“A Study On Survival Analysis and Churn Prediction Using Machine Learning Model for Selected Telecom Companies”

By

Arun Kumar D R

2021-2022

TABLE OF CONTENT

CHAPTERS	CONTENTS	Page No
1	**Chapter 1: Introduction.** 1.1 Introduction about Topic. 1.2 Theoretical background of the study	**07-20**
2	**Chapter 2:** **Review of Literature and Research Design** 1.1 Review of Literature and Gaps 1.2 Statement of the Problem 1.3 Need and Scope of the Study 1.4 Objectives of the Study 1.5 Sampling 1.6 source of Data Collection 1.7 Tools for Data Collection 1.8 Limitations of Study	**21-28**
3	**Chapter3:** **Profile of the industry**	**29-39**
4	**Chapter 4:** **Data Analysis and Interpretation**	**40 -75**
5	**Chapter 5:** **Findings, Conclusions and Suggestions**	**76 -80**
6	**Bibliography**	**81 -82**
7	**WORK DONE DIARY**	**83**
8	**Certificate of Plagiarism**	**84-89**

LIST OF FIGURES

Figures No.	Figures Title	Page No
Figure 1.1	Main causes of customer churn in the telecom industry	9
Figure 1.2	Multiple churn types	12
Figure 1.3	Flow Diagram of Proposed Work	15
Figure 1.4	Steps involved in data cleaning	16
Figure 1.5	Overview of AI, ML, and DL	18
Figure 1.6	The machine learning process	19
Figure 3.1	Global telecom Segmentation	32
Figure 3.2	Telecoms Key Companies & Market Share Insights	33
Figure 3.3	Global forecast of 5G smartphone subscription by region	37
Figure 4.1	Gender Distribution	47
Figure 4.2	Percentage of senior citizen and younger population	48
Figure 4.3	percentage of customer with dependents and partners	48
Figure 4.4	percentage of customers with both partners and dependents.	49
Figure 4.5	Number of Customers by tenure	50
Figure 4.6	Number of Customers by Contract Type	51
Figure 4.7	Number of customers by Month-to-Month Contract, One Year Contract, Two Year Contract	52
Figure 4.8	The distribution of various services used by customers	54
Figure 4.9	checking for churn rate of customers	55
Figure 4.10	Churn rate by Contract Type of customer	56
Figure 4.11	Churn rate by Seniority citizen	57
Figure 4.12	Distribution of monthly charges by churn	58
Figure 4.13	Distribution of total charges by churn	59
Figure 4.14	Survival Analysis of customer based on there their contract type	60
Figure 4.15	Correlation between Quantitative variables	62
Figure 4.16	Checking again Correlation of "Churn" with other variables	63
Figure 4.17	Feature Selection	64
Figure 4.18	Feature Importance using Random Forest Classifier	65
Figure 4.19	Logistic Regression Confusion Matrix	70
Figure 4.20	Decision Tree Confusion Matrix	72
Figure 4.21	Random Forest Confusion Matrix	74
Figure 4.22	Machine learning model comparison	75

Chapter – 1

Introduction

Introduction

Customers are the base for any business success and that is why firms become aware of the significance of acquiring satisfaction of customers. Customer churn is an essential issue and it is regarded as one of the most essential concerns among firms because of increasing rivalry among firms, increased significance of marketing strategies and customers conscious behaviour in present years. Customer churn practice is essential in competitive and rapidly developing in telecom sector. The process of migrating from one service provider to another telecom service provider occurs due to good services or rates or due to various advantages which the rivalry firm provides customers when signing up. Due to the greater cost related with acquiring new customers the prediction of customer churn has developed as an indispensable part of planning process and strategic decision making in telecom sector. Customer churn is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenues of the companies, especially in the telecom Feld.

Customer survival and churn is a term that refers to the rate at which customers stay and leave the business in future. Churn could be due to various factors, including switching to a competitor, or cancelling their subscription because of poor customer service. Long-term relationships with customers are more effective than trying to attract new customers.

In a highly competitive telecom industry, it is essential to keep an efficient customer retention strategy in place even as telecommunications companies to attract in new customers. Successful customer retention helps create brand loyalty and a positive company reputation in addition to saving money on the expensive process of acquiring new customers. We developed and presented a method to use a survival analysis technique that predicts customer survivals and hazards, which was inspired by real telecom service operator data sets. We want to look at how factors like campaign, tariff, tenure, age, and auto-payment affect survival rates and hazards. Following the determination of hazard ratios and survival experiences for each predictor, the data allow telecom service providers to target the right customers and offer incentives to encourage them to stick with their present provider. Customer retention depends on proactive measures that are sparked by the outcomes of the survival model.

The telecom sector has the highest acquisition cost among other sectors. Due to high acquisition cost, mobile service operators are more focused on retention than customer acquisition. Thus, it is crucial to implement an effective retention strategy (ERS) to retain customers, as retention is far easier; less expensive and more achievable than acquisition. In addition, successfully retained customers acquire more new customers; they provide cash flow and profit and they are less sensitive to new pricing. In comparison with new customers, existing old customers may be more loyal, and churning may take a longer time.

Main causes of customer churn in the telecom industry

Customers leave one telecoms provider for another for six main reasons. Companies in the telecom sector should address the following issues to prevent customer churn.

Figure 1.1 Main causes of customer churn in the telecom industry

1). Service quality

Telco managers could successfully determine the root cause of the problem if a network quality issue is the cause of the churn. For example, a defective cell phone, weak reception, or problems with the communication towers could be the main causes of poor service. Telecommunications companies should monitor and manage both their dropped-call rate and their call completion rate in order to accurately track this issue.

2). Features and content availability

When a customer enrols in a plan that isn't best suited to their needs, it happens in marketing or sales. Most frequently, such a decision results in the contract being terminated. In order to spot an issue with acquisition, it's important to analyze your customer churn rate for particular services or bundles.

3). Lower cost substitutes from competitors

Telecommunications companies are constantly changing and even the smallest change—such as a price increase or the introduction of a new product—can prompt customers to switch to a different service provider. These factors also contribute to customer churn in the telecom business, from the issuance of new telco licences introducing new competitors to increasing taxes paid by consumers.

4). Negative customer service experiences

Customer happiness has always been impacted by customer service, regardless of industry. Since communication services are so important in the telecom sector, this is even more obvious. The idea is to keep your customer service issues to a minimal, from those that reach the call centres to those that result in the real resolution of the customer's concerns.

Strategies for Reducing Churn Rate in Telecom Industry

Customer Relationship Management and Churn Models:

Customer Relationship Management (CRM) is the whole set of processes and systems applied to establish long-lived and profitable relationships with specific customers to support business strategies. In telecommunication companies, retention of customers is one of the key activities of the CRM. The main objective of CRM is to establish a win-win relationship between the customer and the firm. Many companies face the customer-losing problem frequently because of not predicting the customers who are going to churn on time. The primary goal of this study is to give telecom businesses a simple and efficient technique to identify consumers who are likely to leave.

Customer retention became very simple and successful for the organisation that helps businesses save consumers because of the customer churn prediction. For the telecom

company, customer loss means losing consumers to competitors, which will result in a significant loss. Therefore, by using churn prediction, a company may provide customers more effective programmes to keep them from churning and using their services.

Retaining current customers within an organisation is the key objective of customer retention. The causes are as follows:

- Due to overcrowding, there is little to no chance of gaining a new customer in the telecommunications industry.
- For a company, acquiring a new customer is expensive for a variety of reasons.

The cost of acquiring a new customer is ten times more than the cost of keeping an existing one. The survival analysis and churn is defined as the inclination of a customer to stay and leave a services provider. We can identify expected churning customers using survival analysis and churn prediction, which allows businesses to target only those customers rather than all customers and offer incentives to keep them as customers.

WHY CHURN RATE IS IMPORTANT?

Churn rates are frequently used to assess a company's potential for general growth as well as the strength of its customer service department. A company is, or will be, in a better or stronger competitive situation if its churn rates are lower. Companies suffer a great deal from customer loss because they frequently invest a lot of money to attract customers. Every online business has a significant additional potential revenue source when it can predict that a certain customer is at a high risk of churning while there is still time to take action.

The impact of the churn rate is clear, so we need strategies to reduce it. Predicting churn is a good way to create proactive marketing campaigns targeted at the customers that are about to churn.

Big data makes it possible for machine learning to predict consumer churn. Churn may be recognised and predicted with the help of machine learning and data analysis. Churn prediction also involves:

- Identifying customers that are at risk,
- determining the problems that customers face,
- Coming up with techniques and strategies to keep customers and reduce churn.

Types of Churners

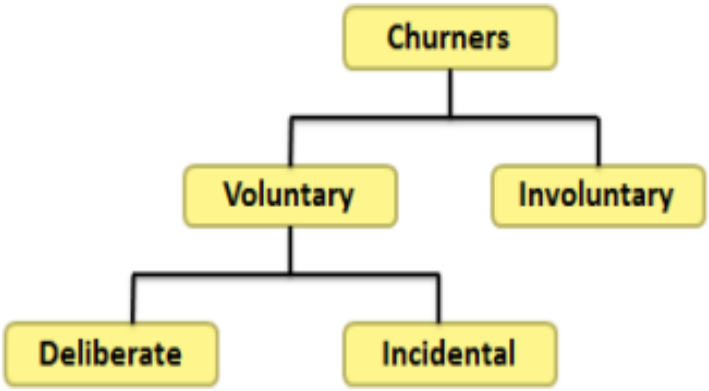

Figure 1.2: Multiple churn types

Churners are classified into two main categories that are

1) voluntary Churners

when a customer decides to cancel a subscription. Voluntary churn can occur for a variety of reasons, such as a shift in the product's perceived value or a bad customer relationship. The voluntary churners are difficult to find due to those customers want to terminate his services from his services providers. The deliberate and inadvertent churners are a further division of the voluntary churners.

a) Deliberate Churners: Customers that need to switch to a different technology or pricing range are the ones who voluntarily churn. The last one is the most significant sort of churners, and most businesses concentrate on it.
b) Incidental Churners : The incidental churners appear as a result of an incident since they have no intention of leaving, but this one has done so for various reasons, such as a change in location or financial circumstance.

A consumer may have made their choice for a variety of reasons, however the following are the most frequent ones:

- The product or service did not live up to their expectations or resolve the issues they had hoped it would.

- The consumer looked at other options after having a negative experience with your product or service.
- The competitor attracted their interest by offering a more appealing alternate solution that fits their requirements or price range.
- The customer no longer requires your service since they are discontinuing operations or going out of business

It's important to truly understand your customers' engagement and satisfaction with the product in order to prevent and reduce voluntary churn. Making your product and customer service indispensable and providing frequent value that they can't live without is a tested way to plug the revenue leak due to voluntary churn.

2) involuntary Churners

Involuntary churners are customers who are removed from the customer list because they don't pay their bills, commit fraud, or don't want to use the phone. refers to the cancellation of a subscription without the customer's knowledge. A failed payment could be the root of the problem. This is one form of churn that can be readily reduced because up to half of all initially unsuccessful payments are successful on a subsequent attempt. These are the main causes of involuntary churn:

- Card expiration
- Hard payment declines that prevent fraudulent attempts when the card is lost or stolen
- Soft payment declines when a credit card is used beyond its maximum capacity
- Network failures

Businesses attempt to prevent involuntary churn. By optimising their checkout page and using clever dunning or payment recovery tactics to successfully connect with customers to recover payments.

Importance of machine learning in customer survival analysis and churn prediction

The telecommunications industry across the world is becoming one of the major sectors. Because of the globalization and developments of tele-com industry and tremendously increased the number of operators in the market that escalates the competition. The customer churn prediction is one of the challenging problems in the tele-com industry. The capability to predict customer survival and churn has significantly increased because of advancement in machine learning and artificial intelligence.

Many churn prediction models have been introduced over the course of many years utilising various methodologies. For churn prediction, the use of machine learning has become increasingly popular. Machine learning uses effective ways to learn from previous observations in order to make accurate prediction. The study of machine learning has grown in significance in the field of telecom churn prediction. To help the telecommunications industry save its customers, researchers have developed very effective experiments for churn prediction. Businesses are enthusiastically creating churn prediction algorithms and making attempts to keep possible churners. Therefore, identifying the causes of churn is essential for creating a better churn prediction model.

The primary goal of the research is to investigate how machine learning models may be used to predict customer churn in the telecom industry. Because of the rise in competitive services, keeping customers in the telecom industry has already become a problem. The primary contribution of this work is the presentation of a churn prediction model that helps telecom businesses in identifying customers who are more likely to leave. Machine learning may be the kind of tool that telecom businesses need to develop a churn prediction model. To create a churn prediction machine learning model telecom data is collected from Kaggle for this study and cleaning data and creating a machine learning algorithm to predict the churn.

Theoretical background of the study

This study shows customer survival analysis and churn prediction in telecom using machine learning. This study makes use of logistic-regression, decision tree and random forest for predicting consumer churn in the telecom sector.

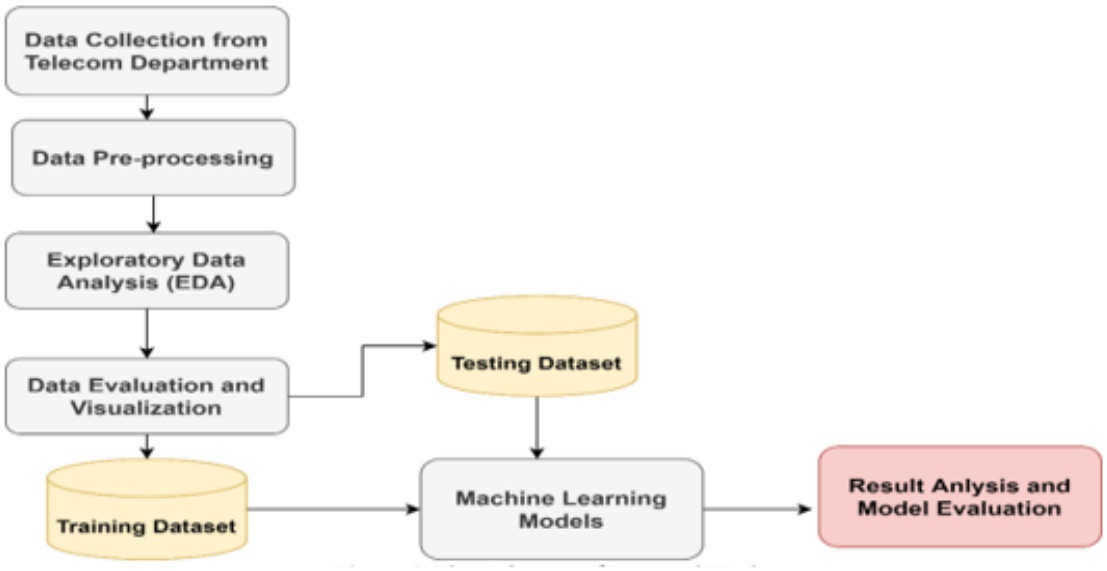

Figure 1.3: Flow Diagram of Proposed Work

Data collection

The data set used in this study is published in the Kaggle website. The data that is feasible for analysis in telecommunication dataset has been used and the prediction has been carried out for the same.

Data Pre-processing

A data set consists of features and N rows. There are many formats used for values. Duplicate values and null values can lead to loss of accuracy in a dataset and dependent values. The pre-processing of data involves 3 steps namely data cleaning, feature selection and data transformation

Exploratory data analysis (EDA)

(EDA) is a powerful and simple tool that can be used to improve your understanding of your data, by exploring its properties. Exploratory data analysis involves examining the important features of a data set, typically using visualisation techniques and summary statistics. The

objective is to understand the data, discover patterns and anomalies, and check assumptions before performing further evaluations.

Data Cleaning

There may be a lot of useless data and gaps in the data. Data cleaning is done to handle this part. It involves dealing with incorrect data, noisy data, missing data, etc. One of the important components of machine learning is data cleaning. It plays a significant part in building a model

Figure 1.4: Steps involved in data cleaning

Removal of unwanted observations

This includes eliminating variables from your dataset that are redundant, duplicate, or irrelevant. The most frequent examples of duplicate observations occur while gathering data, and irrelevant observations are ones that don't actually apply to the particular issue you're trying to solve.

Fixing Structural errors

Structural errors are errors that occur during measurement, data transfer, or other related processes. Mislabelled classes, the same attribute with a different name, typos in feature names, and other structural errors are examples of structural errors. For example, different classes that should be the same or inconsistent capitalization

Managing Unwanted outliers

Depending on the model type, outliers can be problematic. For example, decision tree models are more robust to outliers than linear regression models. In general, we shouldn't get rid of outliers until we have a good reason to. Removing them can sometimes enhance performance, but not always. In order to delete the outlier, one must have a valid explanation, such as suspicious measurements that are unlikely to be a part of actual data.

Handling missing data

A deceptively difficult problem in machine learning is missing data. The missing observation cannot be simply disregarded or eliminated. They need to be addressed carefully because they can be a sign of something significant.

Missing data is a deceptively tricky issue in machine learning. We cannot just ignore or remove the missing observation. They must be handled carefully as they can be an indication of something important. The two most typical methods for handling missing data are.

- Dropping observations with missing values.
- Imputing the missing values from past observations.

Data Visualization

To make data easier for the human brain to grasp and draw conclusions from data visualisation is the practise of putting information into a visual context, like a map or graph. Data visualization's major objective is to make it simpler to detect patterns, trends, and outliers in big data sets.

Feature Engineering

The technique of selecting, modifying, and transforming unprocessed data into features that can be applied in supervised learning is known as feature engineering. It might be important to create and train better features in order to make machine learning effective on new tasks. important phase in machine learning is feature engineering.

Splitting the data in training and testing sets

The train-test split approach is used to evaluate how well machine learning algorithms perform when used to predict outcomes from data that was not used to train the model. it is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modelling problem.

The telecom dataset is divided into two subsets: the training set, used to train Machine learning model, and the test set, used to evaluate the performance of the trained machine learning model.

Machine learning

As a subfield of artificial intelligence known as "machine learning," this term refers to computer algorithms that learn to solve a variety of prediction or detection tasks through the process of data-driven learning. There are three main methods for teaching machines: supervised, unsupervised, and reinforcement training methodologies . Massive amounts of data may now be analysed by Machine learning algorithms in a way that would otherwise be impossible or take a very long time for humans to do manually. Artificial intelligence (AI) and machine learning (ML) algorithms have found widespread use in a variety of industries and fields, including computer vision, agriculture, healthcare, communication, text and speech analysis, and so forth.

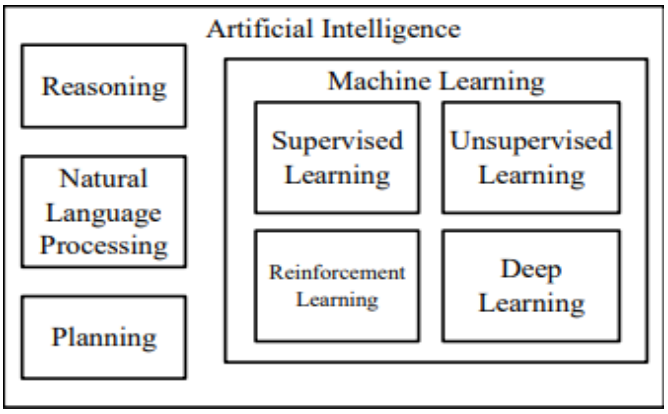

Figure 1.5: Overview of AI, ML, and DL

Supervised learning is a type of machine learning in which the output is predicted by the machines based on well-labelled training data. The term "labelled data" refers to input data that has already been assigned the correct output. In supervised learning, the training data that is

given to the computers serves as the supervisor, instructing them on how to correctly predict the output. We are using this algorithm in this study.

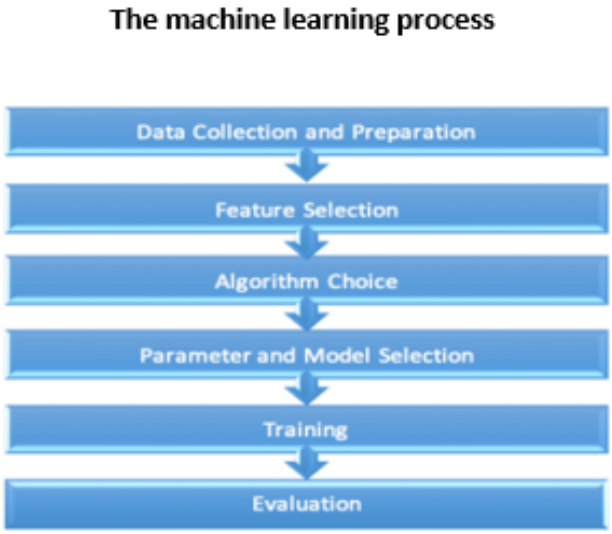

Figure 1.6: The machine learning process

Types of supervised Machine learning Algorithms

- Regression
- Classification

Classification

Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc. we are using this algorithm in our study

Algorithms Used

1 Logistic Regression

Logistic Regression is a supervised machine learning classification technique that is used to model the binary response. It can take an input variable that are either categorical, numeric or heterogeneous data types. Logistic regression is the proper model of regression analysis to utilize when the dependent variable is binary. A predictive analysis called logistic regression is used to explain the relationship between a group of independent binary variables and a dependent binary variable. For customer churn, logistic regression has been used to determine the probability of churn as a function of the characteristics or features of the customers. 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.

2 Decision Tree

The decision tree is the most prominent predictive model that is used for the purpose of classification of upcoming trial. The decision tree consists of two steps, tree building and tree pruning. In tree building the training set data is recursively partitioned in accordance with the values of the attributes. This process goes on until there is no one partition is left to have identical values. In this process some values may be removed from the data due to noisy data.

3 Random Forest

The Random Forest classifier is an ensemble learner, which is considered to be a supervised learning algorithm. The Random Forest classifier utilizes the bagging approach consisting of multiple classifiers Random Forest is an ensemble learning method meaning that it involves a combination of multiple models. RF operates by constructing a number of decision trees D = (D1, D2, · · · , DB) at training time using Bootstrap Aggregating or bagging and 15 random feature selection. Bagging tries to reduce the variance by averaging the outputs of many classifiers. Each individual decision tree in the Random Forest contributes to the final output by majority voting. The class with the most votes is chosen as the final output

Interpretation/ Evaluation

The model's accuracy can be evaluated based on the output. There are many different evaluation metrics available for predictive models created using data mining techniques, and each one has advantages and disadvantages.

Chapter -2

Review of Literature and Research Design

LITERATURE REVIEW

Jain, H., Khunteta, A., & Srivastava, S. (2020).

The problem of customer churn exists today in every industry, whether it is ISP, IT services, social networks, or mobile services (Customers changing their services from one service to another service provider). However, user churn is pretty common in the telecommunications industry. Due to the strong competition in the telecom market, businesses must proactively identify customer churn by analysing customer behaviour and attempt to invest time, effort and money in keeping consumers. The two machine-learning methods utilised in this suggested model to predict customer churn were logistic regression and logit boost. A actual database from the American company Orange and the WEKA Machine-learning technology were used in the experiment.

Rani, K. S., Thaslima, S., Prasanna, N. G. L., Vindhya, R., & Srilakshmi, P. (2021)

Customers are essential to the operation of business. Customer churn can have a variety of effects. Predicting customer churn must be a key component of every organization. This helps to identify users who are about to end their subscription to a service. The mobile telecommunications market recently transitioned from being one that was expanding quickly to one that was overcrowded. The goal of telecommunications firms is to shift their attention from attracting new, large customers to maintaining existing ones. Knowing which customers are likely to switch to a competitor in the future is important because of this. The model is put forth utilising machine learning methods, specifically logistic regression, for churn prediction for telecommunications organizations.

Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016, March).

One of the most popular study subjects in recent years has been the prediction of customer churn in the telecom business. It consists of identifying customers who are willing to withdraw their service subscriptions. The mobile telecommunications market has underwent a transformation from one of rapid growth to one of saturation and high competition. As a result, telecom firms are now more concerned with retaining existing customers than with acquiring new ones. Knowing which customers are most likely to switch to a competitor in the near future is useful because of this. Acquired data from the telecom sector can be used to study customer

churn factors and apply that knowledge to keep the customers. We have suggested using data mining and machine learning methods, such as logistic regression and decision trees, to create a model for predicting churn for telecommunications businesses.

Ahmed, A., & Linen, D. M. (2017, January).

Customer churn forecast has received increased attention from businesses, particularly those in the telecoms sector. The churn prediction models, which are heavily based on data mining principles and use machine learning and meta-heuristic algorithms, have been presented by several authors in various versions. This paper's goal is to examine some of the most significant churn prediction technologies created in recent years. The main goal is to accurately predict customer survival and customer hazard functions in the telecom industries in order to fully understand churn across the customer tenure. The goal of this research is to analyse churn prediction approaches in order to recognise churn behaviour and confirm the causes of customer churn. In order to have a better knowledge of customer churn, this paper provides an overview of churn prediction approaches.

Jain, H., Yadav, G., & Manoov, R. (2021).

One of the key considerations in optimising any organization's income is preventing customer churn. It happens when customers quit using a company's goods or services, and is also referred to as customer churn. We predict customer churn rate in our paper so that appropriate customer retention measures can be done with the support of exploratory data analysis and the implementation of targeted offers. Our implementation of the churn prediction uses comparative analysis of four algorithmic models—logistic regression, random forest, SVM, and XGBoost—on data from three different industries—banking, telecom, and IT. This comparison analysis was conducted since there aren't too many research studies that assess the effectiveness of different algorithms across various domains.

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019).

Customer Churn Prediction (CCP) is a hard process for decision-makers and the machine learning community because churn and non-churn users commonly have similar characteristics. It is clear from various research on customer churn and related data that a classifier demonstrates different levels of accuracy for various zones of a dataset. The accuracy of the anticipated classifier can then be predicted even before the classification. The datasets from the Telecommunication Industry (TCI) demonstrate that I the distance factor is strongly correlated with the classifier's level of certainty and (ii) the classifier performed better in the zone with a higher value of the distance factor (i.e., customer churn and non-churn with great certainty) than in the zone with a lower value of the distance factor (i.e., customer churn and non-churn with low assurance).

Singh, M., Singh, S., Seen, N., Kaushal, S., & Kumar, H. (2018, November).

In the telecom industry, customer churn is one of the most difficult and demanding problems. Currently, businesses' top priorities include keeping their current customers as well as seeking new users. In fact, due of the significant expenditures involved, customer retention is more essential. The current research analyses different machine learning techniques to a publicly accessible telecommunication dataset in the context of churn prediction modelling. Two key conclusions may be drawn from the results: (i) the Random Forest technique outperforms other primary classification models, and (ii) feature engineering is important to the model's accuracy.

Sai, B. K., & Sasikala, T. (2019, April).

Customer Churn is becoming one of the top concerns for many business organisations in the modern business world. Losing a customer will affect the bottom line and profit or loss of the company and pose a threat to everyday activities. Due to the large amount of passionate and competitive service providers, the telecommunications industry has recently faced considerable hurdles. Therefore, keeping current customers becomes very difficult. In this study, we used a Logistic Regression Model to execute an Exploratory Data Analysis using Visualization, Statistical Tests for Feature Selection, and Data Mining Methods for Predicting the Likely Churners.

Hota, L., & Dash, P. K. (October 2021)

The telecommunications sector is very competitive, in particular because of a wide range of active service providers. As a result, they have struggled to keep their current customers. Now is the moment to make sure the telecom business maintains value by prioritising customer retention over customer acquisition because getting new consumers is significantly more expensive than keeping existing ones. There have been several potential machine learning and data mining techniques. This study provides a brief overview of the customer churn problem and investigates the use of different machine learning techniques to predict customer churn using models like XGBoost, Gradient Boost, AdaBoost, ANN, Logistic Regression, and Random Forest. It also compares the accuracy of the models' and performance.

Kumar, M., & Kumar Yadav, D. (2022)

Many different fields are using machine learning to address the challenges. One of the main uses of machine learning is predictive analytics. Numerous methods, including machine learning, data mining, artificial intelligence, and statistics, are used in predictive analytics. It has the ability to predict the future using current data. Predicting customer churn is one of the uses for predictive analytics. The term "churn" has many distinct meanings depending on the context, however in this study, it refers to a situation where a consumer stops utilising any company's goods or services. Customer Churn Prediction (CCP) is significant since it has a direct impact on a company's profitability. The Gradient Boosting Machine (GBM) classifier is used in this paper's churn prediction model to identify potential consumers. Following classification, a variety of performance metrics are used to evaluate the customer's performance.

Statement of the problem

Predicting customer churn is critical for telecommunication companies to be able to effectively retain customers. Getting new customers is more expensive than keeping old ones so companies want to retain their existing customers. For this reason, large telecommunications corporations are seeking to develop models to predict which customers are more likely to stay and churn in future. Telecom is enduring the problem of ever-increasing churn rate. Because of this using machine learning models to predict the survival and churn analysis in telecommunication companies.

Need for the study

Customer retention can be achieved with good customer service and products. Keeping customers satisfied is truly essential for saying that business is successful especially in the tele-com industry. since the cost of acquiring a new customer is much higher than the cost of retaining the existing one. Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low. Accordingly, the current study employs machine learning algorithm. Machine learning algorithm techniques help the telecom firms to protect the loyalty customer.

Objectives

1) To study the factors influencing customers churn and survival of a business organization.

2) To determine the impact of customer churn and survival analysis on telecom industry.

3) To develop an effective machine learning model for customer churn prediction and survival analysis for telecom industry.

Research Methodology

This study uses applied research design and it is analytical in nature

Sample Design

In this research, we have used publicly available telecommunication customer datasets [IBM Sample Data Sets]. The raw data collected from the online resource Kaggle be used to create machine learning models in this research. Each row represents a customer, each column contains customer's attributes described on the column Metadata. The raw data contains 7043 rows (customers) and 21 columns (features). Moreover, the data is heterogeneous meaning it contains mixed-data types; categorical and numeric features.

The data contains customer information including

- Demographic Information (gender, Senior Citizen, Partner, Dependents).
- Customer Account Information (how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges)
- Services that each customer has signed up for – (phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies)
- Customers who left within the last month – the column is called Churn

Sources of data

Secondary Data

The secondary data are the second-hand information which are already collected by someone (organization) for some purpose and are available for the present study. For this study data is collected from Kaggle. Telecommunication customer churn datasets [IBM Sample Data Sets].

The secondary data needed for this study will be gathered from various sources

- Websites
- Journals
- Research articles

TOOLS OF DATA COLLECTION

MS Excel

PLAN OF ANALYSIS

It is planned to undertake the following for developing churn prediction model

- Data collection and Reading
- Exploratory Data Analysis
- Data Cleaning
- Data Visualization
- Identifying significant Features
- Feature Engineering
- Create Machine learning algorithms
- Splitting the data in training (80%) and testing sets (20%)
- Prediction using the model
- Interpretation
- Python has been used to conduct data analysis

Limitation of study

- This is based on secondary data.
- The time was limited
- There was limited access to information.
- The current study is limited to telecom industry only .
- The study does not use other techniques rather than machine learning techniques

Chapter -3

Profile of the industry

Introduction of telecom industry

The telecommunications industry is made up of businesses that enable global communication, whether it be through the phone or the Internet, over airways or cables, via wires or wirelessly. These businesses developed the network that enables the transmission of data in the form of text, voice, audio, or video throughout the globe. Internet service providers, satellite companies, cable companies, and telephone operators are the major firms in the industry.

- The telecommunications industry is made up of firms that send data in the form of words, voice, audio, or video around the world.
- The three primary subsectors of telecommunications are wireless communication, telecom services, and telecom equipment.
- The concentration in telecom has shifted away from voice and toward data, text, and video.

Telecommunications Industry Segments

The largest sub-sector of the telecommunications industry is telecom equipment (the largest), followed by telecom services (next largest) and wireless communication.

The major segments within these sub-sectors include the following:

- Broadband
- Broadcast
- Network equipment
- Wireless communications
- Communications equipment
- Processing systems and products
- Long-distance carriers
- Domestic telecom services
- Foreign telecom services
- Diversified communication services

Telecommunications' wireless communications market is expanding very quickly as more and more people switch to mobile devices and cloud-based computing. The projected foundation for the telecoms sector's continued global expansion is this area of the industry.

The primary issue for the industry going forward is to meet consumer demand for faster data access, higher resolution, quicker video streaming, and a wide range of multimedia applications.

Global telecommunication industry

The global telecommunications market has been undergoing constant transformation. With the adoption and integration of cutting-edge communication technologies like 5G, IoT, NB-IoT, and many others. Through a wired or wireless channel, telecommunication is an infrastructure that enables the transfer of various forms of data, including voice, text, audio, and video, at any time and anywhere in the globe. The field of telecommunications has developed into a competitive industry that has a big impact on worldwide economic and social advancement.

The telecom sector includes telephone, telecommunication, and internet service providers. Over the past few decades, the sector has grown from telephone voice calls to mobile data calls, internet calls, text messages, internet, and TV broadcasting services. The usage of smartphones is growing along with the increase of population and social modernisation.

REGIONAL ANALYSIS

The expansion of the telecom market is driven by the existence of active players globally. The enormous population of Asian nations like China and India is contributing to the region's huge increase in the number of telecom subscribers. North American market growth is anticipated to be accelerated by the rapid technology advancements in the region. The growth of the industry throughout the region is also significantly attributed to the presence of major telecom companies. The regional telecom market's continued expansion was also facilitated by network investment optimization for customer satisfaction.

The implementation of the Internet of Things and growing use of Artificial intelligence And machine learning support the expansion of the telecom sector. The introduction of next-generation high-speed networks has resulted in an increase in the demand for telecom services in business applications. To access faster data bandwidth and reduce network latency, businesses are building 5G small cell networks as well as private LTE and 5G networks. Businesses would be able to meet customer requirements with the least amount of waiting time

because to the improved bandwidth connectivity, which would also enhance the entire consumer experience. It would be necessary for businesses worldwide to have unified and quick network connectivity in order to handle a rapidly growing portion of customer datasets. From 2021 to 2028, this is expected to accelerate the segment's expansion.

SEGMENTATION

ATTRIBUTE	DETAILS	

By Product	• Telecommunication Hardware
	  ○ Mobile Devices
	  ○ PBX Equipment
	  ○ Networking Equipment
	• Telecommunication Services
	  ○ Fixed Networking Services
	  ○ Mobile Networking Services	

By Telecommunication Services, by Type	• Voice • Data	
By Application	• Commercial • Residential	
By Geography	• North America (U.S. and Canada) • Europe (UK, Germany, France, Scandinavia and Rest of Europe) • Asia Pacific (Japan, China, India, Southeast Asia and Rest of Asia Pacific) • Middle East & Africa (South Africa, GCC and Rest of the Middle East & Africa) • Latin America (Brazil, Mexico and Rest of Latin America)	

Figure 3.1: Global telecom Segmentation

Regional Insights

Asia Pacific maintained the highest market share in 2020, accounting for more than 32.0%, and is anticipated to have strong growth between 2021 and 2028. By 2028, the region will probably draw more than half of all new mobile customers. Platforms for e-commerce and retailer buy-in, the increasing use of smartphones, and investments in 5G networks are the main factors

driving the regional market. The rise of the regional market has seen significant contribution from China, Japan, and India. China and India were among the countries with the highest numbers of internet users worldwide in 2020, with more than 854 million and 560 million users, respectively, according to statistics released by the International Telecommunication Union (ITU).

Key Companies & Market Share Insights

Telecom Services Market Report Scope

Report Attribute	Details
Market size value in 2021	USD 1,707.96 billion
Revenue forecast in 2028	USD 2,467.01 billion
Growth Rate	CAGR of 5.4% from 2021 to 2028
Base year for estimation	2020
Historical period	2017 - 2019
Forecast period	2021 - 2028
Quantitative units	Revenue in USD billion and CAGR from 2021 to 2028
Report coverage	Revenue forecast, company ranking, competitive landscape, growth factors, and trends
Segments covered	Service type, transmission, end-use, region
Regional scope	North America; Europe; Asia Pacific; Latin America; MEA
Country scope	U.S.; Canada; U.K.; Germany; France; Italy; China; Japan; India; South Korea; Brazil; Mexico
Key companies profiled	AT&T Inc.; Verizon Communications Inc.; NTT; China Mobile Ltd.; Deutsche Telekom AG; SoftBank Group Corp.; China Telecom Corp Ltd.; Telefonica SA; Vodafone Group; KT Corporation; Bharati Airtel Limited; Reliance Jio Infocomm Limited; KDDI Corporation; Orange SA; BT Group plc; Comcast Corporation

Figure 3.2: Telecoms Key Companies & Market Share Insights

Telecommunication industry of India

With 1.2 billion subscribers, India is currently the second-largest telecoms market in the world. The Indian mobile economy is predicted to significantly contribute to India's GDP because of its rapid expansion.

The subsectors of the telecommunications business that are broken down into these groups. Infrastructure, equipment, mobile virtual network operators (MNVO), white space spectrum, 5G, telephone service providers, and broadband. Nearly the past ten years, India has added over 500 million additional smartphone users. We are expected to have 850 Mn smartphone users by 2026, representing ~55% of the total population. The Indian telecom tower market has expanded dramatically by 65% during the past seven years.

Telecom Milestones in India

- The establishment of the first "telegraph" in Calcutta in the 1850s began the process of telecom development in India (today known as Kolkata).
- Mumbai and Chennai started operating telephone exchanges in 1882. With 93 first subscribers, Kolkata's exchange was given the name "Central Exchange."
- In 1981, the Indian telecommunications sector began to liberalise. The Centre for Development of Telematics (C-DOT) was established to produce electronic telephone exchanges In India.
- As the need for telephones increased in the 1990s, houses received landline telecommunications equipment. As part of LPG (Liberalisation, Privatisation, Globalisation) policy, the Indian government also opened up the telecom sector for private investment during this time.
- In 1994, the government launched the National Telecommunications Policy (NTP), which altered infrastructure ownership, service, and regulation with the goal of spreading telecom infrastructure to all of India's villages.
- From 2005 on, internet acceptance and usage increased, but last-mile reachability remained challenging. To overcome this limitation, the government promoted the use of 3G and 4G spectrum, creating a competitive environment in which private companies could provide value-added services.
- The Indian telecom industry is currently in a good position to attract more than 900 million Internet users by 2025.
- It is predicted that 5G technology will boost the Indian economy by $450 billion between 2023 and 2040. IMT/5G spectrum auctions are currently scheduled to take place by the end of July 2022.

The largest companies involved in the Indian telecom market are,

- BSNL
- MTNL
- Reliance Communications
- Bharti Airtel
- Vodafone Idea
- Mahanagar Telephone Nigam Limited
- Aircel
- Tata Teleservices
- Telenor
- Jio

Top 5 Telecom Companies in India | Operators

1. Reliance Jio

Reliance Jio Info COMM Limited, also known as Jio, is an Indian telecommunications services provider with its headquarters in Mumbai, Maharashtra, with a nationwide LTE network that covers all 22 telecom circles. Jio is a wholly owned subsidiary of Reliance Industries.

Market share: **36.5 %**

According to market share, Reliance Jio is India's leading telecom provider. With more than 372.99 million users as of May 31, 2019, it is the biggest mobile network operator in India and the third-largest in the entire world.

2. Bharti Airtel Limited

Leading international telecommunications provider Bharti Airtel Limited has operations in 18 nations across Asia and Africa. The Company, which has its headquarters in New Delhi, India, is one of the top three mobile service providers worldwide in terms of users.

2G, 3G, and 4G wireless services, mobile commerce, fixed line services, high-speed home internet, DTH, and enterprise services are among the company's product offerings in India

(including national & international long distance services to carriers). In India, it is a private telecom operator.

Market share: 27.5 %

It provides mobile commerce and 2G, 3G, and 4G wireless services in various nations. More than 403 million people use Bharti Airtel services worldwide.

3. Vodafone Idea Limited

The Aditya Birla Group and the Vodafone Group have partnered to form Vodafone Idea Limited. It is the top telecom service provider in India. On 2G, 3G, and 4G platforms, the company offers voice and data services across the entirety of India.

Market Share: 23.5 %

4. Bharat Sanchar Nigam Ltd.

The Bharat Sanchar Nigam Ltd. company was established on September 15th, 2000. With effect from 1 October 2000 on an ongoing basis, it takes over the business of providing telecom services and network management from the former Central Government Departments of Telecom Services (DTS) and Telecom Operations (DTO).

It is one of India's largest and most prestigious public sector organisations, offering a wide variety of telecom services. The business ranks fourth among India's largest telecom service providers.

Market share: 9.9 %

5. MTNL (Mahanagar Telephone Nigam Limited)

The Government of India founded MTNL on April 1st, 1986 with the goal of improving the quality of telecom services, expanding the communication network, introducing new services, and generating income for the growth of telecom in India's two major metropolises, Delhi and Mumbai.

Market share: 0.29 %

The business is one of India's top 10 telecom service providers. MTNL now offers fixed telephone service, GSM (including 3G services), and CDMA-based mobile service, as well as internet, broadband, ISDN, and leased line services.

The Top 8 Technological Trends That Will Shape The Telecom Sector in future.

The telecom sector is determined to go through a digital transition in order to stay competitive in future. They are making use of technologies like Mobile Edge Computing (MEC), Internet - of - things (IoT), and 5G to bounce back from the COVID-19 problem and take a bold position in the future.

1. 5G network – The network of Today and Tomorrow

3 billion people use the internet worldwide. The primary way of internet connection for billions of these individuals is their smartphones.

The biggest growth driver for the telecom industry is 5G. Due to this telecom companies are banking on 5G faster data speed and better connectivity. Network efficiency will increase because to technologies like Software-defined networking (SDN) and Network Function Virtualization (NFV).

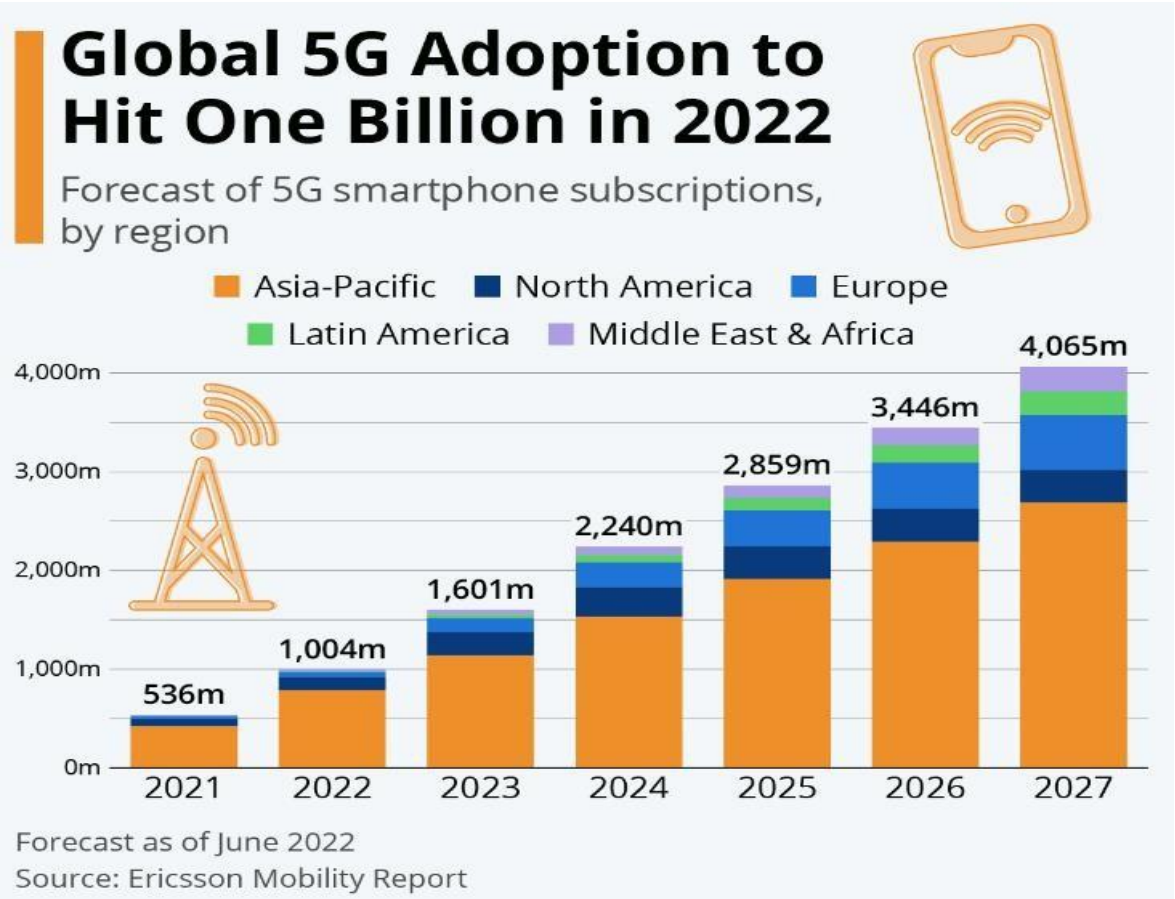

Figure 3.3: Global forecast of 5G smartphone subscription by region

2. Artificial Intelligence and Machine Learning

The growth of Artificial intelligence and Machine Learning will be extremely beneficial for the telecommunications industry. They already use Artificial Intelligence tools like chatbots, virtual assistants, and virtual assistants to customize the user experience and increase consumer experience.

Social media and technology have evolved to such an extent that there is an enormous growth of consumer data. Artificial intelligence and machine learning have made it possible to get consumer data at our fingertips. Consumer data gives the ability to conduct consumer analysis. With the help of consumer analysis, it has become possible to identify patterns and trends to make informed decisions.

3. Wi-Fi 6

A Wi-Fi network with an increased range that offers improved network performance, connects more devices, and offers faster communications is known as Wi-Fi 6. With additional technologies, it is possible to expand network capacity, decrease latency, and offer a variety of wide range of options for raising revenues and savings.

4. Internet of Things

Telecom service providers' responsibilities for facilitating communication between people and devices have changed as a result of the Internet of Things. IoT adoption enables the telecom sector to remotely monitor base stations and data centres. As a result, the network will have little downtime, business processes will be improved, and income will increase.

5. Big Data

With the use of IoT sensors in mobile devices and apps, telecom companies are able to collect vast amounts of data, building informational pools. As a result, they must make sure that their network can move huge amounts of data effectively and that it can handle new technologies.

The gathered data can be used by the telecom sector to develop critical business insights and understand customer usage trends. In the end, the data can be used to enhance customer service, evaluate innovative services, and keep an eye on and develop the network. Big data enables firms to grow stronger and use it to their advantage in the market.

6. Edge Computing

The worldwide demand for connected edge computing services will reach $7 billion by 2025. The Edge Computing ecosystem is rapidly evolving. It brings processing capabilities closer to the device's location in real-time, eliminating the need to transport data to a cloud data center and lowering waiting time. The data becomes more relevant and beneficial to the end-user with faster speed and on-time delivery. Edge computing also helps businesses, and their customers reduce total network traffic and improve performance. Not only that, but it also boosts network security by offering low-cost local encryption and security features.

7. Cloud Computing

According to a Mordor Intelligence analysis, the telecom cloud industry had a value of $25.33 billion in 2020 and was expected to grow to $74.36 billion by 2026. Therefore, during the course of the next several years, the majority of communication service providers (CSPs) will move their operations to the cloud. The majority of telecommunications companies rely on a massive computing infrastructure to deliver a wide range of apps, manage data, and bill for services. Moving to the cloud decreases the demand for internal computing resources, internal costs, and increases revenue streams.

8. Cyber Security

When it comes to cybersecurity, one of the most sensitive industries is the telecommunications industry. Due to the size of their subscriber base, telecom businesses present several options for malicious attempts to access their data without authorization.

When a major telecom service provider's infrastructure is attacked, the results might potentially have an impact on an entire country, as well as on businesses, consumers, and governmental organisations. Additionally, it has a significant impact on the trustworthiness and reputation of the telecom brand.

Chapter -4

Data Analysis and Interpretation

DATA DESCRIPTION

This sample data module tracks a fictional telco company's customer churn based on various factors. The churn column indicates whether the customer departed within the last month. Other columns include gender, dependents, monthly charges, and many with information about the types of services each customer has. Predicting customer churn is critical for telecommunication companies to be able to effectively retain customers.

Data Collection

The data set available in Kaggle is an adaptation of the original IBM data.

7043 observations with 21 variables

Following are the features present in the dataset:

Fields Names	DATA DESCRIPTION
Customer ID	Unique id of each customers
Gender	Whether the customer is a male or a female
Senior Citizen	Whether the customer is a senior citizen or not (1, 0)
Partner	Whether the customer has a partner or not (Yes, No)
Dependents	Whether the customer has dependents or not (Yes, No)
Tenure	Number of months the customer has stayed with the company
Phone Service	Whether the customer has a phone service or not (Yes, No)
Multiple Lines	Whether the customer has multiple lines or not (Yes, No, No phone service)
Internet Service	Customer's internet service provider (DSL, Fiber optic, No)
Online Security	Whether the customer has online security or not (Yes, No, No internet service)
Online Backup	Whether the customer has an online backup or not (Yes, No, No internet service)
Device Protection	Whether the customer has device protection or not (Yes, No, No internet service)
Tech Support	Whether the customer has tech support or not (Yes, No, No internet service)
Streaming TV	Whether the customer has streaming TV or not (Yes, No, No internet service)
Streaming Movies	Whether the customer has streaming movies or not (Yes, No, No internet service)

Contract	The contract term of the customer (Month-to-month, One year, Two years)
Paperless Billing	Whether the customer has paperless billing or not (Yes, No)
Payment Method	The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
Monthly Charges	The amount charged to the customer monthly
Total Charges	The total amount charged to the customer
Churn	Whether the customer churned or not (Yes or No)

1)Data Reading

The first step of the analysis consists of reading and storing the data in a Pandas data frame using the pandas.read_csv function.

```
# import telecom dataset into a pandas data frame
telco_data = pd.read_csv('Telco-Customer-Churn.csv')

# visualize column names
telco_data.columns
```

```
Index(['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents',
       'tenure', 'PhoneService', 'MultipleLines', 'InternetService',
       'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport',
       'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling',
       'PaymentMethod', 'MonthlyCharges', 'TotalCharges', 'Churn'],
      dtype='object')
```

2. Exploratory Data Analysis and Data Cleaning

Check the various attributes of data like shape (rows and cols), Columns, datatypes

```
telco_data.shape
```

```
(7043, 21)
```

```
# summary of the data frame
telco_data.info()
```

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

Interpretation: The data set, which is displayed above, includes 21 columns and 7043 observations. Apparently, the data set contains no null values. However, we see that the column Total Charges was incorrectly identified as an object.

The data set contains details about:

- Customer demographic information, including gender, age range, and the status of partners and dependents
- The Churn column lists customers who have left in the past month.
- Account details for each customer, including their contract, payment method, paperless billing information, monthly costs, and total charges;

- The subscription services each customer has chosen, such as phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.

```
# Check the descriptive statistics of numeric variables
telco_data.describe()
```

	SeniorCitizen	tenure	MonthlyCharges
count	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692
std	0.368612	24.559481	30.090047
min	0.000000	0.000000	18.250000
25%	0.000000	9.000000	35.500000
50%	0.000000	29.000000	70.350000
75%	0.000000	55.000000	89.850000
max	1.000000	72.000000	118.750000

Interpretation: Senior Citizen is actually a categorical hence the 25%-50%-75% distribution is not proper 75% customers have tenure less than 55 months Average Monthly charges are USD 64.76 whereas 25% customers pay more than USD 89.85 per month.,

```
100*telco_data['Churn'].value_counts()/len(telco_data['Churn'])
```

```
No     73.463013
Yes    26.536987
Name: Churn, dtype: float64
```

```
telco_data['Churn'].value_counts()
```

```
No     5174
Yes    1869
Name: Churn, dtype: int64
```

Data is highly imbalanced, ratio = 73:27

So we analyse the data with other features while taking the target values separately to get some insights.

3.Data Cleaning

1. Create a copy of base data for manipulation & processing

```
df = telco_data.copy()
```

2. Total Charges should be numeric amount. Let's convert it to numerical data type

```
df.TotalCharges = pd.to_numeric(df.TotalCharges, errors='coerce')
df.isnull().sum()
```

```
customerID           0
gender               0
SeniorCitizen        0
Partner              0
Dependents           0
tenure               0
PhoneService         0
MultipleLines        0
InternetService      0
OnlineSecurity       0
OnlineBackup         0
DeviceProtection     0
TechSupport          0
StreamingTV          0
StreamingMovies      0
Contract             0
PaperlessBilling     0
PaymentMethod        0
MonthlyCharges       0
TotalCharges        11
Churn                0
dtype: int64
```

3. As we can see there are 11 missing values in Total Charges column. Let's check these records

4. Missing Value Treatement

Since the % of these records compared to total dataset is very low ie 0.15%, it is safe to ignore them from further processing.

```
#Removing missing values
df.dropna(how = 'any', inplace = True)

#df.fillna(0)
```

```
# checking for row and cols after removing na values
df.shape
```

```
(7032, 21)
```

Interpretation: Removed the 11-missing value from total chargers column. After removing null values remaining row are 7032 and 21 columns.

5. Remove columns not required for processing

```
#drop column customerID and tenure
df.drop(columns= ['customerID'], axis=1, inplace=True)
df.head()
```

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetServ
0	Female	0	Yes	No	1	No	No phone service	D
1	Male	0	No	No	34	Yes	No	D

4. Data exploration and Visualisation

1. Demographics to understand the gender, age range, partner and dependent status of the customers.

```
colors = ['#4D3425','#E4512B']
ax = (df['gender'].value_counts()*100.0 /len(df)).plot(kind='bar',
                                                        stacked = True,
                                                        rot = 0,color = colors)

ax.set_ylabel('% Customers')
ax.set_xlabel('Gender')
ax.set_ylabel('% Customers')
ax.set_title('Gender Distribution')

# create a list to collect the plt.patches data
totals = []

# find the values and append to list
for i in ax.patches:
    totals.append(i.get_width())

# set individual bar lables using above list
total = sum(totals)

for i in ax.patches:
    # get_width pulls left or right; get_y pushes up or down
    ax.text(i.get_x()+.15, i.get_height()-3.5, \
            str(round((i.get_height()/total), 1))+'%',
            fontsize=12,
            color='white',
           weight = 'bold')
```

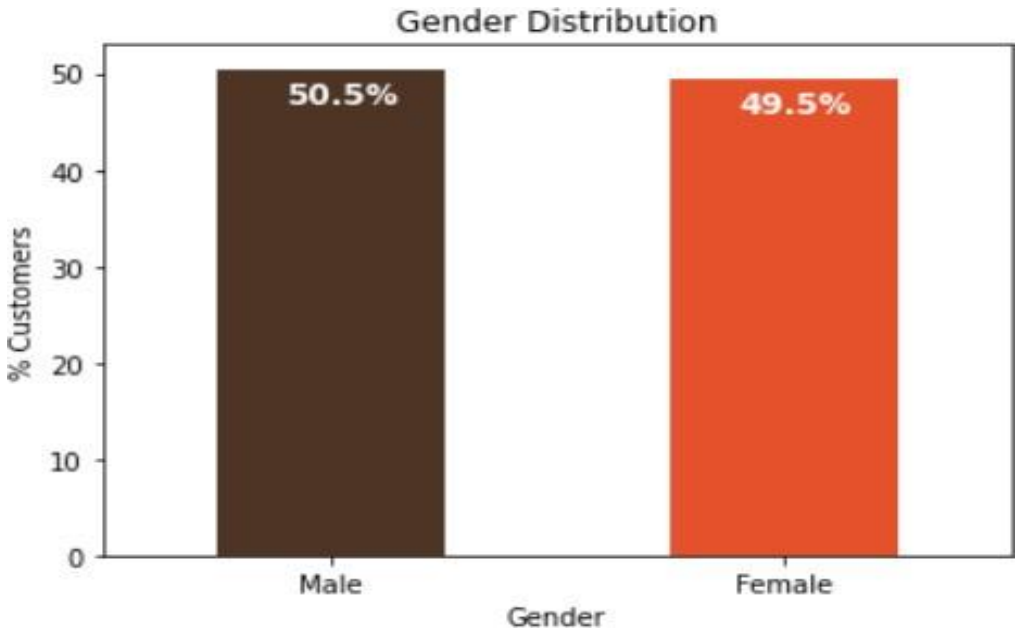

Figure 4.1: Gender Distribution

Interpretation: Gender Distribution - About half of the customers in our data set are male(50.50%) while the other half are female (49.5%).

```
ax = (df['SeniorCitizen'].value_counts()*100.0 /len(df))\
.plot.pie(autopct='%.1f%%', labels = ['Younger', 'Senior'],
          figsize =(5,5), fontsize = 13 )

ax.set_ylabel('Senior Citizens',fontsize = 13)
ax.set_title('% of Senior Citizens', fontsize = 13)
```

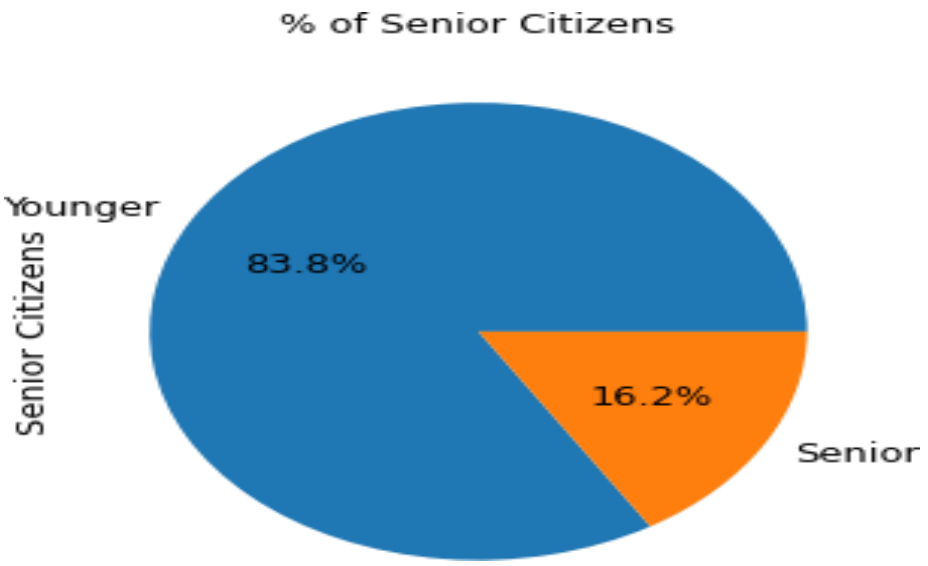

Figure 4.2: Percentage of senior citizen and younger population

Interpretation: % Senior Citizens - There are only 16% of the customers who are senior citizens. Thus most of our customers in the data are younger people like (83%).

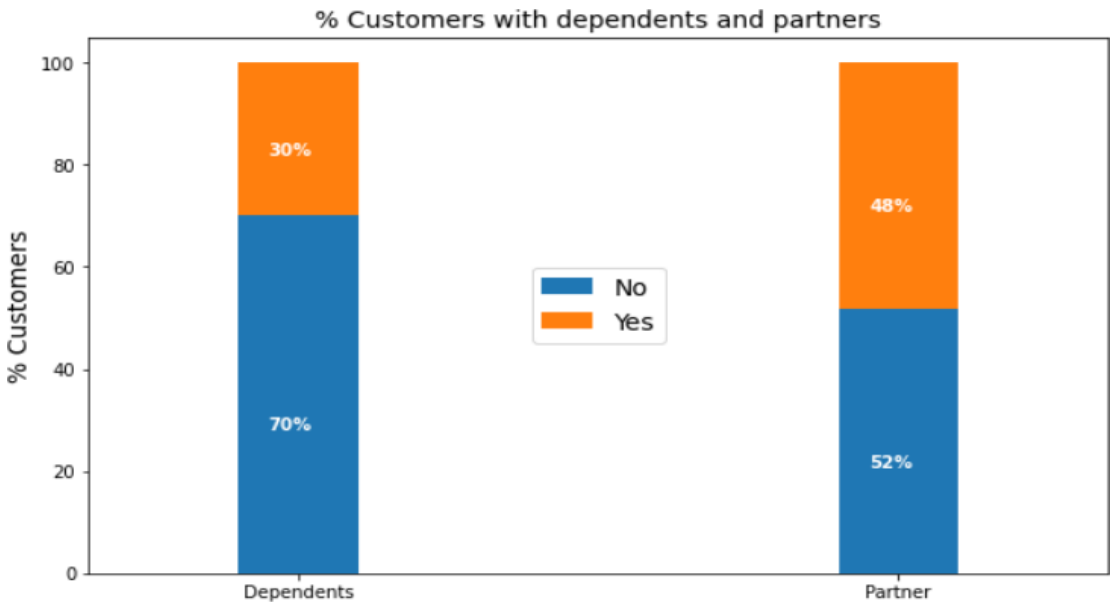

Figure 4.3: percentage of customer with dependent's and partners

Interpretation: From above plot, it can be seen that, 48% of the customers have a partner channel, while only 30% of the total customers have dependents.

It would be interesting is to look at the % of customers with both partners and dependents.

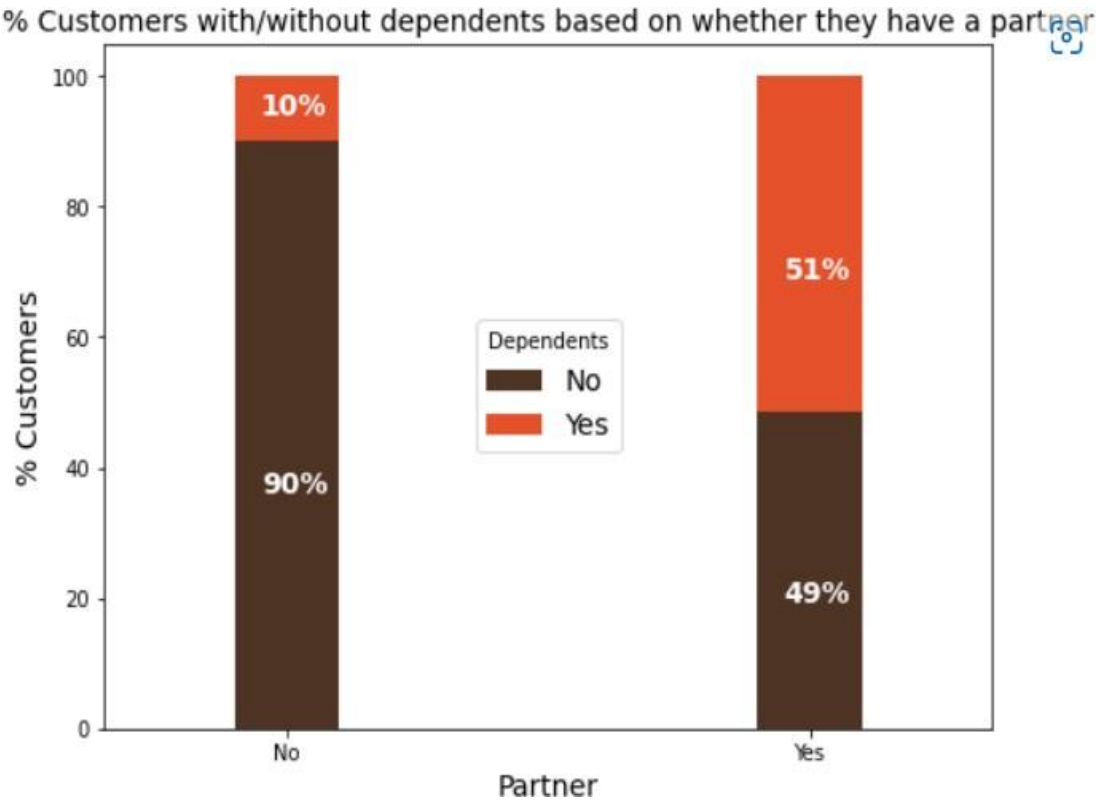

Figure 4.4: % of customers with both partners and dependents.

Interpretation: Interestingly, among the customers who have a partner, only about half of them (51%) also have a dependent, while other half (49%) do not have any independents. Additionally, as expected, among the customers who do not have any partner, a majority (90%) of them do not have any dependents

A review of any differences between the % of customers with/without dependents and partners by gender reveal that, there is no difference in their distribution by gender. Additionally, there is no difference in senior citizen status by gender.

2. Customer Account Information:

```
ax = sns.distplot(telco_data['tenure'], hist=True, kde=False,
                  bins=int(180/5),  color = 'darkred',
                  hist_kws={'edgecolor':'black'},
                  kde_kws={'linewidth':25 })
ax.set_ylabel('Number of Customers')
ax.set_xlabel('Tenure (months)')
ax.set_title('Number of Customers by tenure')
```

```
Text(0.5, 1.0, 'Number of Customers by tenure')
```

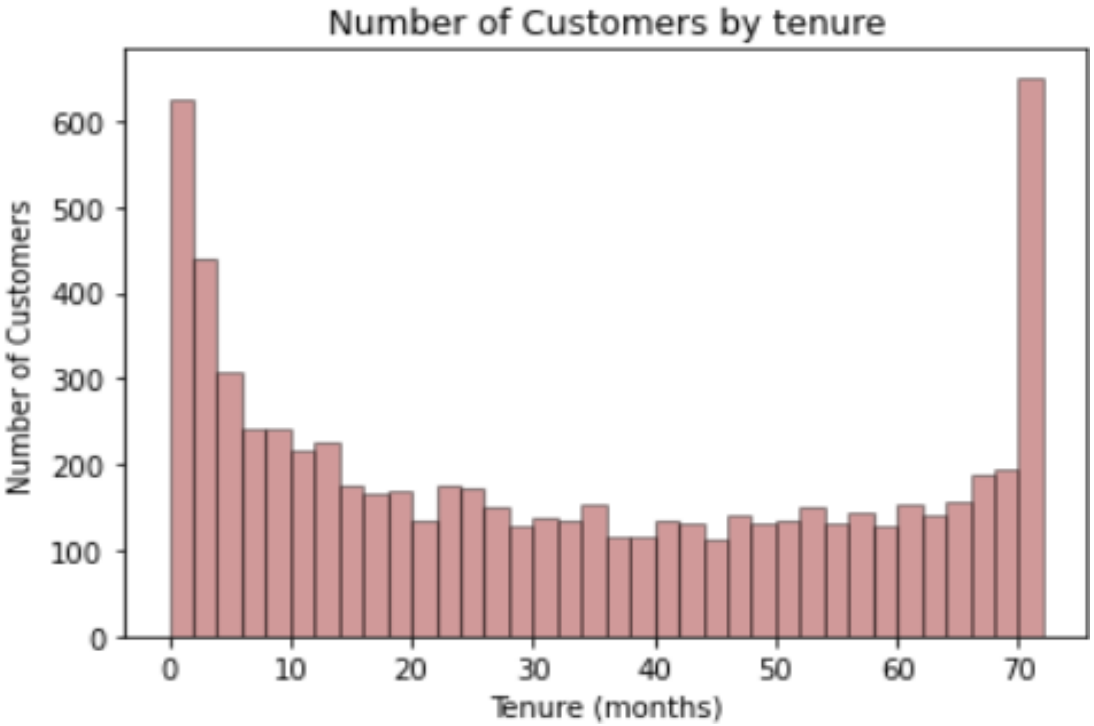

Figure 4.5: Number of Customers by tenure

Interpretation: The above histogram displays a lot of customers have been with the telecom company for just a month, while quite a many are there for about 70 months. This could be potentially because of different contractual agreement some customers might have.

3. Checking the customers by tenure of contract

```
ax = telco_data['Contract'].value_counts().plot(kind = 'bar',rot = 0,
                                                width = 0.3)
ax.set_ylabel('Number of Customers')
ax.set_title('Number of Customers by Contract Type')
```

```
Text(0.5, 1.0, 'Number of Customers by Contract Type')
```

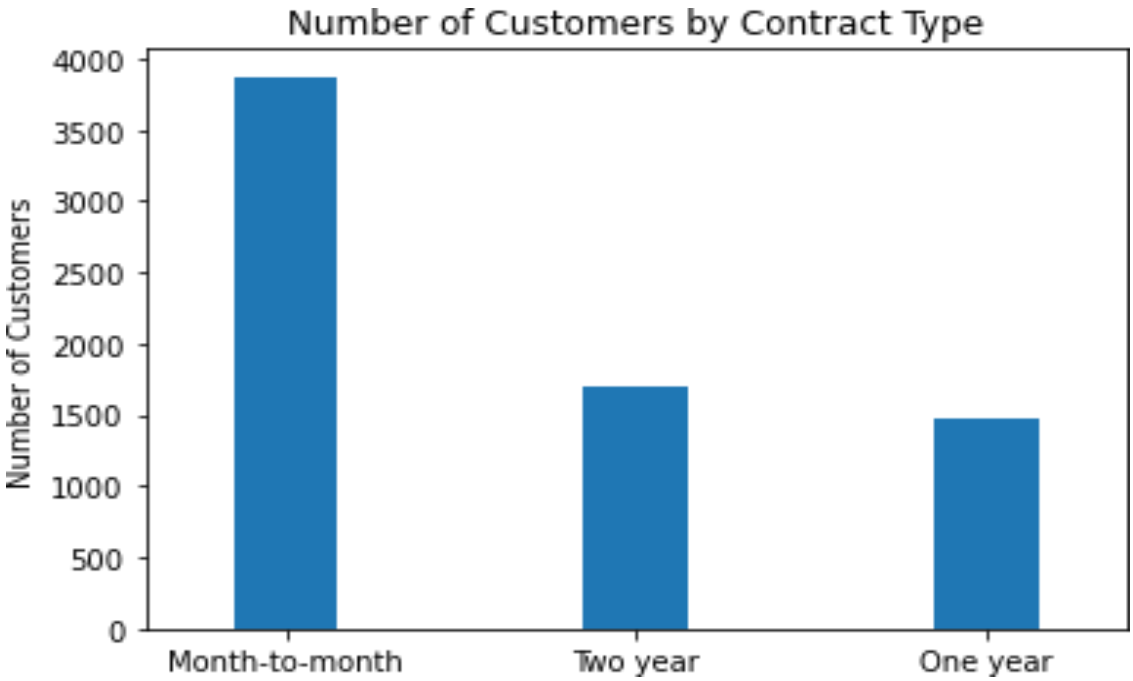

Figure 4.6: Number of Customers by Contract Type

Interpretation: Interestingly most of the monthly contracts last for 1-2 months, while the 2 year contracts tend to last for about 70 months. This shows that the customers taking a longer contract are more loyal to the company and tend to stay with it for a longer period of time. This is also visible on the correlation chart with the churn rate. More of less equal number of customers in the 1 year and 2 year contracts; around 1500 customers are on 1 year contract and close to around 1700 customers are on 2 years contracts.

Below we will understand the tenure of customers based on their contract type.

```
fig, (ax1,ax2,ax3) = plt.subplots(nrows=1, ncols=3, sharey = True, figsize = (20,6))

ax = sns.distplot(telco_data[telco_data['Contract']=='Month-to-month']['tenure'],
                  hist=True, kde=False, bins=int(180/5), color = 'turquoise',
                  hist_kws={'edgecolor':'black'},kde_kws={'linewidth': 4},ax=ax1)
ax.set_ylabel('# of Customers')
ax.set_xlabel('Tenure (months)')
ax.set_title('Month to Month Contract')

ax = sns.distplot(telco_data[telco_data['Contract']=='One year']['tenure'],
                  hist=True, kde=False,bins=int(180/5), color = 'steelblue',
                  hist_kws={'edgecolor':'black'},kde_kws={'linewidth': 4},ax=ax2)
ax.set_xlabel('Tenure (months)',size = 14)
ax.set_title('One Year Contract',size = 14)

ax = sns.distplot(telco_data[telco_data['Contract']=='Two year']['tenure'],
                  hist=True, kde=False,bins=int(180/5), color = 'darkred',
                  hist_kws={'edgecolor':'black'},kde_kws={'linewidth': 4},ax=ax3)
ax.set_xlabel('Tenure (months)')
ax.set_title('Two Year Contract')
```

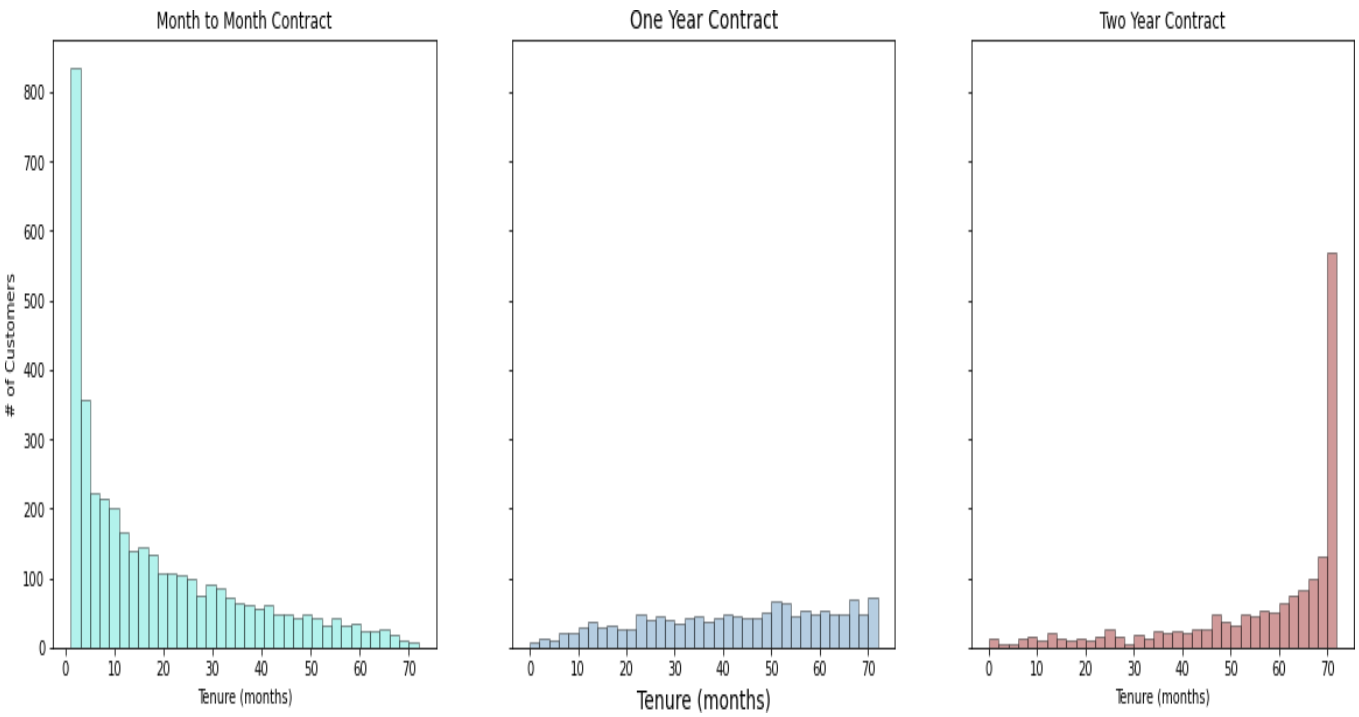

Figure 4.7: Number of customers by Month to Month Contract, One Year Contract, Two Year Contract

Interpretation: From this plot we will understand the tenure of customers based on their contract type like month – month contract, one year contract, and two-year contract. There are more customers in month – month contract. Interestingly most of the monthly contracts last for

1-2 months, while the 2 year contracts tend to last for about 70 months. This shows that the customers taking a longer contract are more loyal to the company and tend to stay with it for a longer period of time.

Let us now look at the distribution of various services used by customers

```
telco_data.columns.values
```

```
array(['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents',
       'tenure', 'PhoneService', 'MultipleLines', 'InternetService',
       'OnlineSecurity', 'OnlineBackup', 'DeviceProtection',
       'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract',
       'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges',
       'TotalCharges', 'Churn'], dtype=object)
```

```
services = ['PhoneService','MultipleLines','InternetService','OnlineSecurity',
            'OnlineBackup','DeviceProtection','TechSupport','StreamingTV','StreamingMovies']

fig, axes = plt.subplots(nrows = 3,ncols = 3,figsize = (15,12))
for i, item in enumerate(services):
    if i < 3:
        ax = telco_data[item].value_counts().plot(kind = 'bar',ax=axes[i,0],rot = 0,color = colors)

    elif i >=3 and i < 6:
        ax = telco_data[item].value_counts().plot(kind = 'bar',ax=axes[i-3,1],rot = 0,color = colors)

    elif i < 9:
        ax = telco_data[item].value_counts().plot(kind = 'bar',ax=axes[i-6,2],rot = 0,color = colors)
    ax.set_title(item)
```

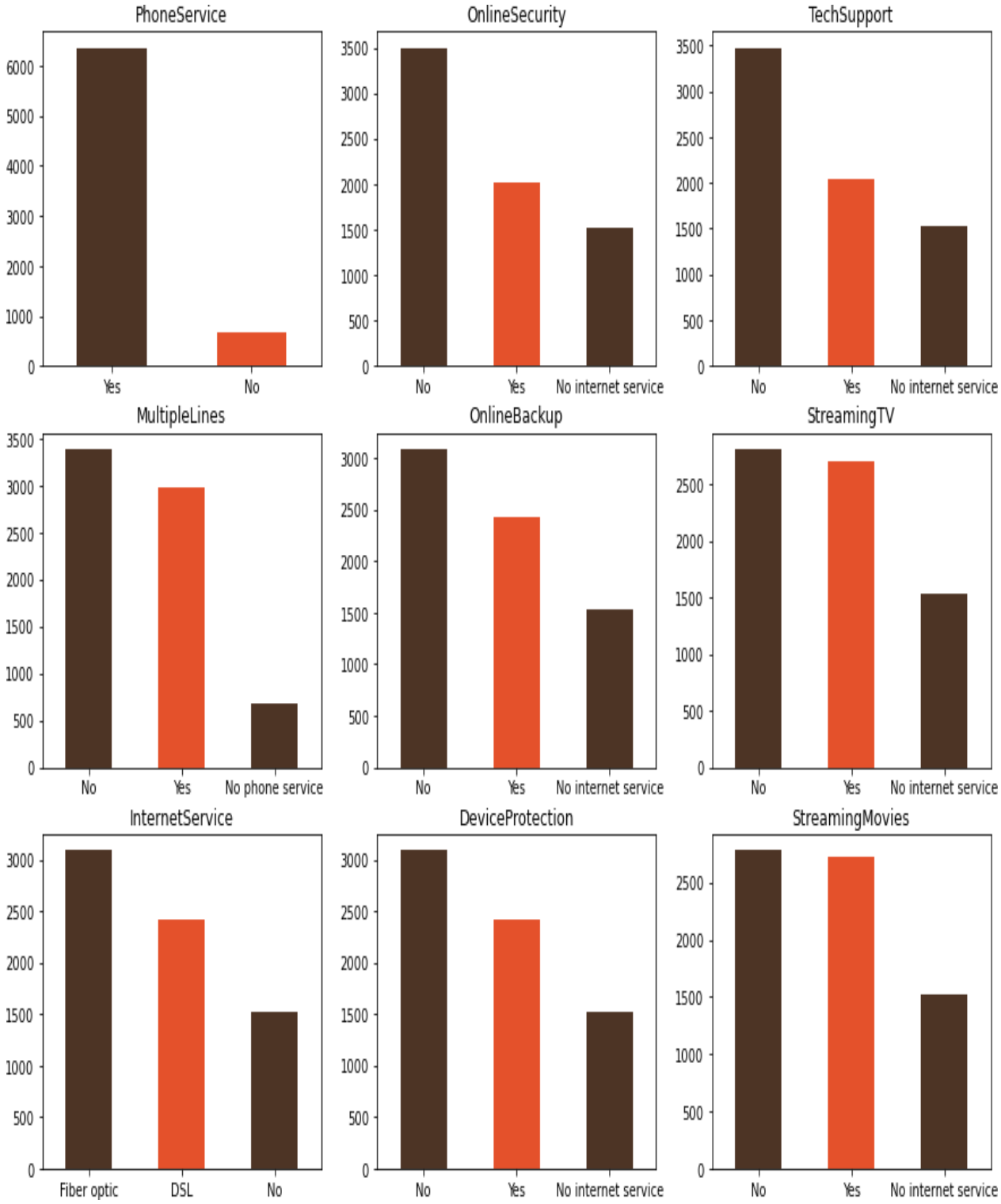

Figure 4.8: The distribution of various services used by customers

Interpretation: Above plot displays a clear visibility of the different services provided by the Telecom company and avail abled by the customers.

Datamining to check prediction variable and it's interaction with other variables

```
ax = (df['Churn'].value_counts()*100.0 /len(df)).plot(kind='bar',
                            stacked = True,rot = 0,color = colors )
ax.set_ylabel('% Customers')
ax.set_xlabel('Churn')
ax.set_title('Churn Rate')

# create a list to collect the plt.patches data
totals = []

# find the values and append to list
for i in ax.patches:
    totals.append(i.get_width())

# set individual bar lables using above list
total = sum(totals)

for i in ax.patches:
    # get_width pulls left or right; get_y pushes up or down
    ax.text(i.get_x()+.15, i.get_height()-4.0, \
            str(round((i.get_height()/total), 1))+'%', color='white')
```

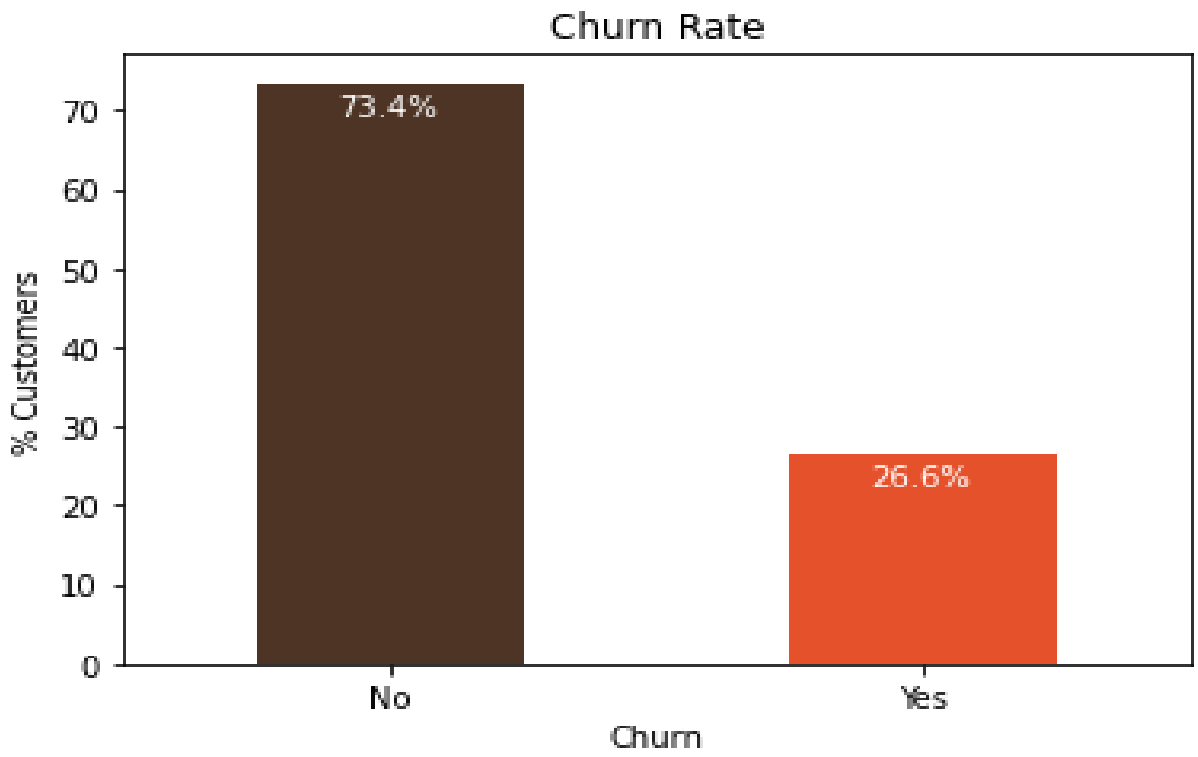

Figure 4.9: checking for churn rate of customers

Interpretation: 26.5% churn rate is considered quite high. However, 73.5 % of the customers do not churn or 73.5 % survived in the service. Clearly the data is skewed as it is always expected that a large majority of the customers to not churn. Skewness' could lead to a lot of false negatives in predictive modelling.

Churn by Contract Type

```
contract_churn = telco_data.groupby(['Contract','Churn']).size().unstack()

ax = (contract_churn.T*100.0 / contract_churn.T.sum()).T.plot(kind='bar',
                                                              width = 0.3,
                                                              stacked = True,
                                                              rot = 0,
                                                              figsize = (10,6))
ax.legend(loc='best',prop={'size':14},title = 'Churn')
ax.set_ylabel('% Customers',size = 14)
ax.set_title('Churn by Contract Type',size = 14)

# Code to add the data labels on the stacked bar chart
for p in ax.patches:
    width, height = p.get_width(), p.get_height()
    x, y = p.get_xy()
    ax.annotate('{:.0f}%'.format(height), (p.get_x()+.25*width, p.get_y()+.4*height),
                color = 'white',
               weight = 'bold',
               size = 14)
```

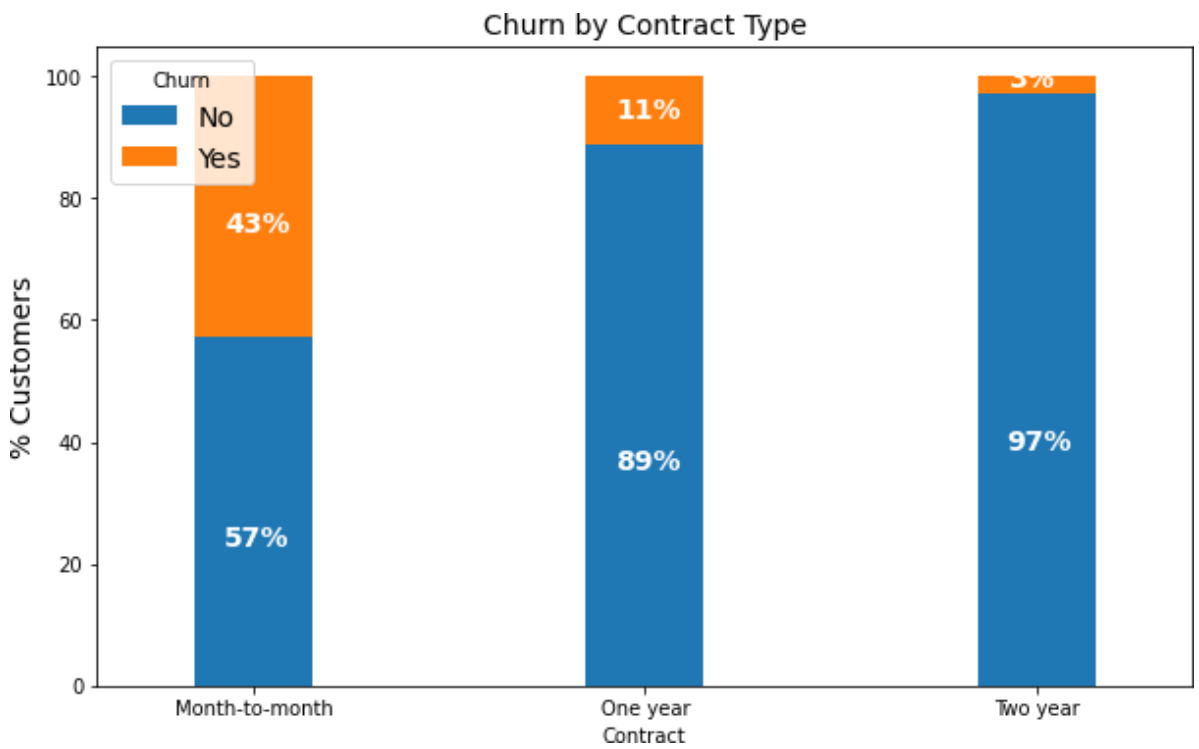

Figure 4.10: Churn rate by Contract Type of customer

Interpretation: From above graph the customers who have a month-to-month contract have a very high churn rate like 43%. The customers who have a two year contract have a very high survival rate like 97%.

Churn by Seniority

```
seniority_churn = telco_data.groupby(['SeniorCitizen','Churn']).size().unstack()

ax = (seniority_churn.T*100.0 / seniority_churn.T.sum()).T.plot(kind='bar',
                                                                width = 0.2,
                                                                stacked = True,
                                                                rot = 0,
                                                                figsize = (8,6),
                                                                color = colors)
ax.legend(loc='center',prop={'size':14},title = 'Churn')
ax.set_ylabel('% Customers')
ax.set_title('Churn by Seniority Level',size = 14)

# Code to add the data labels on the stacked bar chart
for p in ax.patches:
    width, height = p.get_width(), p.get_height()
    x, y = p.get_xy()
    ax.annotate('{:.0f}%'.format(height), (p.get_x()+.25*width, p.get_y()+.4*height),
                color = 'white',
               weight = 'bold',size =14)
```

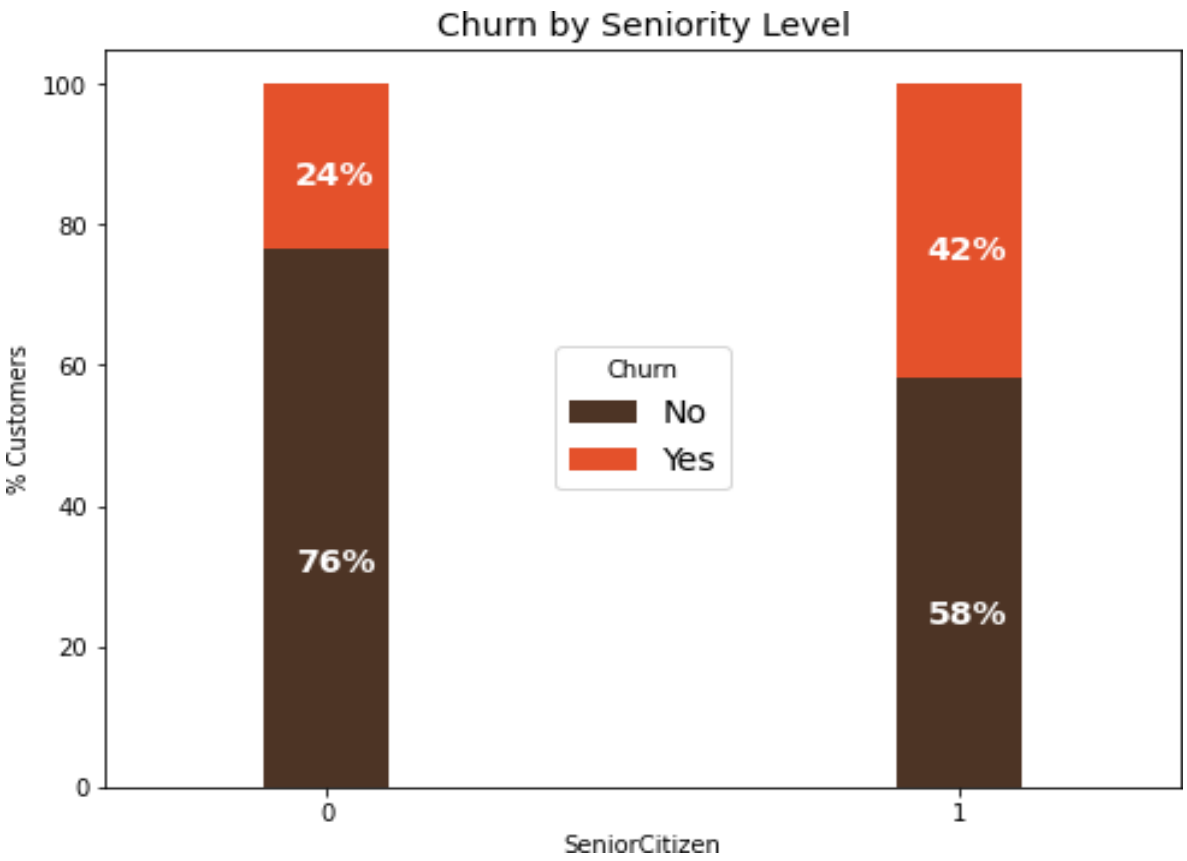

Figure 4.11: Churn rate by Seniority citizen

Interpretation: From above plot, it is apparent that Senior Citizens have almost double the churn rate than younger population. The younger population have more survival rate 76%.

Churn by Monthly Charges

```
ax = sns.kdeplot(telco_data.MonthlyCharges[(telco_data["Churn"] == 'No') ],
                 color="Red", shade = True)
ax = sns.kdeplot(telco_data.MonthlyCharges[(telco_data["Churn"] == 'Yes') ],
                 ax =ax, color="Blue", shade= True)
ax.legend(["Not Churn","Churn"],loc='upper right')
ax.set_ylabel('Density')
ax.set_xlabel('Monthly Charges')
ax.set_title('Distribution of monthly charges by churn')
```

```
Text(0.5, 1.0, 'Distribution of monthly charges by churn')
```

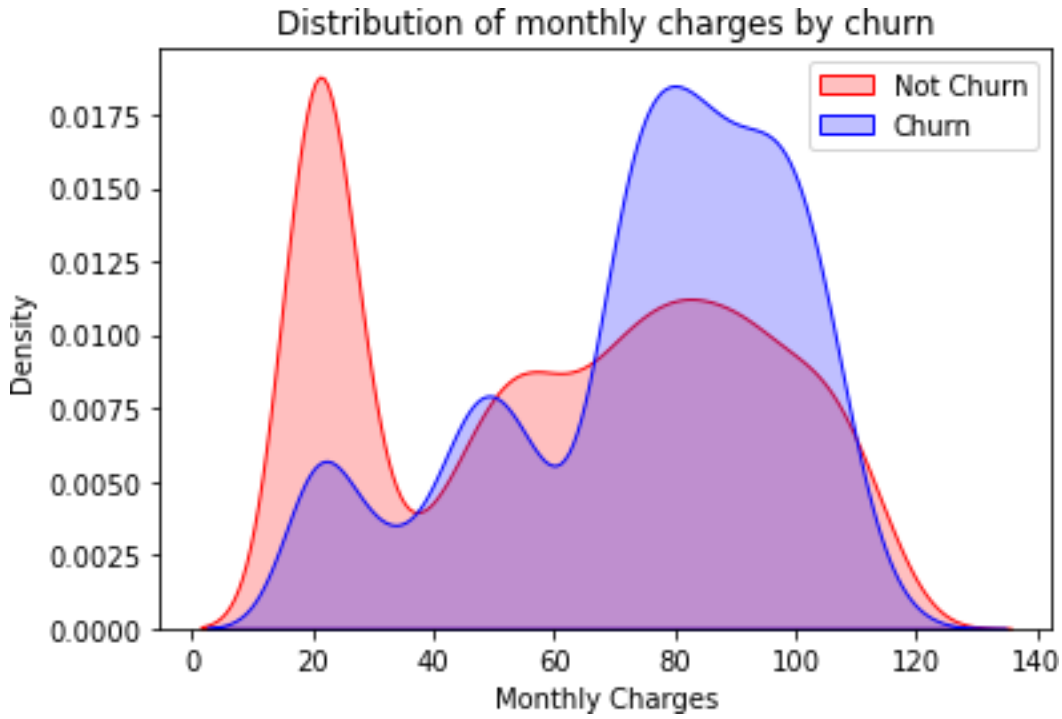

Figure 4.12: Distribution of monthly charges by churn

Interpretation: From above plot, we can understand that Higher % of customers churn when the monthly charges are high. And we can understand that Higher % of customers stayed when the monthly charges are less.

Churn by Total Charges

```
ax = sns.kdeplot(df.TotalCharges[(df["Churn"] == 'No') ],
                 color="Red", shade = True)
ax = sns.kdeplot(df.TotalCharges[(df["Churn"] == 'Yes') ],
                 ax =ax, color="Blue", shade= True)
ax.legend(["Not Churn","Churn"],loc='upper right')
ax.set_ylabel('Density')
ax.set_xlabel('Total Charges')
ax.set_title('Distribution of total charges by churn')
```

```
Text(0.5, 1.0, 'Distribution of total charges by churn')
```

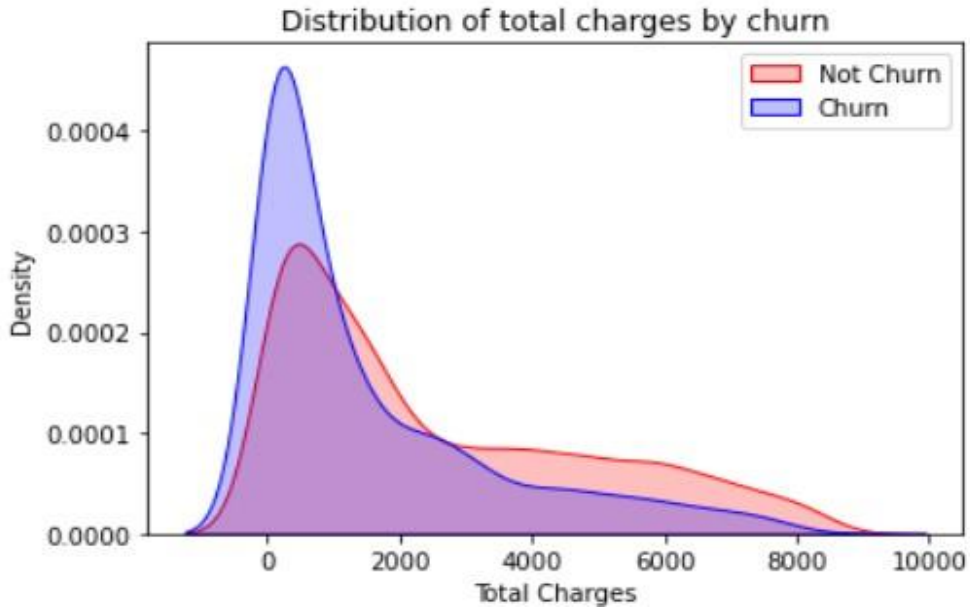

Figure 4.13: Distribution of total charges by churn

Interpretation: It seems that there is higher churn when the total-charges are lower. there is higher survival rate when the total charges are higher.

Findings

1) The majority of the company's contracts are month-to-month contracts.

2) The rate of customer churn is significant for people with month-to-month contracts. Almost four times higher than one-year contract and fourteen times higher than two-year contract.

3) Longer-term contracts are more likely to keep customers from switching sides.

4) Customers with higher monthly charges are more likely to churn than those with lower charges in all contract types

5) Customers who have been with the business for a while are less likely to leave.

6) It's interesting to observe that once total charges reach a certain amount, the probability of churning starts to decline.

Survival Analysis of customer based on there their contract type

```
fig_contract, ax_contract = plt.subplots(figsize=(12,7))
plot_categorical_survival(feature='Contract', ax=ax_contract)
ax_contract.set_title('Customer Churn by Contract Type')
ax_contract.set_xlabel('Customer Tenure (Months)')
ax_contract.set_ylabel('Customer Survival Chance (%)')
plt.show()
```

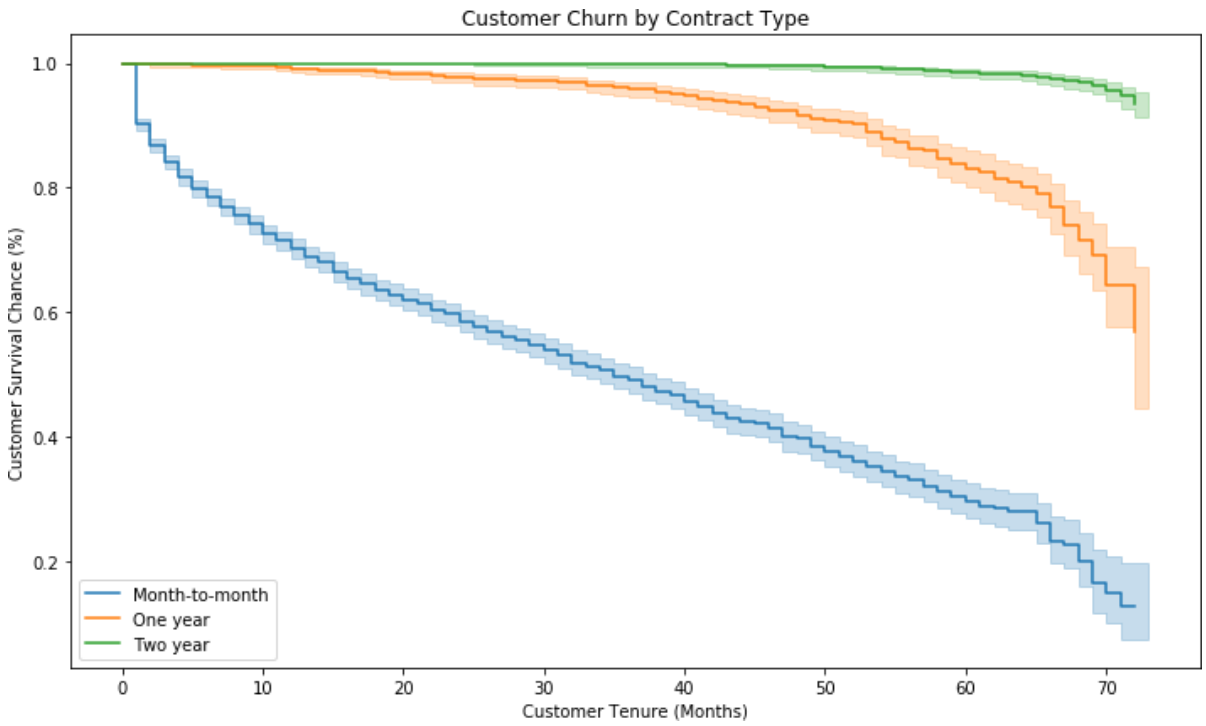

Figure 4.14: Survival Analysis of customer based on there their contract type

Interpretation: These are the survival curves for customers based on the predictor, 'contract type'. As we can notice: The probability of survival is the highest among customers whose contract period is two years. We can say with more than 98% probability that these customers will stay active (not churn) even after 50 months

For yearly contract customers, the probability of survival starts dropping after 50 months (after 4 years). It indicates a need to introduce some customer retention interventions to prevent their churn -For monthly contract customers, there's a significant drop immediately after a couple of months

Correlation Matrix

Correlations between customer data features and customer churn Examining the correlation between churn and each customer feature to decide which features of the data to include in the predictive churn model.

```
df['Churn'].replace(to_replace='Yes', value=1, inplace=True)
df['Churn'].replace(to_replace='No',  value=0, inplace=True)

# converting all the categorical variables into dummy variables
df_dummies = pd.get_dummies(df)
df_dummies.head()
```

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges	Churn	gender_Female
0	0	1	29.85	29.85	0	1
1	0	34	56.95	1889.50	0	0
2	0	2	53.85	108.15	1	0
3	0	45	42.30	1840.75	0	0
4	0	2	70.70	151.65	1	1

Correlation between Quantitative variables

```
# correlation between quantitative variables
plt.figure(figsize=(12, 6))
sns.heatmap(data=df[['tenure','MonthlyCharges','TotalCharges']].corr(),
            annot=True)

plt.title("Correlation between Quantitative variables",fontsize=16)
```

```
Text(0.5, 1.0, 'Correlation between Quantitative variables')
```

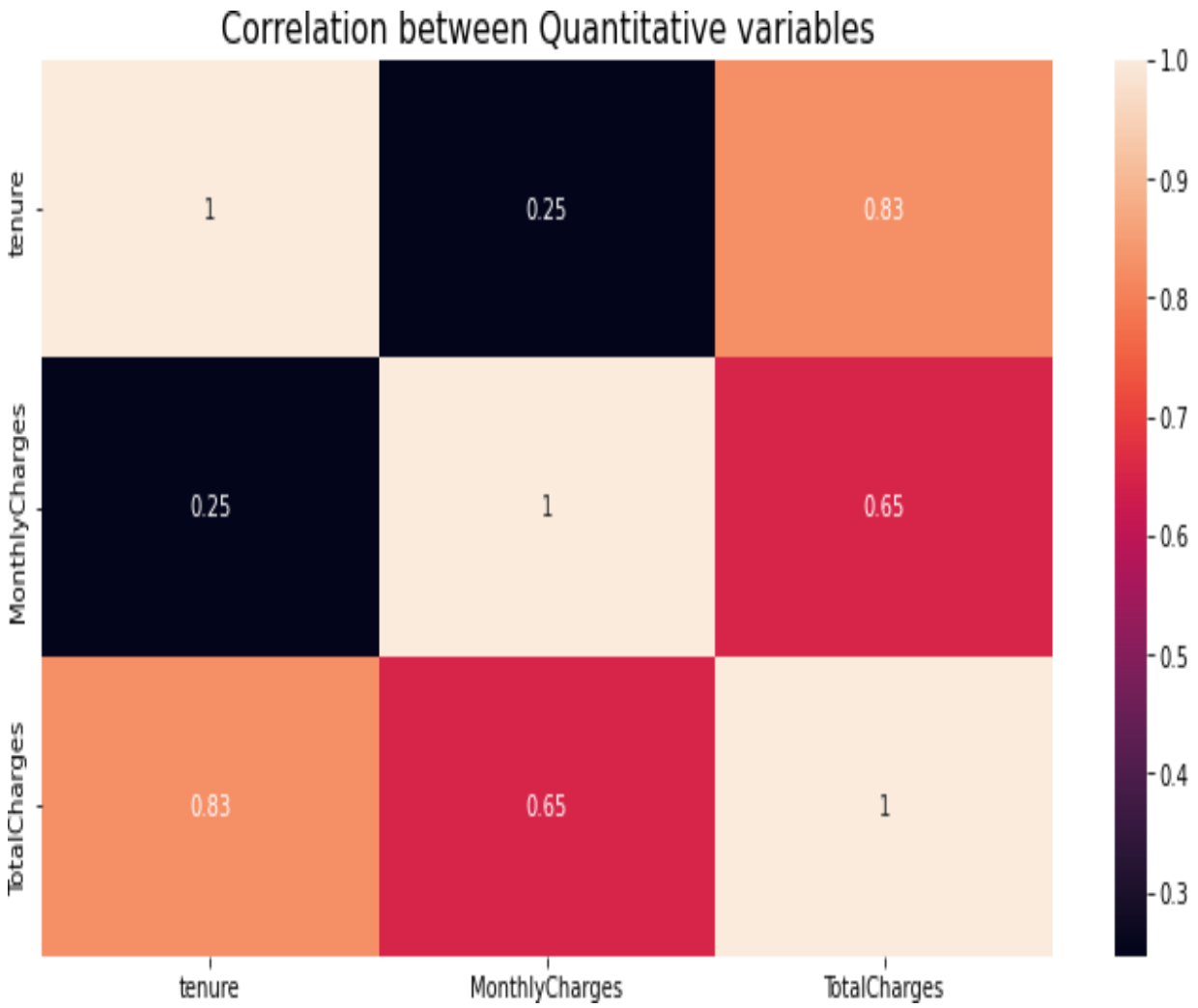

Figure 4.15: Correlation between Quantitative variables

Interpretation:

- Again, as expected, strong correlation exists between tenure and total charges, which means that as tenure increases so does total charges.
- Strong correlation exists between monthly charges and total charges as well.

Checking again Correlation of "Churn" with other variables on a different plot

```
# Checking again Correlation of "Churn" with other variables on a different plot

plt.figure(figsize=(15,8))
df_dummies.corr()['Churn'].sort_values(ascending = False).plot(kind='bar')
```

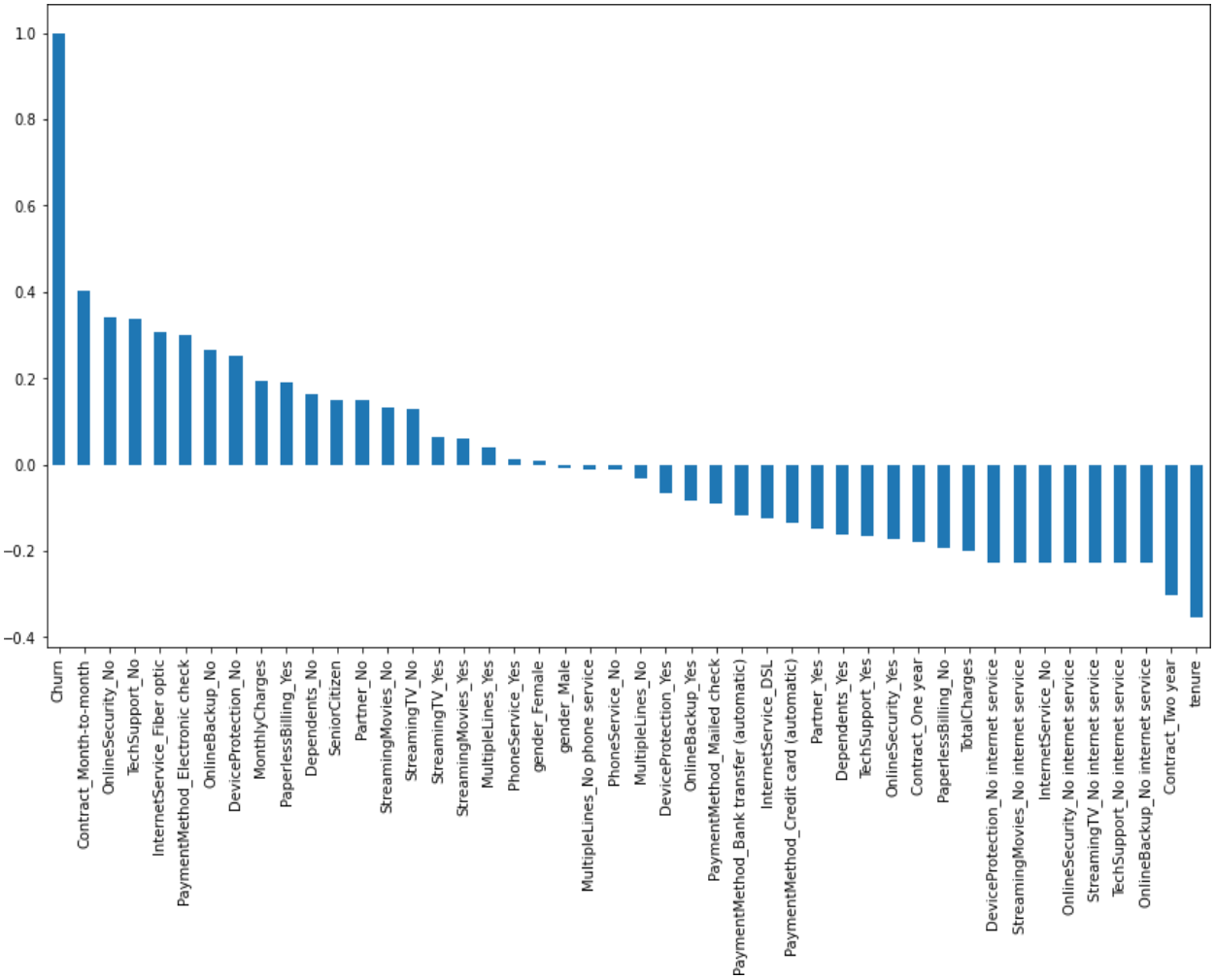

Figure 4.16: Checking again Correlation of "Churn" with other variables

Interpretation: Month to month contracts, absence of online security and tech support seem to be positively correlated with churn. While, tenure, two year contracts seem to be negatively correlated with churn.

Interestingly, services such as Online security, streaming TV, online backup, tech support, etc. without internet connection seem to be negatively related to churn.

We will explore the patterns more for the above correlations below before modelling and identifying the important variables.

FEATURE-IMPORTANCE

Mutual information — analysis of linear and nonlinear relationships

Mutual information measures the mutual dependency between two variables based on entropy estimations. In machine learning, we are interested in evaluating the degree of dependency between each independent variable and the response variable. Higher values of mutual information show a higher degree of dependency which indicates that the independent variable will be useful for predicting the target.

Feature Selection

```
## PLOT HEATMAP FOR ALL FEATURES
plt.figure(figsize=(25, 10))

corr = telco_data.apply(lambda x: pd.factorize(x)[0]).corr()

mask = np.triu(np.ones_like(corr, dtype=bool))

ax = sns.heatmap(corr, mask = mask, xticklabels = corr.columns, yticklabels =
                 annot = True, linewidths = .2)
```

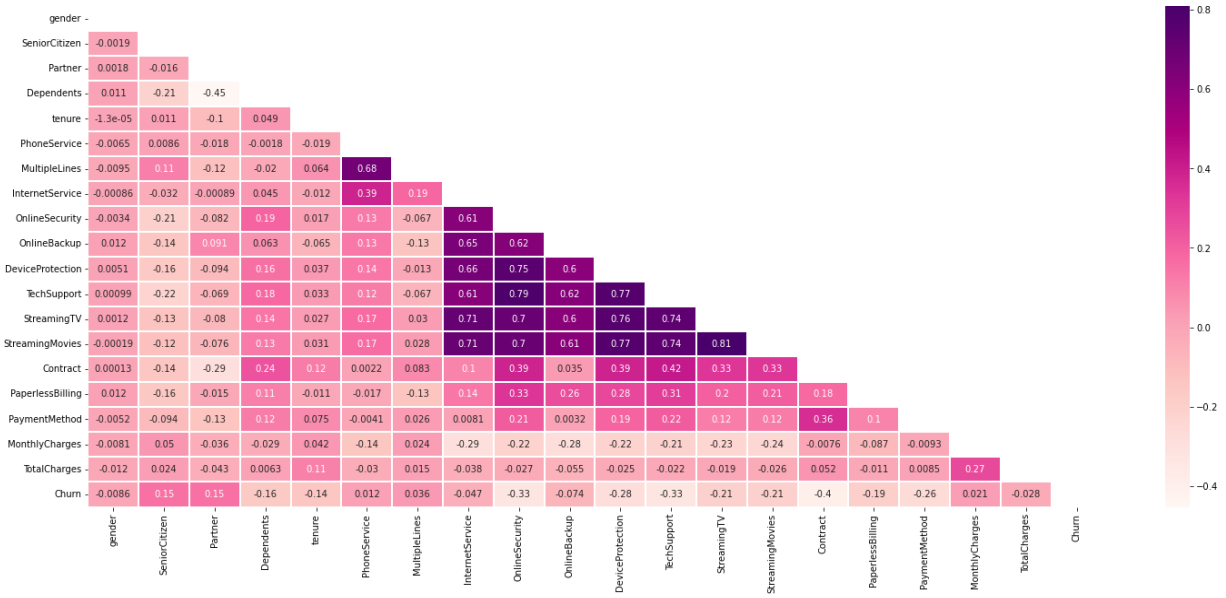

Figure 4.17: Feature Selection

```python
## FIND THE FEATURE IMPORTANCE
from sklearn.ensemble import RandomForestClassifier
import plotly.graph_objects as go

X = df_dummies.drop(columns=['Churn'])
y = df_dummies['Churn']
clf = RandomForestClassifier(n_estimators = 100,random_state = 0).fit(X,y)

impFeature = pd.Series(data = clf.feature_importances_, index = X.columns).sort_values(ascending = False)

fig = go.Figure(go.Bar(
            x = impFeature.values,
            y = impFeature.index,
            orientation = 'h'))
fig.show()
```

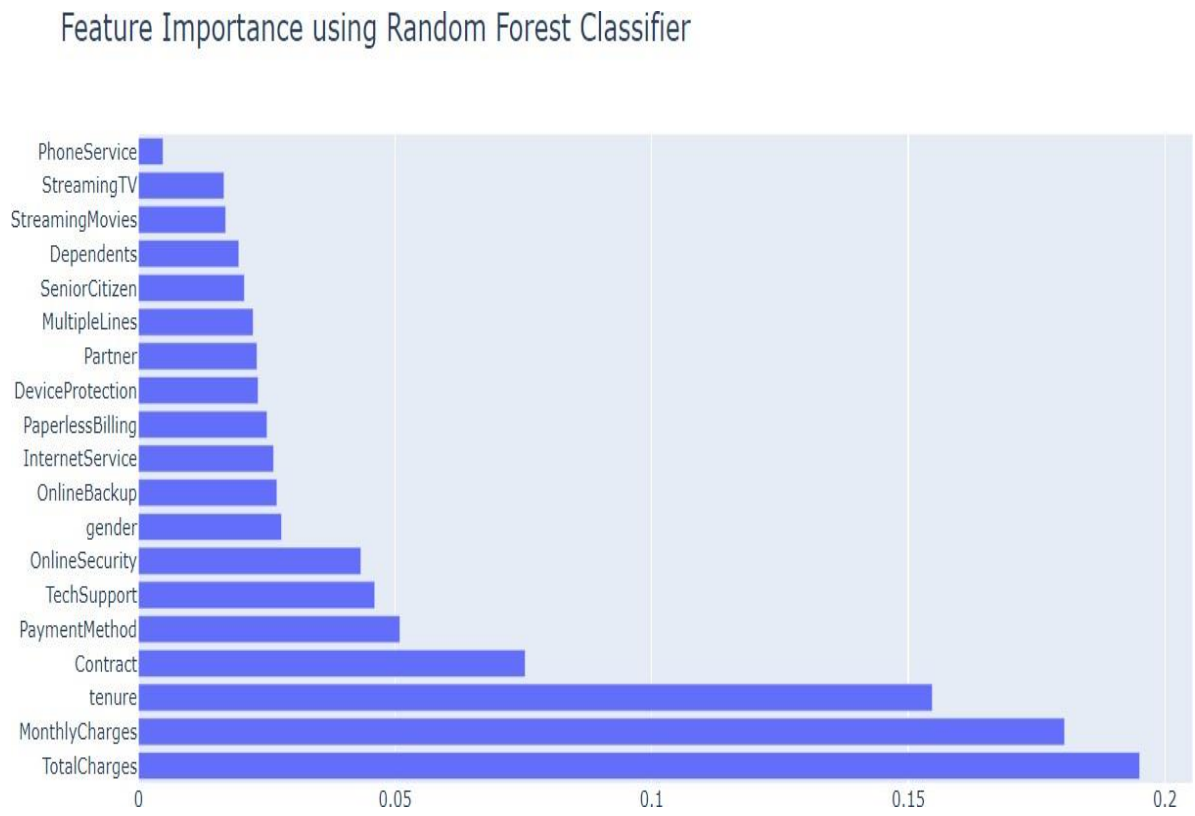

Figure 4.18: Feature Importance using Random Forest Classifier

Interpretation: According to the previous EDA, the heatmap and the feature importance chart, it was decided to exclude the phone service attribute as it had little relationship to the target attribute.

Feature Engineering

Feature engineering is the process of extracting features from the data and transforming them into a format that is suitable for the machine learning model. In this project, we need to transform both numerical and categorical variables.

Checking format consistency for categorical attributes

```
for i in df.columns:
    if df[i].dtypes=="object":
        print(f'{i} : {df[i].unique()}')
        print("*****************************************************")
```

```
*****************************************************
gender : ['Female' 'Male']
*****************************************************
Partner : ['Yes' 'No']
*****************************************************
Dependents : ['No' 'Yes']
*****************************************************
PhoneService : ['No' 'Yes']
*****************************************************
MultipleLines : ['No phone service' 'No' 'Yes']
*****************************************************
InternetService : ['DSL' 'Fiber optic' 'No']
*****************************************************
OnlineSecurity : ['No' 'Yes' 'No internet service']
*****************************************************
OnlineBackup : ['Yes' 'No' 'No internet service']
*****************************************************
DeviceProtection : ['No' 'Yes' 'No internet service']
*****************************************************
TechSupport : ['No' 'Yes' 'No internet service']
*****************************************************
StreamingTV : ['No' 'Yes' 'No internet service']
*****************************************************
StreamingMovies : ['No' 'Yes' 'No internet service']
*****************************************************
Contract : ['Month-to-month' 'One year' 'Two year']
*****************************************************
PaperlessBilling : ['Yes' 'No']
*****************************************************
PaymentMethod : ['Electronic check' 'Mailed check' 'Bank transfer (automatic)'
 'Credit card (automatic)']
*****************************************************
Churn : ['No' 'Yes']
*****************************************************
```

Label Encoding

Label encoding is used to replace categorical values with numerical values. This encoding replaces every category with a numerical label. In this project, we use label encoding with the following binary variables: (1) gender, (2) Partner, (3) Dependents, (4)PaperlessBilling, (5)PhoneService , and (6)Churn .

```
df2 = df.copy()

# label encoding (binary variables)
label_encoding_columns = ['gender', 'Partner', 'Dependents',
                          'PaperlessBilling', 'PhoneService', 'Churn']

# encode categorical binary features using label encoding
for column in label_encoding_columns:
    if column == 'gender':
        df2[column] = df2[column].map({'Female': 1, 'Male': 0})
    else:
        df2[column] = df2[column].map({'Yes': 1, 'No': 0})
```

One-Hot Encoding

One-hot encoding creates a new binary column for each level of the categorical variable. The new column contains zeros and ones indicating the absence or presence of the category in the data.In this project, we apply one-hot encoding to the following categorical variables:

(1) Contract, (2) PaymentMethod, (3) MultipleLines, (4) InternetServices, (5) OnlineSecurity, (6) OnlineBackup, (7) DeviceProtection, (8) TechSupport, (9) StreamingTV, and (10)StreamingMovies.

```
#One-Hot Encoding
# one-hot encoding (categorical variables with more than two levels)
one_hot_encoding_columns = ['MultipleLines', 'InternetService', 'OnlineSecurit
                            , 'OnlineBackup', 'DeviceProtection',
                            'TechSupport', 'StreamingTV',
                            'StreamingMovies', 'Contract', 'PaymentMethod']

# encode categorical variables with more than two levels using one-hot encodin
df2 = pd.get_dummies(df2, columns = one_hot_encoding_columns)
```

Normalization

Data Normalization is a common practice in machine learning which consists of transforming numeric columns to a common scale. Data normalization transforms multiscale data to the same scale. After normalization, all variables have a similar influence on the model, improving the stability and performance of the learning algorithm.

There are multiple normalization techniques in statistics. In this project, we will use the min-max method to rescale the numeric columns (tenure, MontlyCharges, and TotalCharges) to a common scale. The min-max approach (often called normalization) rescales the feature to a fixed range of [0,1] by subtracting the minimum value of the feature and then dividing by the range.

```
#Normalization
# min-max normalization (numeric variables)
min_max_columns = ['tenure', 'MonthlyCharges', 'TotalCharges']

# scale numerical variables using min max scaler
for column in min_max_columns:
        # minimum value of the column
        min_column = df2[column].min()
        # maximum value of the column
        max_column = df2[column].max()
        # min max scaler
        df2[column] = (df2[column] - min_column) / (max_column - min_column)
```

Machine learning algorithms

We will consider several different models to predict customer churn. To ensure we are not over-fitting to our data, we will split the 7,032 customer records into a training and test set, with the test set being 20% of the total records and the training set being 80% of the total records. It is important to scale the variables in regression so that all of them are within a range of 0 to 1.

Splitting the data in training and testing sets

Split the data set as train and test with %20-%80 ratio.

X is the data with the independent variables, Y is the data with the dependent variable. The test size variable determines in which ratio the data will be split. It is quite common to do this in a 80 Training / 20 Test ratio. Also ned to stratify the train-test-split to have a balanced split

```
# select independent variables
X = df2.drop(columns='Churn')

# select dependent variables
y = df2.loc[:, 'Churn']

# prove that the variables were selected correctly
print(X.columns)

# prove that the variables were selected correctly
print(y.name)
```

Then, we can use the train_test_split function from the sklearn.model_selection package to create both the training and testing sets.

```
# split the data in training and testing sets
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20,
                                                    random_state=40, shuffle=True)

print('length of X_train and x_test: ', len(X_train), len(X_test))
print('length of y_train and y_test: ', len(y_train), len(y_test))
```

```
length of X_train and x_test:  5625 1407
length of y_train and y_test:  5625 1407
```

Logistics Regression

Using machine learning algorithm and the dependent variable here churn 1 or churn 0 is categorical. The trained model can be used to predict if a customer churned or not for the test dataset.

1). Importing Logistic Regression

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix,
        accuracy_score, classification_report
```

2). creating object for Logistics Regression model

```
model_lg = LogisticRegression(max_iter=120,random_state=0, n_jobs=20)
```

3). Logistics Regression Model Training using training dataset

```
model_lg.fit(X_train, y_train)
```

```
LogisticRegression(max_iter=120, n_jobs=20, random_state=0)
```

4). Making Predictions using testing data set

```
pred_lg = model_lg.predict(X_test)
```

5). Calculating Accuracy of theLogistics Regression model

```
lg = round(accuracy_score(y_test, pred_lg)*100,2)
print("Logistic Regression accuracy is :" ,lg)
```

```
Logistic Regression accuracy is : 79.82
```

6). Classification Report of the Logistics Regression model

```
print(classification_report(y_test, pred_lg))
```

```
              precision    recall  f1-score   support

           0       0.86      0.88      0.87      1050
           1       0.61      0.56      0.59       357

    accuracy                           0.80      1407
   macro avg       0.73      0.72      0.73      1407
weighted avg       0.79      0.80      0.80      1407
```

```
import seaborn as sns
import matplotlib.pyplot as plt
# confusion Maxtrix
cm1 = confusion_matrix(y_test, pred_lg)
ax= plt.subplot()
sns.heatmap(cm1, annot=True, fmt='g', ax=ax)
# labels, title and ticks
ax.set_xlabel('Predicted labels',fontsize=14)
ax.set_ylabel('Actual Values',fontsize=14)
ax.set_title('Logistic Regression Confusion Matrix',fontsize=16)
ax.xaxis.set_ticklabels(['Not churn ', 'Churn'],fontsize=12)
ax.yaxis.set_ticklabels(['Not churn', 'Churn'],fontsize=12)
```

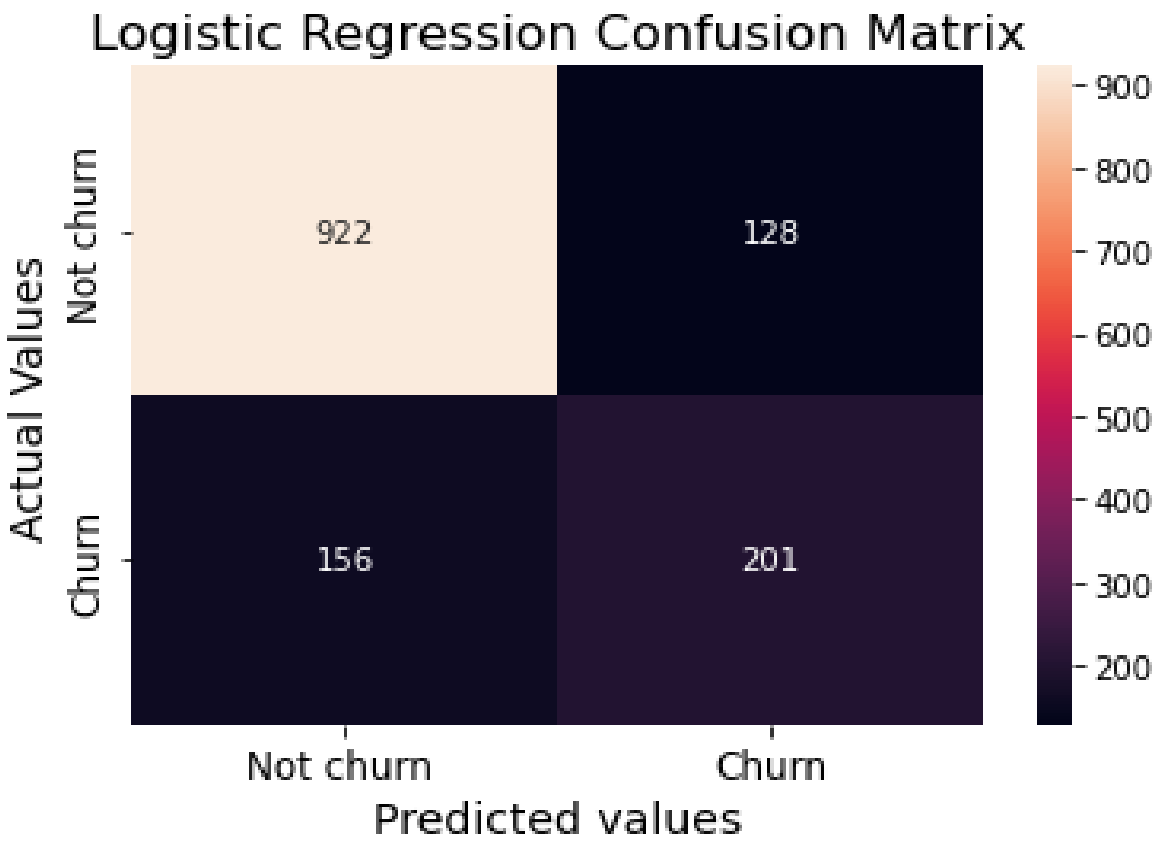

Figure 4.19: Logistic Regression Confusion Matrix

Interpretation: From the Logistics Regression confusion matrix we can see that: There are total 922+128=1050 actual not-churn values and the algorithm predicts 922 of them as not-churn and 128 of them as churn. While there are 156+201=357 actual churn values and the algorithm predicts 156 of them as not-churn values and 201 of them as churn values. The Logistics Regression model give the accuracy score of 79.82%.

2. Decision Tree

1). importing decision Tree

```
from sklearn.tree import DecisionTreeClassifier
```

2). Creating object of the decision Tree model

```
model_dt = DecisionTreeClassifier(max_depth=4, random_state=42)
```

3). Decision Tree Model Training using training dataset

```
model_dt.fit(X_train, y_train)
```

```
DecisionTreeClassifier(max_depth=4, random_state=42)
```

4). Making Predictions using testing data set

```
pred_dt = model_dt.predict(X_test)
```

5). Calculating Accuracy of the decision Tree model

```
dt  = round(accuracy_score(y_test, pred_dt)*100, 2)
print("Decision Tree accuracy is",dt)
```

```
Decision Tree accuracy is 78.75
```

6). Classification Report of the decision Tree model

```
print(classification_report(y_test, pred_dt))
```

```
              precision    recall  f1-score   support

           0       0.84      0.88      0.86      1050
           1       0.59      0.51      0.55       357

    accuracy                           0.79      1407
   macro avg       0.72      0.70      0.71      1407
weighted avg       0.78      0.79      0.78      1407
```

```
import seaborn as sns
import matplotlib.pyplot as plt
# confusion Maxtrix
cm2 = confusion_matrix(y_test, pred_dt)
ax= plt.subplot()
sns.heatmap(cm2, annot=True, fmt='g', ax=ax)
# labels, title and ticks
ax.set_xlabel('Predicted values',fontsize=14)
ax.set_ylabel('Actual Values',fontsize=14)
ax.set_title(' Decision Tree Confusion Matrix',fontsize=16)
ax.xaxis.set_ticklabels(['Not churn ', 'Churn'],fontsize=12)
ax.yaxis.set_ticklabels(['Not churn', 'Churn'],fontsize=12)
```

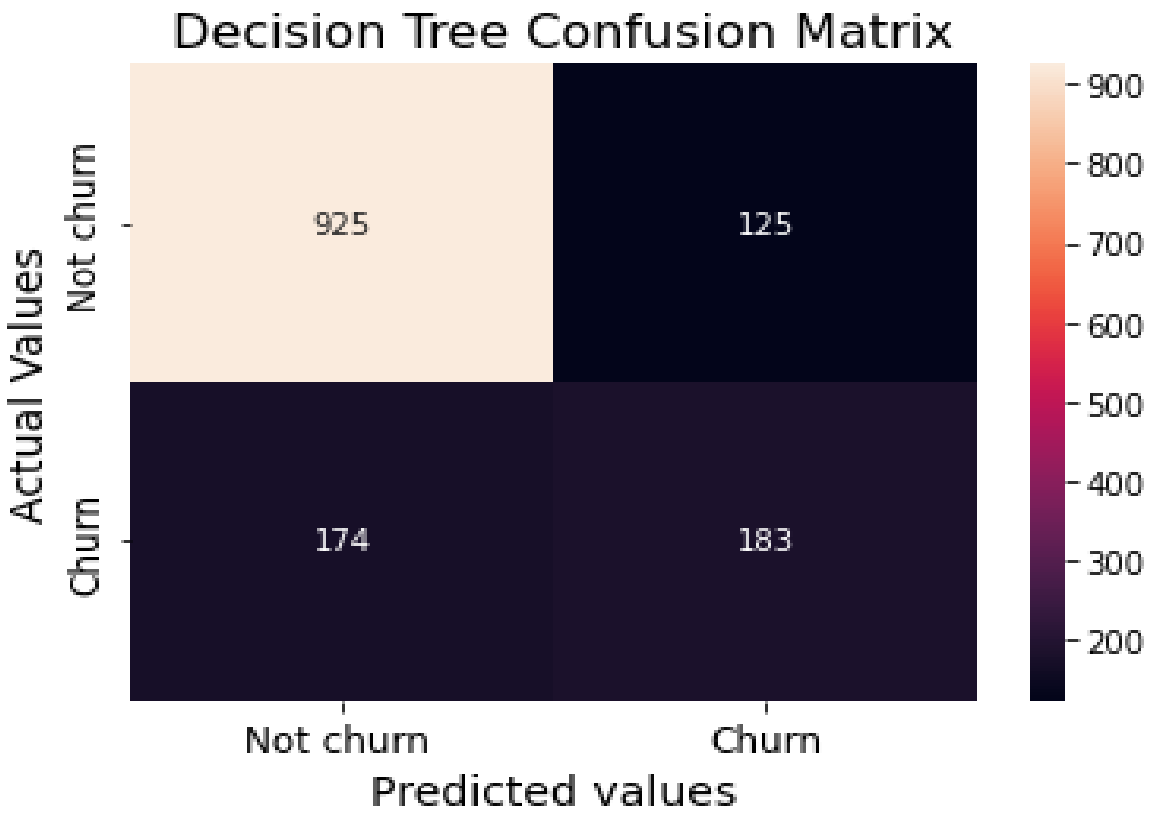

Figure 4.20: Decision Tree Confusion Matrix

Interpretation: From the Decision Tree confusion matrix we can see that: There are total 925+125=1050 actual not-churn values and the algorithm predicts 925 of them as not churn or stayed in the services and 125 of them as churn. While there are 174+183=357 actual churn values and the algorithm predicts 174 of them as not-churn values and 183 of them as churn values. The Decision Tree model give the accuracy of score 78.75%.

Random Forest

1). importing Random Forest

```
from sklearn.ensemble import RandomForestClassifier
```

2). Creating object of the Random Forest model

```
# Creating model object
model_rf = RandomForestClassifier(n_estimators=300,min_samples_leaf=0.16,
                                  random_state=42)
```

3). Random Forest Model Training using training dataset

```
# Training Model
model_rf.fit(X_train, y_train)
```

```
RandomForestClassifier(min_samples_leaf=0.16, n_estimators=300, random_state=42)
```

4). Making Predictions using testing data set

```
# Making Prediction
pred_rf = model_rf.predict(X_test)
```

5). Calculating Accuracy of the Random Forest model

```
# Calculating Accuracy Score
rf = round(accuracy_score(y_test, pred_rf)*100, 2)
print('Random Forest model accuracy is ',rf)
```

```
Random Forest model accuracy is  74.63
```

6). Classification Report of the Random Forest model

```
print(classification_report(y_test,pred_rf))
```

```
              precision    recall  f1-score   support

           0       0.75      1.00      0.85      1050
           1       0.00      0.00      0.00       357

    accuracy                           0.75      1407
   macro avg       0.37      0.50      0.43      1407
weighted avg       0.56      0.75      0.64      1407
```

```
import seaborn as sns
import matplotlib.pyplot as plt
# confusion Maxtrix
cm3 = confusion_matrix(y_test, pred_rf)
ax= plt.subplot()
sns.heatmap(cm3, annot=True, fmt='g', ax=ax)
# labels, title and ticks
ax.set_xlabel('Predicted values',fontsize=14)
ax.set_ylabel('Actual Values',fontsize=14)
ax.set_title('Random Forest Confusion Matrix',fontsize=16)
ax.xaxis.set_ticklabels(['Not churn ', 'Churn'],fontsize=12)
ax.yaxis.set_ticklabels(['Not churn', 'Churn'],fontsize=12)
```

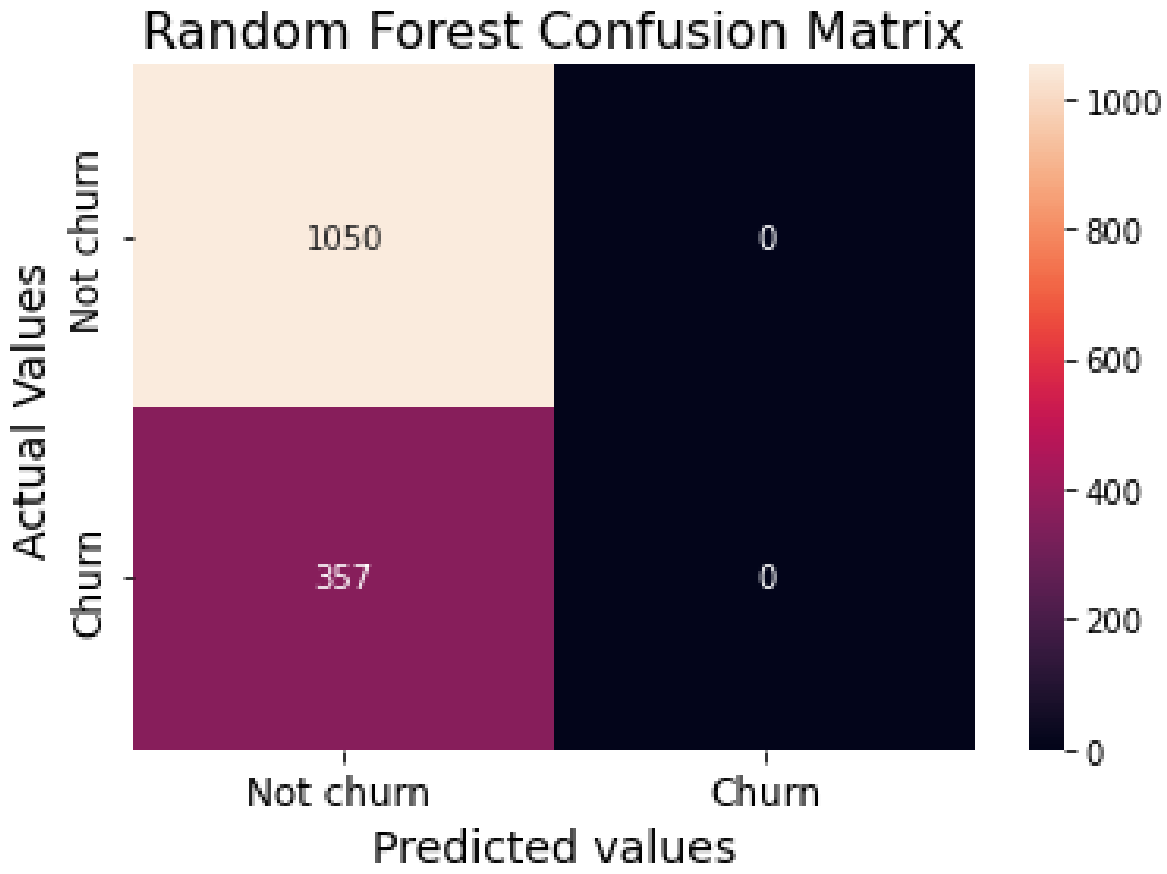

Figure 4.21: Random Forest Confusion Matrix

Interpretation: From the Random Forest confusion matrix we can see that: There are total 1050+0=1050 actual non-churn values and the algorithm predicts 1050 of them as not churn or stayed in the services and 0 of them as churn. While there are 357+0=561 actual churn values and the algorithm predicts 357 of them as non-churn values and 0 of them as churn values. The Random Forest model give the accuracy of 74.63%.

Machine learning model comparison to see which has the best Accuracy score

```
models = pd.DataFrame({
    'Model':['Logistic Regression', 'Decision Tree', 'Random Forest'],
    'Accuracy_score' :[lg, dt, rf]
})
models
sns.barplot(x='Accuracy_score', y='Model', data=models)

models.sort_values(by='Accuracy_score', ascending=False)
```

	Model	Accuracy_score
0	Logistic Regression	79.82
1	Decision Tree	78.75
2	Random Forest	74.63

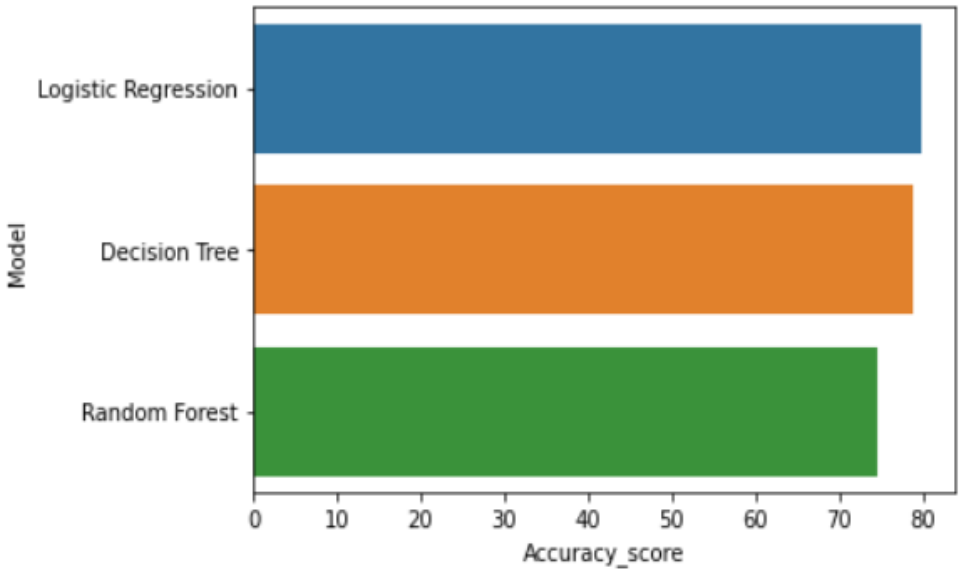

Figure 4.22: Machine learning model comparison

Interpretation: Above graph shows the performance and comparison of different models used in this project. Having created our models with default and optional parameters, Logistic Regression (LR) had the best accuracy score of (79.82%) among the other models all most 80%. The decision tree has the second highest accuracy score of (78.75%) . compared to this two-model random forest has less accuracy score (74.63%).

Chapter 5:

Findings, Conclusions and Suggestions

Summary of Findings

- One of the most significant problems confronting the telecom industry today is customer churn. Recognizing potential customer churn is essential for avoiding losses.
- In the telecom sector, keeping consumers is a crucial task because doing so lowers customer churn by raising customer satisfaction.
- The ability to predict churn is much more important because acquiring new customers has far higher expenses than keeping existing ones.
- Customer churn has a negative impact on a company's profitability. There are numerous tactics that can be used to reduce client churn. Knowing a company's customers well is the best strategy to prevent customer churn.
- In the telecom industry, machine learning makes churn prediction much easier. The machine learning can help in understanding the customers who might churn or survive in their service.
- The scope of this project is to find the relationships between the available input attributes and whether the customer churned, and select the most important attributes that can be utilised by supervised machine learning models and find the best model to predict if a customer will STAY or LEAVE the company, so the company can take corrective action.
- 73.4% of the data in the dataset is recorded as not churning customers and only 26.6% of the records are churning customers.
- Customers with month-to-month contracts have a relatively high likelihood of churning.
- Long-term customers are more loyal and less likely to churn.
- - About half of the customers in our data set are male(50.50%) while the other half are female (49.5%).
- The churn rate of senior citizens is almost double that of young citizens.
- Month-to-month contracts are the most dominant contract for the company.
- Customers in month-to-month contracts have a high churn rate. Almost four times higher than one-year contract and fourteen times higher than two-year contract.
- Customers who sign longer contracts are less likely to leave the company.
- Customers with higher monthly charges are more likely to churn than those with lower charges in all contract types

- The probability of survival is the highest among customers whose contract period is two years. We can say with more than 98% probability that these customers will stay active (not churn) even after 50 months
- The best way to avoid customer churn is for a company to truly know its customers. This includes identifying customers who are at risk of churning and working to improve their satisfaction. Improving customer service is, of course, at the top of the priority for tackling this issue.

Conclusion

As the telecommunications industry continues to grow, the problem of customer churn has significant growth as well. Retaining customers is a critical challenge in the telecommunications industry since it reduces customer churn through increased customer satisfaction. With a subscription-based business model, customer retention is one of the key performance indicators for a telecommunication company. Competition is tough in the industry as customers can choose from a variety of providers.

The use of predictive analytics can help tackle this threat by identifying vulnerable customers and implementing customer-centric retention measures. The proposed prediction analysis can be solved with the help of machine learning models.

- IBM telecom dataset consists of 7033 observations, 21 attribute
- Several techniques and methods have been applied across the entire supervised machine learning process and used machine learning models to predict.
- Data exploration was performed across 20 features and the target variable (churn). A feature engineering process was applied including feature encoding, feature selection, and feature importance.
- Three machine learning models logistics regression, decision tree, Random Forest implemented with the python to predicting customer churn
- To better understand the performance of the models, confusion matrices, classification accuracy, precision, recall, F1-score were applied.

- The obtained results show that our proposed churn model produced better results and performed better by using machine learning techniques. logistics regression, |produced better accuracy score around 80% among the other models.
- As noticed that the churned customers have some similar services, which means that any telecom company can detect the predictors and retain their customers.
- Since the costs related with obtaining a new customer is much greater than retaining older customer the prediction of churn becomes even more essential

Suggestions

- Telecom service providers must spend on 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 model can be deployed to run on a regular basis to understand the changes in the behaviour of the customers
- Collect more data from customers going forward such as reviews, ratings etc. to understand more about why a certain group of customers might be leaving – Opportunity to apply Natural Language processing to free form textual reviews.
- To detect early signs of potential churn, one must first develop a holistic view of the customers and their interactions across numerous channels, including store/branch visits, product purchase histories, customer service calls, Web-based transactions, and social media interactions, to mention a few.
- This includes identifying customers who are at risk of churning and working to improve their satisfaction. Improving customer service is, of course, at the top of the priority for tackling this issue.
- Building customer loyalty through relevant experiences and specialized service is another strategy to reduce customer churn.
- A company could provide some targeted incentives such as exclusive promotional pricing or gift vouchers to those high-risk customers, in order to keep them for another one or two years and extend their lifetime value to the company.

- Some firms survey customers who have already churned to understand their reasons for leaving in order to adopt a proactive approach to avoiding future customer churn.
- Churn rate is an important indicator for subscription-based companies. Identifying customers who aren't happy can help managers identify product or pricing plan weak points, operation issues, as well as customer preferences and expectations. When you know all that, it's easier to introduce proactive ways of reducing churn.

Bibliography

Ahmed, A., & Linen, D. M. (2017, January). A review and analysis of churn prediction methods for customer retention in telecom industries. In *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 1-7). IEEE.

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research*, *94*, 290-301.

Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016, March). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. In *2016 Symposium on Colossal Data Analysis and Networking (CDAN)* (pp. 1-4). IEEE.

Hota, L., & Dash, P. K. (October 2021) Prediction of Customer Churn in Telecom Industry: A Machine Learning Perspective.

Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost. *Procedia Computer Science*, *167*, 101-112.

Jain, H., Yadav, G., & Manoov, R. (2021). Churn prediction and retention in banking, telecom and IT sectors using machine learning techniques. In Advances in Machine Learning and Computational Intelligence (pp. 137-156). Springer, Singapore.

Kumar, M., & Kumar Yadav, D. (2022). Customer Churn Prediction in Telecommunication Using Gradient Boosting Machine. In *International Conference on Innovative Computing and Communications* (pp. 769-776). Springer, Singapore.

Rani, K. S., Thaslima, S., Prasanna, N. G. L., Vindhya, R., & Srilakshmi, P. (2021). Analysis of Customer Churn Prediction in Telecom Industry Using Logistic Regression. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN*, 2347-5552.

Singh, M., Singh, S., Seen, N., Kaushal, S., & Kumar, H. (2018, November). Comparison of learning techniques for prediction of customer churn in telecommunication. In *2018 28th International Telecommunication Networks and Applications Conference (ITNAC)* (pp. 1-5). IEEE.

Sai, B. K., & Sasikala, T. (2019, April). Predictive analysis and modeling of customer churn in telecom using machine learning technique. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 6-11). IEEE.

Reference Website

https://www.kaggle.com/blastchar/telco-customer-churn

https://www.wikipedia.org/

https://www.ibef.org/industry/telecommunications

https://www.investindia.gov.in/sector/telecom

https://www.thebusinessresearchcompany.com/report/telecom-global-market-report

https://www.grandviewresearch.com/industry-analysis/global-telecom-services-market

A Study On Survival Analysis and Churn Prediction Using Machine Learning Model for Selected Telecom Companies

ORIGINALITY REPORT

18%	10%	4%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	2%
2	www.kaggle.com Internet Source	1%
3	towardsdatascience.com Internet Source	1%
4	Submitted to University of Teesside Student Paper	1%
5	Submitted to University of Wales Institute, Cardiff Student Paper	1%
6	Submitted to Bournemouth University Student Paper	<1%
7	www.fortunebusinessinsights.com Internet Source	<1%
8	www.ijres.org Internet Source	<1%

9	Submitted to Queen Mary and Westfield College Student Paper	<1 %
10	Submitted to Christ University Student Paper	<1 %
11	Submitted to University of Southampton Student Paper	<1 %
12	Submitted to University of Bedfordshire Student Paper	<1 %
13	Submitted to University of Witwatersrand Student Paper	<1 %
14	scirp.org Internet Source	<1 %
15	Submitted to Coventry University Student Paper	<1 %
16	dataskunkworks.com Internet Source	<1 %
17	www.abacademies.org Internet Source	<1 %
18	Submitted to Indian School of Business Student Paper	<1 %
19	Submitted to Tias Business School B.V. Student Paper	<1 %
20	lib.dr.iastate.edu	

	Internet Source	<1 %
21	www.javatpoint.com Internet Source	<1 %
22	Submitted to Universiti Teknikal Malaysia Melaka Student Paper	<1 %
23	Submitted to Georgia Institute of Technology Main Campus Student Paper	<1 %
24	Submitted to Institute of Management Technology Student Paper	<1 %
25	Submitted to Technological Institute of the Philippines Student Paper	<1 %
26	Submitted to Galileo Global Education Student Paper	<1 %
27	Khulood Ebrah, Selma Elnasir. "Churn Prediction Using Machine Learning and Recommendations Plans for Telecoms", Journal of Computer and Communications, 2019 Publication	<1 %
28	Submitted to Liverpool John Moores University Student Paper	<1 %

29	Hemlata Jain, Ajay Khunteta, Sumit Srivastava. "Churn Prediction in Telecommunication using Logistic Regression and Logit Boost", Procedia Computer Science, 2020 Publication	<1 %
30	Submitted to Texas A&M University, College Station Student Paper	<1 %
31	Submitted to University of Central Lancashire Student Paper	<1 %
32	"International Conference on Innovative Computing and Communications", Springer Science and Business Media LLC, 2022 Publication	<1 %
33	Submitted to Taylor's Education Group Student Paper	<1 %
34	Submitted to The Robert Gordon University Student Paper	<1 %
35	Submitted to The University of Manchester Student Paper	<1 %
36	Submitted to Harrisburg University of Science and Technology Student Paper	<1 %
37	Submitted to Northcentral Student Paper	<1 %

38	Predictive Analytics with Microsoft Azure Machine Learning, 2015. Publication	<1 %
39	Submitted to SP Jain School of Global Management Student Paper	<1 %
40	Submitted to University of Strathclyde Student Paper	<1 %
41	Submitted to International Management Institute Student Paper	<1 %
42	Submitted to Nanyang Polytechnic Student Paper	<1 %
43	Submitted to Center for Study of Science, Technology and Policy Student Paper	<1 %
44	www.coursehero.com Internet Source	<1 %
45	Submitted to Ahmedabad University Student Paper	<1 %
46	Submitted to University of Stirling Student Paper	<1 %
47	openaccess.hku.edu.tr Internet Source	<1 %
48	Submitted to Sheffield Hallam University	

49	Submitted to University of Adelaide Student Paper	<1 %
50	Submitted to University of York Student Paper	<1 %
51	Pallavi Kamath, Pallavi Patil, Shrilatha S, Sushma, Sowmya S. "Crop yield forecasting using data mining", Global Transitions Proceedings, 2021 Publication	<1 %
52	www.grandviewresearch.com Internet Source	<1 %
53	www.hardware-depot-online.com Internet Source	<1 %
54	www.techgig.com Internet Source	<1 %
55	doaj.org Internet Source	<1 %
56	Jared Mathews, Prosenjit Chatterjee, Shankar Banik. "CoAP-DoS: An IoT Network Intrusion Data Set", 2022 6th International Conference on Cryptography, Security and Privacy (CSP), 2022 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		