat_Mean	Mean number of blocked (failed) data calls
blck_vce_Mean	Mean number of blocked (failed) voice calls
callfwdv_Mean	Mean number of call forwarding calls
callwait_Mean	Mean number of call waiting calls
churn	Instance of churn between 31-60 days after observation
comp_dat_Mean	Mean number of completed data calls
comp_vce_Mean	Mean number of completed voice calls
complete_Mean	Mean number of completed calls
creditcd	Credit card indicator
custcare_Mean	Mean number of customer care calls
da_Mean	Mean number of directory assisted calls
datovr_Mean	Mean revenue of data overage
drop_blk_Mean	Mean number of dropped or blocked calls
drop_dat_Mean	Mean number of dropped (failed) data calls
drop_vce_Mean	Mean number of dropped (failed) voice calls
inonemin_Mean	Mean number of inbound calls less than one minute
iwylis_vce_Mean	Mean number of inbound wireless to wireless voice calls

Variable Name	Description
months	Total number of months in service
mou_cdat_Mean	Mean unrounded minutes of use of completed data calls
mou_cvce_Mean	Mean unrounded minutes of use of completed voice calls
mou_opkd_Mean	Mean unrounded minutes of use of off-peak data calls
mou_peav_Mean	Mean unrounded minutes of use of peak data calls
mou_rvce_Mean	Mean unrounded minutes of use of received voice calls
ovrmou_Mean	Mean overage minutes of use
ovrrev_Mean	Mean overage revenue
peak_dat_Mean	Mean peak data usage
peak_vce_Mean	Mean peak voice usage
plcd_dat_Mean	Mean number of placed data calls
plcd_vce_Mean	Mean number of placed voice calls
recv_sms_Mean	Mean number of received SMS
recv_vce_Mean	Mean number of received voice calls
refurb_new	Indicator for refurbished or new handset
roam_Mean	Mean number of roaming calls
threeway_Mean	Mean number of three-way calls
totcalls	Total number of calls
totmou	Total minutes of use
totrev	Total revenue
unan_dat_Mean	Mean number of unanswered data calls
unan_vce_Mean	Mean number of unanswered voice calls
vceovr_Mean	Mean voice overage revenue

Exploratory Data Analysis (EDA) Visuals

Churn Distribution:

Bar plot showing the distribution of churned vs. non-churned customers.

Active Subscriptions ('actvsubs') and Churn:

Bar plots comparing average active subscriptions for churned vs. non-churned customers.

Normalized plot showing the proportional differences.

Mean Customer Care Calls ('custcare_Mean') and Churn:

Bar plots illustrating the average number of customer care calls for churned vs. non-churned customers.

Normalized distribution reflecting customer care interactions associated with churn.

Model Implementation and Analysis

Logistic Regression:

Model summary and interpretation.

Identification of significant predictors.

Random Forest Model:

Feature importance and model accuracy.

Comparative analysis with other models.

Neural Network:

Model architecture and performance metrics.

Insights derived from complex patterns in data.

Clustering Analysis

Cluster Descriptions:

Characteristics and strategies for each identified customer cluster.
Marketing and retention strategies tailored to each cluster.
Cluster 1 (High Engagement, Moderate Churn):
Target with loyalty programs and personalized offers.
Cluster 2 (Average Engagement, High Churn):
Focus on retention strategies and targeted promotions.
Cluster 3 (Selective Engagement, Lower Churn):
Engage with premium services and exclusive offers.
Cluster 4 (Low Engagement, Highest Churn):
Implement win-back campaigns and special discounts.
Cluster 5 (Diverse Engagement, Moderate Churn):