TELCO CUSTOMER CHURN ANALYSIS USING MACHINE LEARNING ALGORITHMS

Introduction:

Data Science is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes.

It helps us to discover hidden patterns from the raw data.

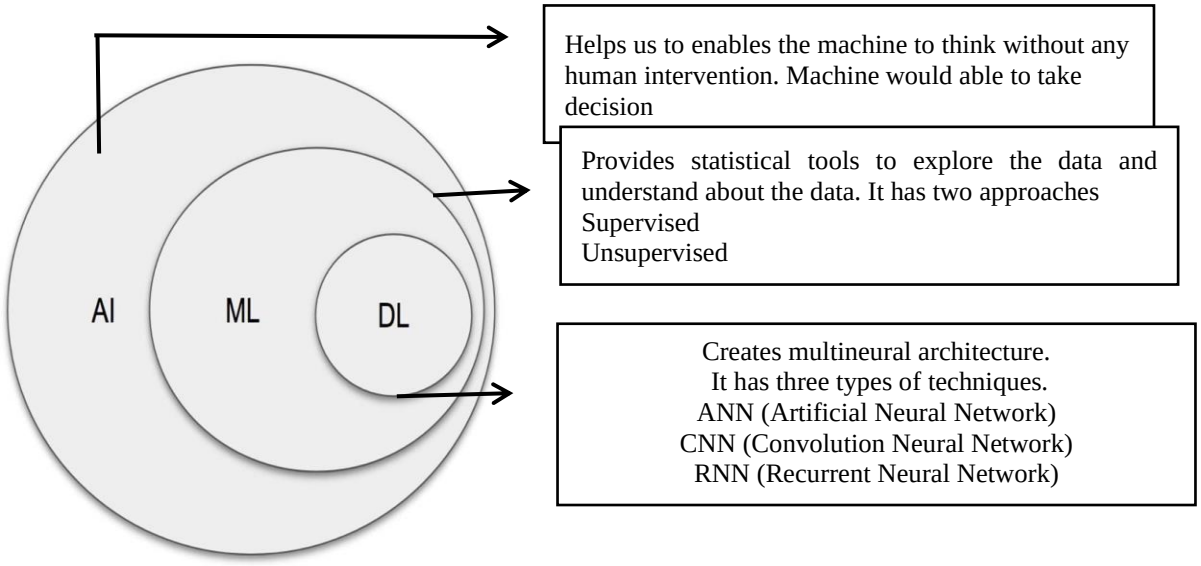
The term Data Science has emerged because of the evolution of mathematical statistics, data analysis, and big data.

Data Science is an interdisciplinary field that allows you to extract knowledge from structured or unstructured data.

Data Science enables us to translate a business problem into a research project and then translate it back into a practical solution.

The following diagram shows the

Artificial Intelligence Vs Machine Learning Vs Deep Learning



Now where DS (Data Science) fit into this?

It's a technique which try to apply all those techniques ML, DL apart form this it also uses some tools like mathematics, statistics, probability and Linear Algebra. Data

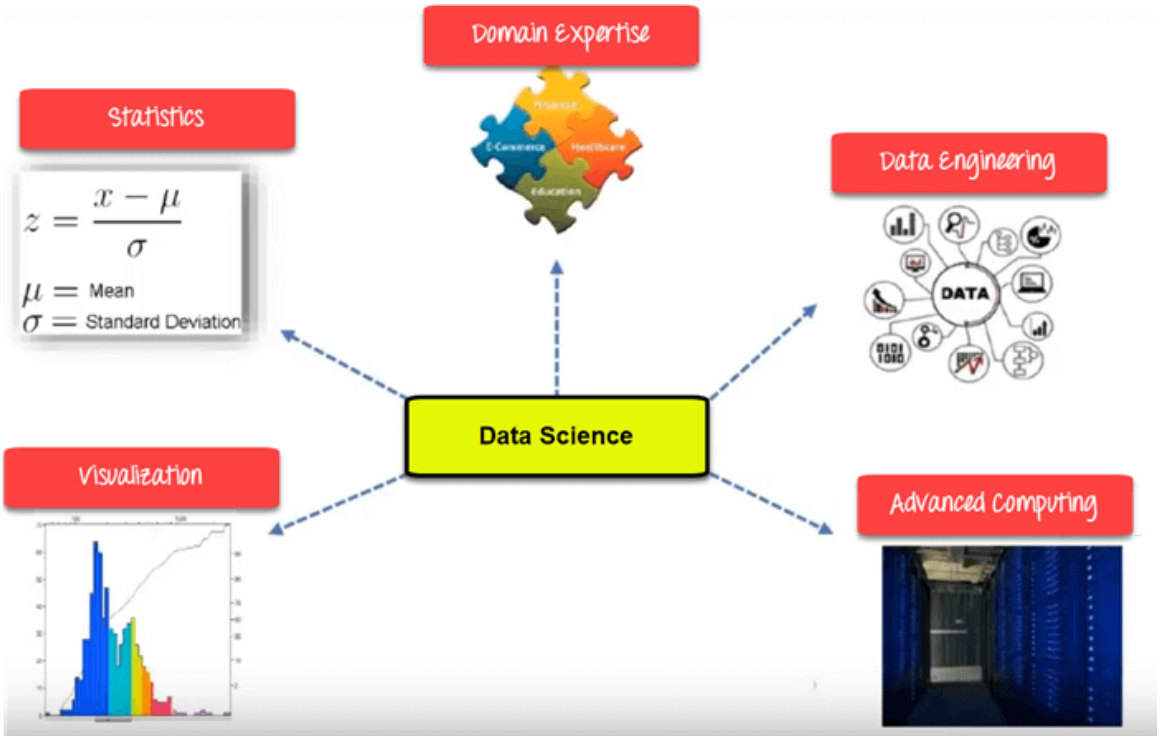
Scientist has to work on ML, DL by using mathematical or statistical and Linear Algebra etc.

Data Science is changing the world for the better.

Data and analytics are used everyday to help business drive efficiencies, glean deeper operational insights and ultimately generate more revenue. However, the impact of data science reaches far beyond the business section and is helping to solve some of the mankind's most pressing issues.

From preventing blindness and treating drug and alcohol addiction to fighting poverty, data science is being utilized not only as a business tool - but for the greater good of society.

Data Science Components



Statistics

Statistics is the most critical unit in the Data Science. It is the method or science of collecting and analyzing numerical data in large quantities to get useful insights.

A Data Scientist analyzes the data through various statistical procedures. In particular, two types of procedures used are:

- Descriptive Statistics
- Inferential Statistics

Descriptive Statistics:

Descriptive Statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data. Descriptive Statistics do not, however, allow us to make conclusions beyond the data we have analyzed or reach conclusions regarding any hypothesis we might have made. They are simply a way to describe our data.

Inferential Statistics:

With inferential statistics, we try to reach conclusions that extend beyond the immediate data alone.
For instance, we use inferential statistics to try to infer from the sample data what the population might think or we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in the study.
Thus we use inferential statistics to make inferences from our data to more general conditions.

Visualization

Visualization technique helps you to access huge amounts of data in easy to understand and meaningful visuals.

Machine Learning

Machine Learning explores the building and study of algorithms which learn to make predictions about unforeseen/future data.

Deep Learning

Deep Learning method is new machine learning research where the algorithms selects the analysis model to follow.

With the emergence of new technologies, there has been an exponential increase in data. This has created an opportunity to analyze and derive meaningful insights from data.
It requires special expertise of a ‘Data Scientist’ who can use various statistical & machine learning tools to understand and analyze data.

A Data Scientists, specializing in Data Science, not only analyzes the data but also uses machine learning algorithms to predict future occurrences of an event.

Therefore, we can understand Data Science as a field that deals with data processing, analysis, and extraction of insights from the data using various statistical methods and computer algorithms.

It is a multidisciplinary field that combines mathematics, statistics, and computer science.

Orange Tool For Data Science:

Orange is an open source data visualization and analysis tool. It features a visual programming front-end for explorative rapid qualitative data analysis and interactive data visualization.

Description:

Orange is a component-based visual programming software package for data visualization, machine learning, data mining, and data analysis.

Orange components are called widgets and they range from simple data visualization, subset selection, and preprocessing to empirical evaluation of learning algorithms and predictive modeling.

Visual programming is implemented through an interface in which workflows are created by linking predefined or user-designed widgets, while advanced users can use Orange as a Python library for data manipulation and widget alteration

Software:

Orange is an open-source software package released under GPL. Versions up to 3.0 include core components in C++ with wrappers in Python are available on GitHub. From version 3.0 onwards, Orange uses common Python open-source libraries for scientific computing, such as numpy, scipy and scikit-learn, while its graphical user interface operates within the cross-platform Qt framework.

The default installation includes a number of machine learning, preprocessing and data visualization algorithms in 6 widget sets (data, visualize, classify, regression, evaluate and unsupervised). Additional functionalities are available as add-ons (bioinformatics, data fusion and text-mining).

Orange is supported on macOS, Windows and Linux and can also be installed from the Python Package Index repository (pip install Orange3).

As of May 2018 the stable version is 3.13 and runs with Python 3, while the legacy version 2.7 that runs with Python 2.7 is still available.

Features:

Orange consists of a canvas interface onto which the user places widgets and creates a data analysis workflow. Widgets offer basic functionalities such as reading the data, showing a data table, selecting features, training predictors, comparing learning algorithms, visualizing data elements, etc. The user can interactively explore visualizations or feed the selected subset into other widgets.

Canvas: Graphical front-end for data analysis

Widgets:

- Data: Widgets for data input, data filtering, sampling, imputation, feature manipulation and feature selection.
- Visualization: Widgets for common visualization (box plot, histograms, scatter plot) and multivariate visualization (mosaic display, sieve diagram)
- Classify: A set of supervised machine learning algorithms for classification.
- Regression: A set of supervised machine learning algorithms for regression.
- Evaluate: Cross-validation, sampling-based procedures, reliability estimation and scoring of prediction methods.
- Unsupervised: Unsupervised learning algorithms for clustering (k-means, hierarchical clustering) and data projection techniques (multidimensional scaling, principal component analysis, correspondence analysis)

Add-ons:

- Associate: Widgets for mining frequent itemsets and association rule learning.
- Bioinformatics: Widgets for gene set analysis, enrichment, and access to pathway libraries.
- Data fusion: Widgets for fusing different data sets, collective matrix factorization, and exploration of latent factors.
- Educational: Widgets for teaching machine learning concepts, such as k-means clustering, polynomial regression, stochastic gradient descent.
- Geo: Widgets for working with geospatial data.
- Image analytics: Widgets for working with images and ImageNet embeddings

- Network: Widgets for graph and network analysis.
- Text mining: Widgets for natural language processing and text mining.
- Time series: Widgets for time series analysis and modeling.
- Spectroscopy: Widgets for analyzing and visualization of (hyper)spectral datasets.

Objectives:

The program provides a platform for experiment selection, recommendation systems, and predictive modeling and is used in biomedicine, bioinformatics, genomic research, and teaching. In science, it is used as a platform for testing new machine learning algorithms and for implementing new techniques in genetics and bioinformatics. In education, it was used for teaching machine learning and data mining methods to students of biology, biomedicine, and informatics.

Introduction to CRISP-DM Framework

CRISP DM is a framework that will help us decide how to approach a problem. We are going to use a methodology that was built initially for data mining problems, but works well for all types of business problems.

This methodology is called the cross industry standard process for data mining. The framework is generalized from the CRISP-DM methodology, to work for all types of business problems.

The CRISP-DM methodology provides a structured approach to planning a data mining and predictive analytics project. It is a robust and well-proven methodology.

Why CRISP-DM Methodology :

CRISP-DM provides a high level of flexibility that helps improve hypotheses and data analysis methods in a regular manner during further iterations.

It allows to create a long-term strategy based on short iterations at the beginning of project development.

During first iterations, a team can create a basic and simple model cycle that can easily be improved in further iterations.

This principle allows to ameliorate a preliminary developed strategy after obtaining additional information and insights.

CRISP-DM Project Management Methodology Famework

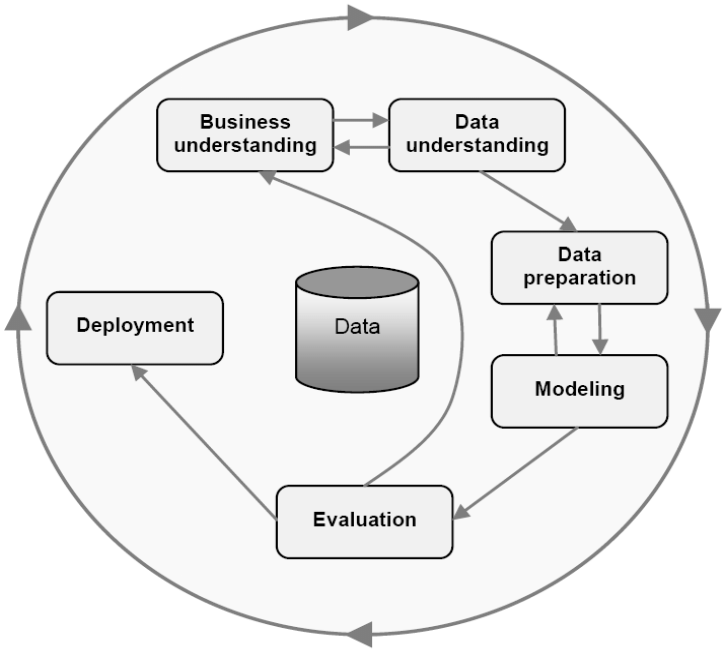


Figure 4 CRISP-DM Frame Work

- **Business Understanding**
  We understand the project objectives and requirements from a business prospective, and then convert this knowledge into a data mining problem

  definition and a preliminary data plan.

- **Data Understanding**
  We collect data in order to get familiar with it. We identify data quality problems, discover first insights into the data, or to detect interesting subsets to form hypothesis for hidden information.

- **Data Preparation:**
  In this phase, we cover all activities to construct the final dataset from the from the initial raw data.

- **Modeling:**
  Modeling techniques such as supervised, unsupervised, NLP, Time series, Neural networks are selected and applied.

- **Evaluation:**
  The models are evaluated based testing on unseen data and the best models are selected.

- **Deployment:**
  The deployment architecture is designed and provisioned on the cloud/on-prem and the models are then deployed.

Phase1: Business understanding:

Customer churn, also known as customer attrition, occurs when customers stop doing business with a company or stop using a company's services.
By being aware of and monitoring churn rate, companies are equipped to determine their customer retention success rates and identify strategies for improvement.
We will use EDA (Exploratory Data Analysis) and implement some algorithms using Orange Canvas Tool to understand the precise customer behaviors and attributes which signal the risk and timing of customer churn.

What is churn Rate?

Churn rate is the percentage of subscribers to a service that discontinue their subscription to that service in a given time period.

Customer churn is a critical metric because it is much more cost effective to retain existing customers than it is to acquire new customers as it saves cost of sales and marketing. Customer retention is more cost-effective as you've already earned the trust and loyalty of existing customers.

Churn rate is an important consideration in the industries like Banking and Insurance, Retail, Financial, online retail, Travel, and telephone and cell phone services industry.

With growing pressure from competition and government mandates improving retention rates of profitable customers has become an increasingly urgent to telecom service providers.

Churn can be categorised into three types:

1) **Involuntary churn**: This occurs when subscribers fail to pay for service and as a result the provider terminates service. Termination of service due to theft or fraudulent usage is also classified as involuntary churn.

2) **Unavoidable churn**: This occurs when a customer dies or moves or is otherwise permanently removed from the marketplace, travels outside the country without roaming, and possibly permanently
3) relocates to places outside the outside network coverage.

4) **Voluntary churn**: Termination of service relationship by the customer, leaving one service operator for another because of better value or dissatisfaction with current service provider.

Why churn rate is an important?

Churn rates are often used to indicate the strength of a company's customer service division and its overall growth prospects. Lower churn rates suggest a company is, or will be, in a better or stronger competitive state. Customer loss impacts carriers significantly as they often make a significant investment to acquire customers.

The ability to predict that a particular customer is at a high risk of churning, while there is still time to do something about it, represents a huge additional potential revenue source for every online business. Besides the direct loss of revenue that results from a customer abandoning the business, the costs of initially acquiring that customer may not have already been covered by the customer's spending to date. (In other words, acquiring that customer may have actually been a losing investment.)

Furthermore, it is always more difficult and expensive to acquire a new customer than it is to retain a current paying customer.

Obviously, a low churn rate is ideal. Companies that experience a high churn rate are under more pressure to generate revenue from other areas or gain new clients. It's almost always cheaper and easier to retain customers than it is to go through the process of acquiring new ones. Monitoring churn is the first step in understanding how good you are at retaining customers and identifying what actions might result in a higher retention rate. The survival of any business is based on its ability to retain customers.

Solutions to reduce churn rate:

Telecom players use a variety of different metrics to determine when customers are about to churn, or leave. It is profitable for companies to explore the reasons why customers are leaving, and then target at-risk customers with enticing offers.

- There are a number of different tactics companies use to maintain their customer bases. One of the most important is simply providing efficient customer service. Providing clients with an easy way to get questions answered and issues handled is the key to maintaining cellular clients.

- Value- added services serve as a subscriber retention tool, especially for established players. While for newer entrants, it will become a part of the marketing strategy to attract customers. If VAS providers leverage the opportunities to tie up with operators, there could be a major increase in the uptake of their services.

- A commonly used tactic is for a carrier to offer upgrades on the client's existing account. Expanding on services offered and giving better rates or discounts to the client often improves customer retention rates.

- Another tactic is offering free access or reduced rates on smartphone applications. The increasing regular use by customers of cellphone applications makes free access to such applications an enticing bonus for many customers.

- Competing cellular providers aggressively market special deals to churn customers away from their current provider. Common practices include offering free phones and buying out any existing service contract. The cellular service business is highly competitive and will likely remain so; therefore, churn rates will continue to be an important focus for cellular providers.

- Personalized Tariff plans and service recommendations to each subset of subscribers because a one-size-fit strategy is no longer suitable for telecom sector, every user has a different purpose and usage pattern.
- By leveraging the user traffic, operators' strategic and technical teams can make clear and decisive decisions to reduce costs by millions of dollars without jeopardizing quality.
- Fighting wireless churn with trendy smart-phones and fast data network
- One-on-one Marketing is one of the best tactics to reduce churn rate. Make sure that customers are communicated the new services offering based on their usage analysis and trends and should be given proactive information on the plans which will benefit the customer.
- Effective communication is one way to reduce churn. Being proactive in addressing difficulties and issues faced by your customers not only helps in building trust and reliability but also ensures a strong working relationship.
- Cultivate loyalty with attractive smartphone portfolios and strong mobile data networks to support those devices. Loyal customers are less likely to churn because they are more invested in your business relationship and companies have built up a long history of delivering good results and keeping promises.

Phase2: Data Understanding:

We collect the data in order to get familiar with it, we identify data quality problems, discover first insights into the data, or to detect interesting subsets to form hypothesis for hidden information.

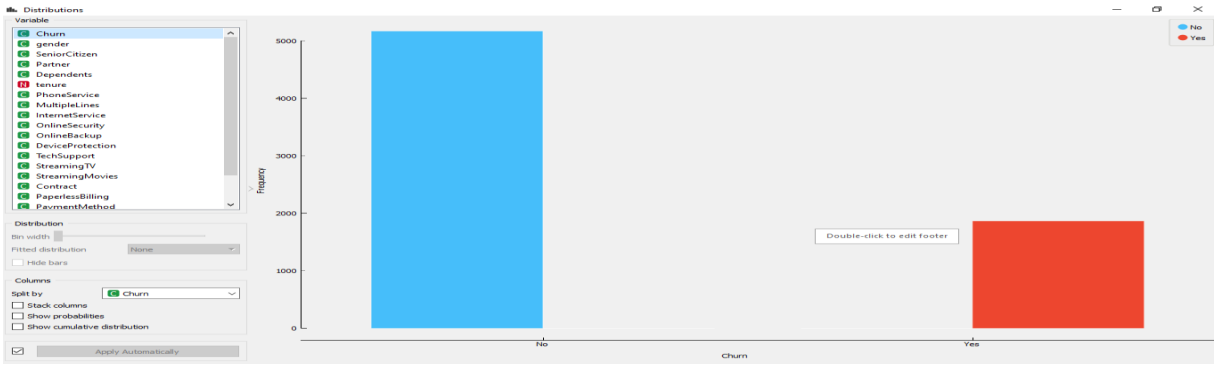
Loading the data into orange tool to understand its features.

The features in this dataset include the following:

Customer Demographic Data	**Subscribed Services**	**Customer Account Information**
● Gender ● SeniorCitizen ● Partner ● Dependents	● PhoneService ● MultipleLine ● InternetService ● OnlineSecurity ● OnlineBackup ● DeviceProtection ● TechSupport ● StreamingTV ● StreamingMovies	● CustomerID ● Contract ● PaerlessBilling ● PaymentMethod ● MonthlyCharges ● TotalCharges ● Tenure

Target is Churn, which has binary classes 1 and 0.
The objective here is to identify and quantify the factors which influence churn rate.
In below diagram it shows 74% of the customers do not churn and customer churn rate is 26%.

Customer Churn Rate



Each row represents a customer; each column contains customer's attributes.
The datasets have the following attributes or features and their datatypes :

Customer ID	Object Customer ID
GenderCustomer	Gender (female, male)
SeniorCitizen	Whether the customer is a senior citizen or not (1, 0)
PartnerWhether	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
PhoneService	Whether the customer has a phone service or not (Yes, No)
MultipleLines:	Whether the customer has multiple lines or not(yes, No, No phone service)
InternetService	Customer's internet service provider (DSL, Fiber optic, No)
OnlineSecurity:	Whether the customer has online security or not (Yes, No, No internet service)
OnlineBackup:	Whether the customer has online backup or not (Yes, No, No internet service)
DeviceProtection:	Whether the customer has device protection or not (Yes, No, No service)
TechSupport:	Whether the customer has tech support or not (Yes, No, No internet service)
StreamingTV:	Whether the customer has streaming TV or not (Yes, No, No internet service)
StreamingMovies:	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 year)
PaperlessBilling:	Whether the customer has paperless billing or not (Yes, No)
PaymentMethod:	The customer's payment method(Electronic check, mailed check ,Bank

	transfer(automatic), Credit card(automatic))
MonthlyCharges:	The amount charged to the customer monthly
TotalCharges:	The total amount charged to the customer
Churn:	Whether the customer churned or not(Yes or No)

The dataset contains 7043 rows(customers) and 21 columns (features). The churn column is our target (Dependent variable).

Churn	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup
No	Female	1	Yes	No	25	Yes	Yes	DSL	Yes	No
Yes	Female	1	Yes	No	8	Yes	Yes	Fiber optic	No	Yes
No	Female	1	Yes	Yes	60	Yes	No	DSL	Yes	Yes
Yes	Male	1	No	No	18	Yes	Yes	Fiber optic	No	No
No	Female	0	Yes	Yes	63	Yes	Yes	Fiber optic	Yes	No
No	Male	1	Yes	Yes	66	Yes	Yes	Fiber optic	No	Yes
No	Female	0	Yes	Yes	34	Yes	Yes	No	No internet ser...	No internet ser...
No	Female	0	No	No	72	Yes	Yes	Fiber optic	No	No
No	Female	0	Yes	No	47	Yes	Yes	Fiber optic	No	No
No	Male	0	No	No	60	Yes	Yes	Fiber optic	No	Yes
No	Male	0	Yes	No	72	No	No phone service	DSL	Yes	Yes
No	Female	0	Yes	Yes	18	Yes	No	DSL	No	No
Yes	Female	0	No	No	9	Yes	Yes	Fiber optic	No	No
No	Female	0	No	No	3	Yes	No	DSL	No	Yes
No	Male	0	Yes	No	47	Yes	Yes	Fiber optic	No	Yes
No	Female	0	No	No	31	Yes	No	DSL	No	Yes
No	Female	0	Yes	Yes	50	Yes	No	No	No internet ser...	No internet ser...
No	Male	0	No	No	10	Yes	No	Fiber optic	Yes	No
No	Male	0	No	No	1	Yes	No	DSL	No	No
No	Female	0	Yes	Yes	52	Yes	No	No	No internet ser...	No internet ser...

DeviceProtectior	TechSupport	StreamingTV	StreamingMovie:	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges
No	Yes	Yes	Yes	Two year	Yes	Electronic check	54.60	3423.50
No	No	Yes	No	Month-to-month	No	Electronic check	89.85	248.40
No	No	No	No	Two year	No	Bank transfer (a...	31.05	1126.35
No	No	Yes	Yes	Month-to-month	Yes	Electronic check	100.25	1064.65
No	Yes	No	No	Month-to-month	Yes	Credit card (aut...	85.20	2151.60
No	No	No	No	Month-to-month	Yes	Electronic check	74.40	229.55
No	No	No	No	Month-to-month	No	Mailed check	50.70	350.35
No	Yes	Yes	No	Two year	Yes	Electronic check	88.95	3027.65
No	No	Yes	Yes	One year	Yes	Electronic check	56.45	3985.35
No	No	No	Yes	Month-to-month	Yes	Electronic check	85.95	1215.65
No	Yes	No	No	Two year	No	Bank transfer (a...	50.55	3260.10
No	No	No	Yes	Month-to-month	Yes	Electronic check	35.45	35.45
No	No	No	No	Month-to-month	No	Electronic check	44.35	81.25
No	No	No	No	Month-to-month	Yes	Bank transfer (a...	75.00	1778.50
No	No	No	No	Month-to-month	Yes	Mailed check	70.45	70.45
No	Yes	No	Yes	Month-to-month	No	Electronic check	71.15	563.65
No	No	Yes	Yes	Two year	Yes	Bank transfer (a...	101.05	5971.25
No	No	No	No	One year	Yes	Credit card (aut...	84.30	5289.05
No	Yes	Yes	No	Two year	Yes	Credit card (aut...	99.05	6416.70
No	No	No	No	Month-to-month	Yes	Electronic check	45.65	45.65
No	Yes	Yes	Yes	Month-to-month	Yes	Electronic check	68.55	564.35
No	No	Yes	Yes	Month-to-month	No	Mailed check	95.00	655.50

The below output shows that there are 11 missing values for Total Charges.

	Name	Distribution	Center	Dispersion	Min.	Max.	Missing
C	StreamingMovies		No	1.07			0 (0%)
C	PaymentMethod		Electronic check	1.37			0 (0%)
N	MonthlyCharges		64.7617	0.4646	18.25	118.75	0 (0%)
N	tenure		32.37	0.76	0	72	0 (0%)
N	TotalCharges		2283.3004	0.9927	18.80	8684.80	11 (0%)

Data visualization of descriptive statistics

Churn with reference to gender

Here are the bar plots of demographic data of our sample.
From this below demographic plots, we notice that the sample is evenly split across gender. With reference to churn, churn is more among males than female. There is no effect of Gender on Churn.

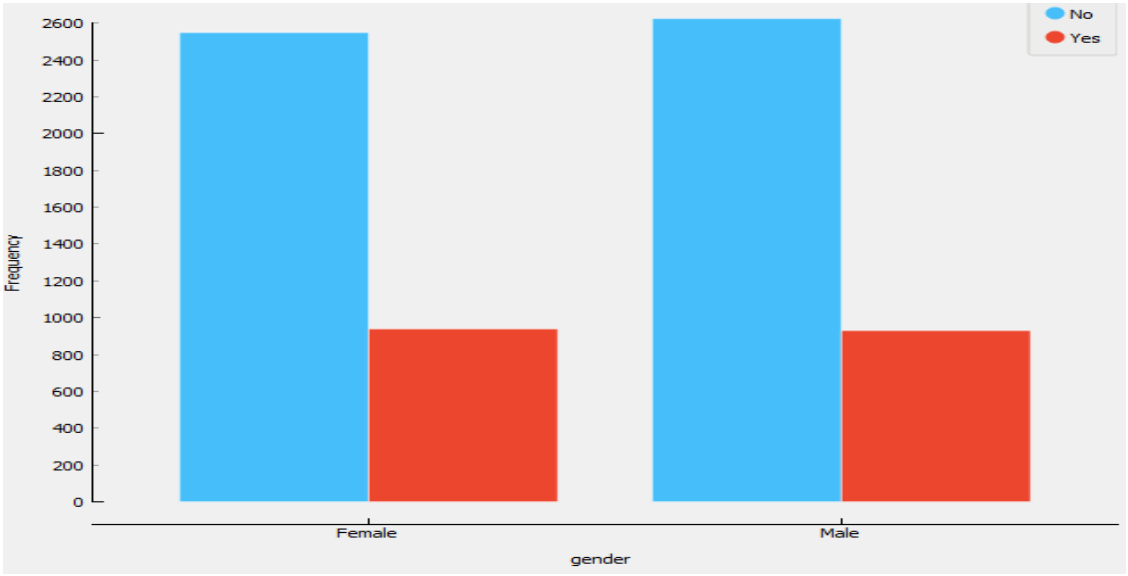
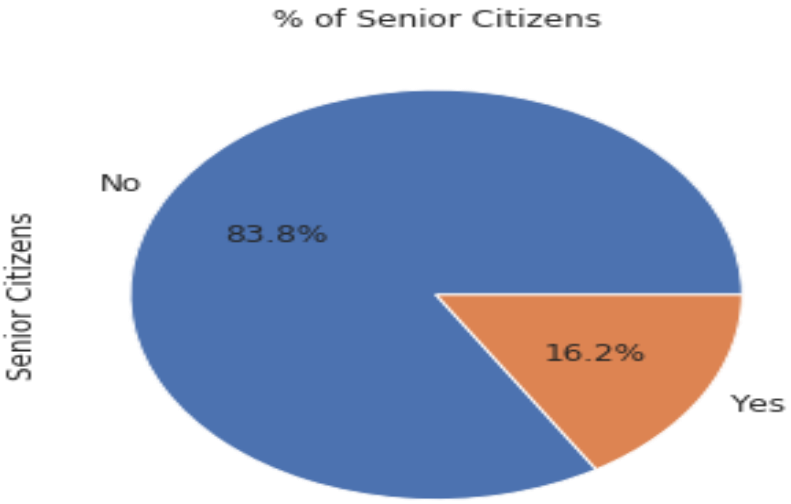


Figure 6 Churn with reference to Gender

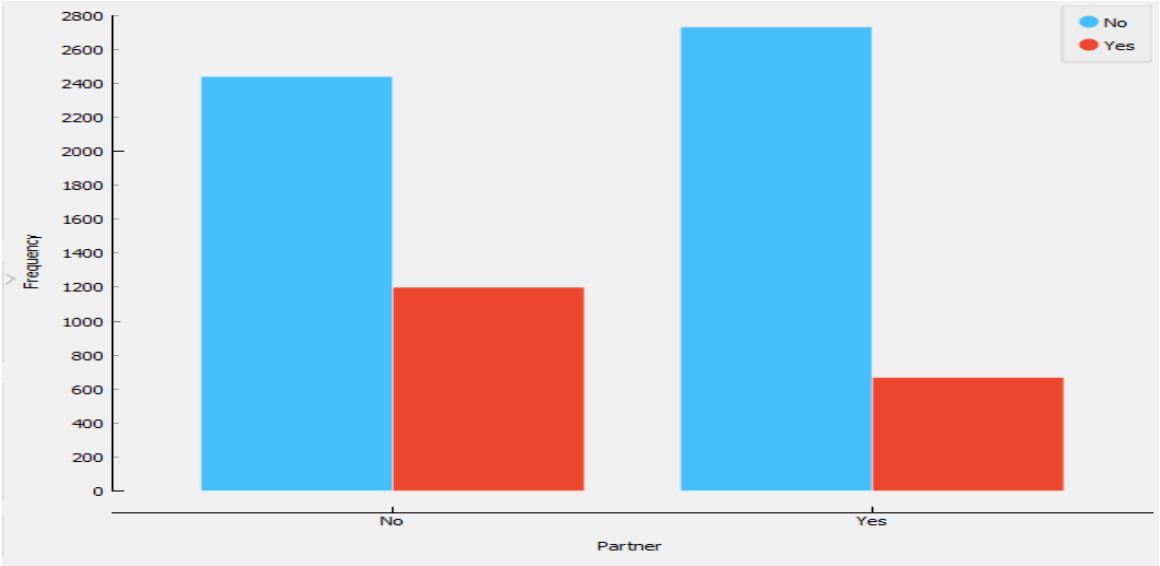
Churn with reference to SeniorCitizen

In this below diagram it shows that there are only 16% of the customers who are senior citizens. Thus most of our customers in the data are younger people. Senior citizen tends to churn more



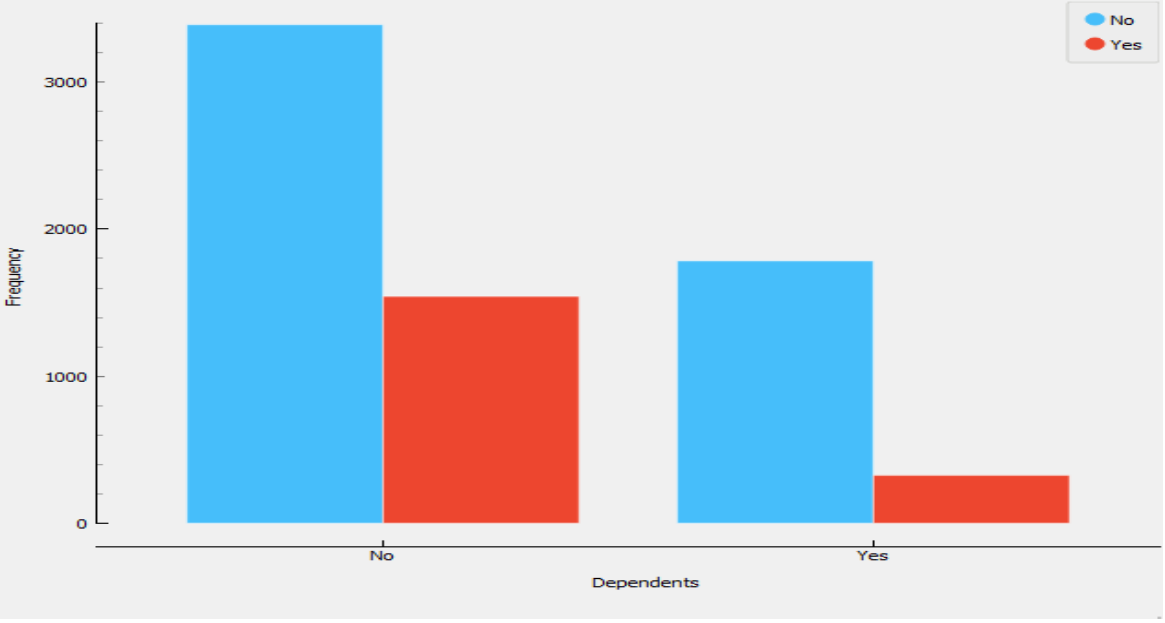
Churn with reference to Partner

In this below diagram it shows that the sample is evenly split across partner status. About 52% of the customers have a partner.



Churn with reference to dependents

In this below diagram it shows that only 30% of the total customers have dependents. Customer with no dependents tends to churn more.

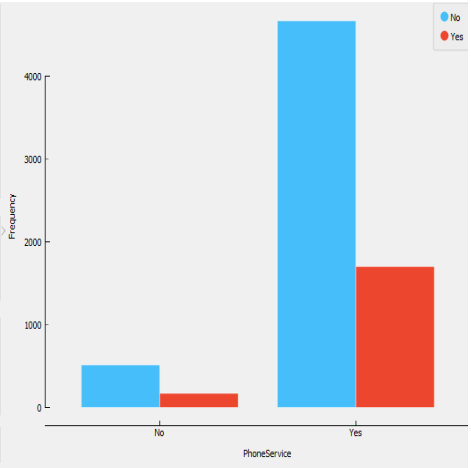


The various offered services are plotted below.

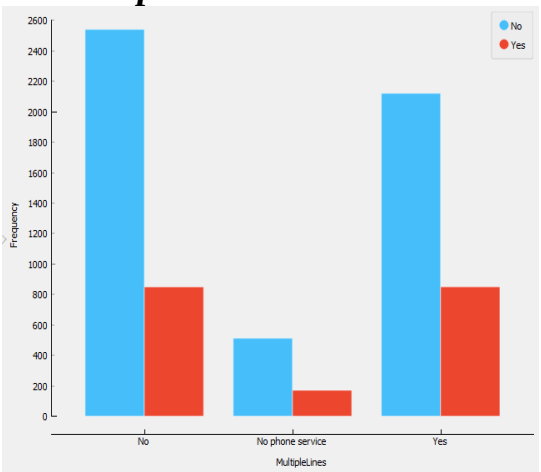
Churn with reference to PhoneService and MultipleLines

Multiple lines of internet connectivity doesn't effect churn that much.

PhoneService



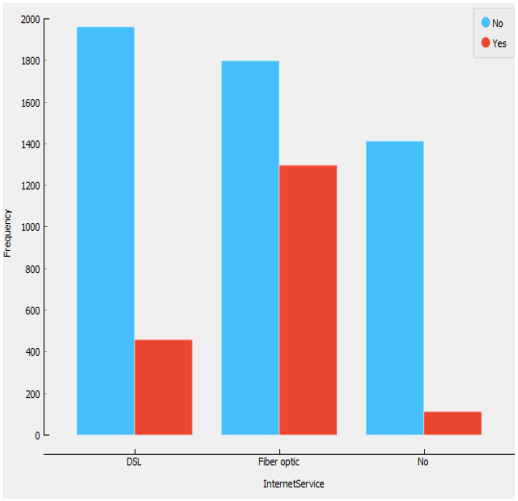
MultipleLines



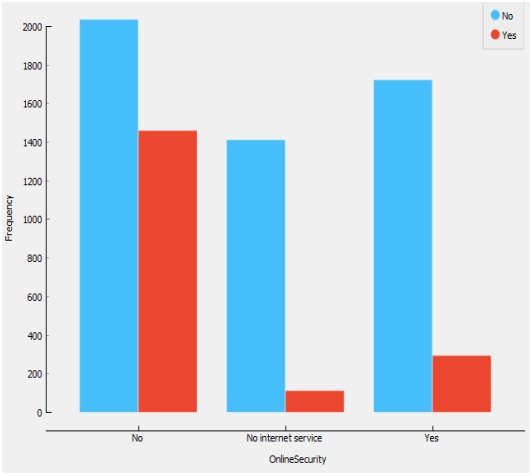
Churn with reference to InternetService and Online Security

Fiber optic internet connection is more popular than DSL internet connection and each online service has a minority of users. The Fiber optic internet is costly and thus should either be promoted to appropriate target audience or better technology can be implemented to cut cost on this service. Ultimately the market research team has to decide the break even point for this service, whether it is profiting as much as the loss of customers it is causing.

InternetService



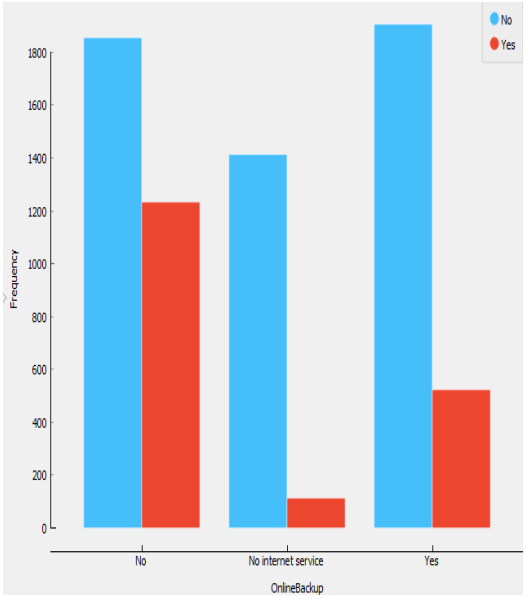
Online Security



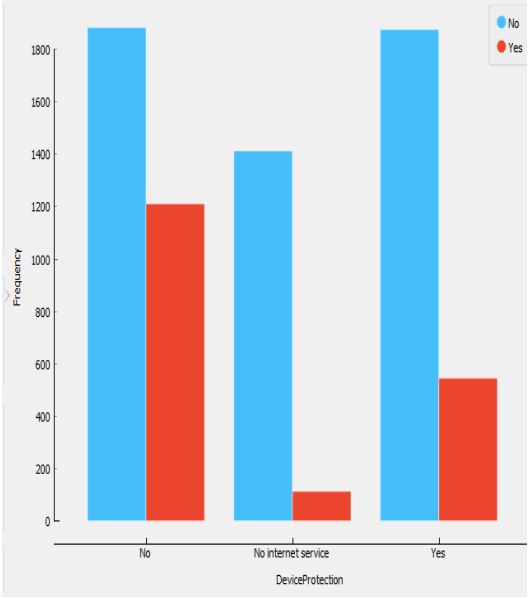
Churn with reference to onlineBackup and DeviceProtection

Customer opted for Online Backup churn less than who have not opted and also customers opted for Device protection churn less than who have not opted.

OnlineBackup



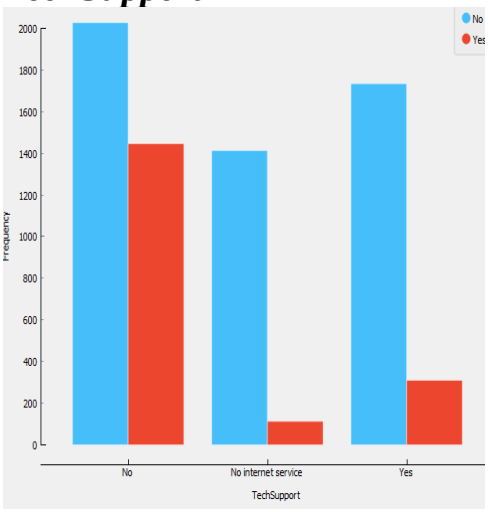
DeviceProtection



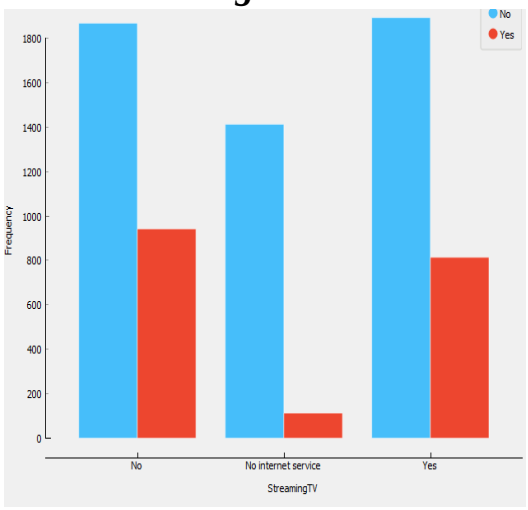
Churn with reference to TechSupport and StreamingTV

Customer opted for Tech support churn less than who have not opted and Streaming TV doesn’t make such impact on churning.

TechSupport

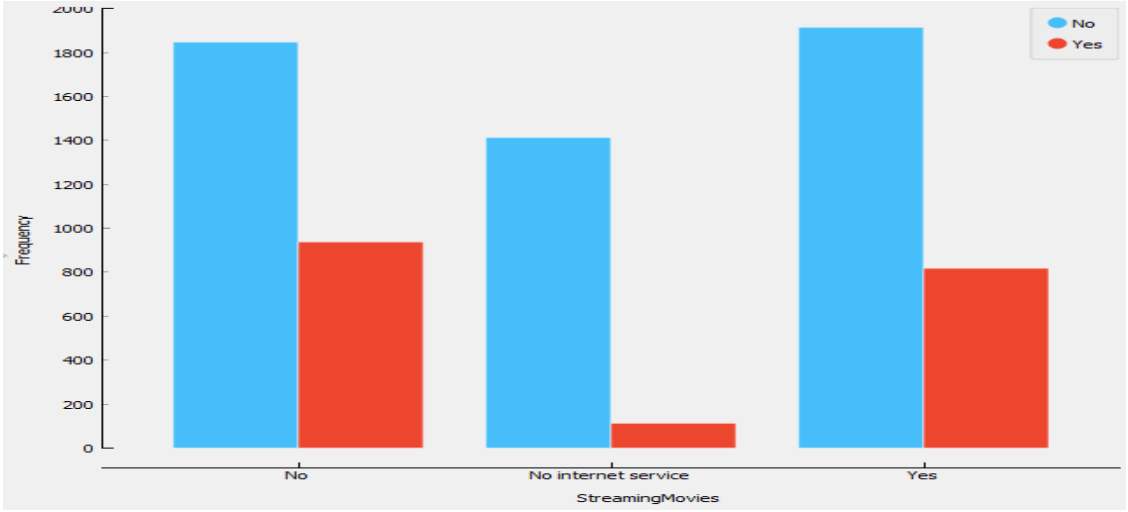


StreamingTV



Churn with reference to StreamingMovies

Streaming Movies doesn’t make such impact on churning.

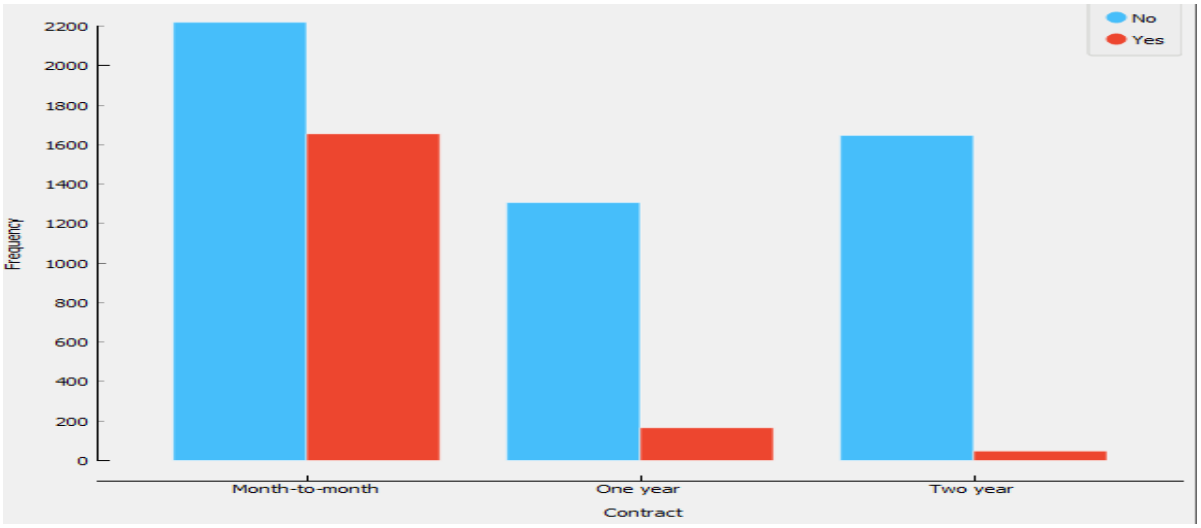


<u>The remaining categorical variables are related to contract and payment status.</u>

Churn with reference to Contract

In this below graph it is observed that most customers who are on month-to-month contracts are more likely to churn. It can be hypothesized that the reason is to be attributed to personal reasons of customers to have reservations about long term contracts or higher costs per unit time resulting from monthly contracts.

From the below graph we can see that most of the customers are in the month-to-month contract. Interestingly most of the monthly contracts last for 1-2 months, while the 2 years contracts tend to last about 70 months. This shows that customers taking a longer contract are more loyal to the company and tend to stay with it for longer time.



Churn with reference to PaperlessBilling

Churn rate is higher for the customers who opted for paperless billing.

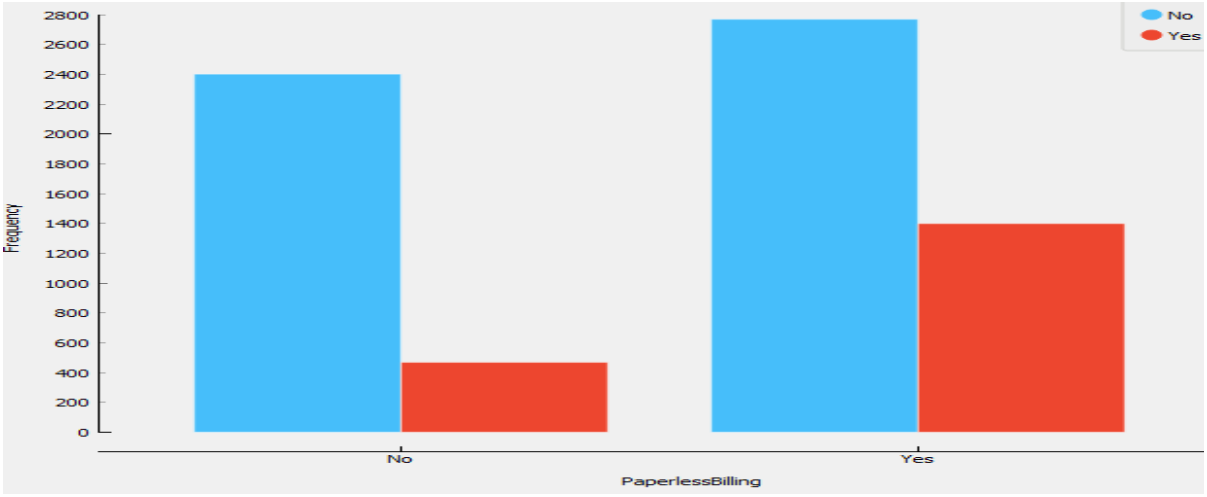
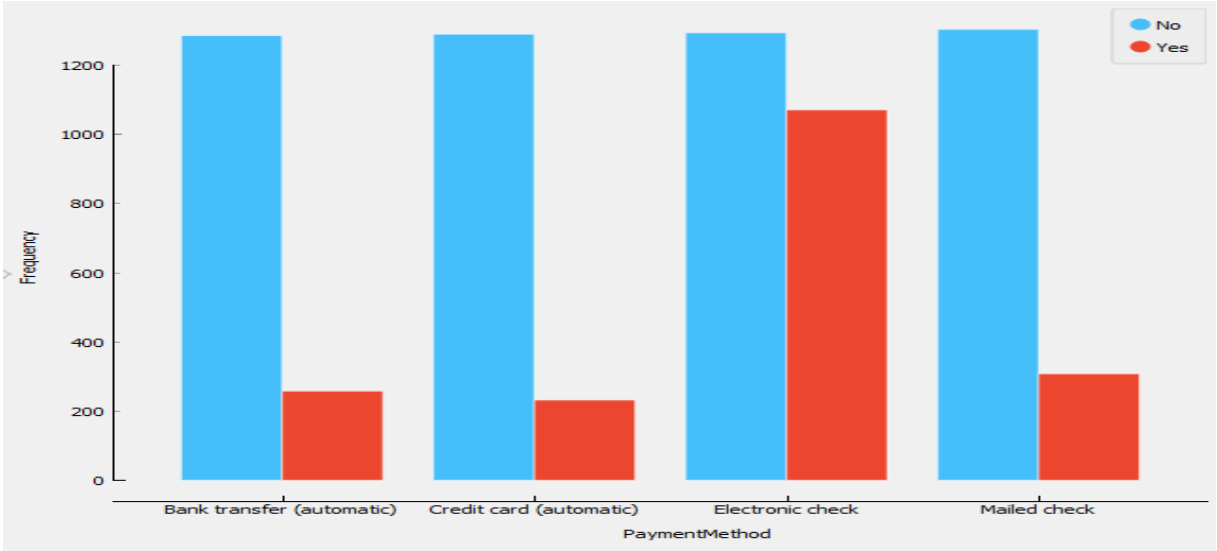


Figure 16 Churn with reference to PaperlessBilling

Churn with reference to PaymentMethod

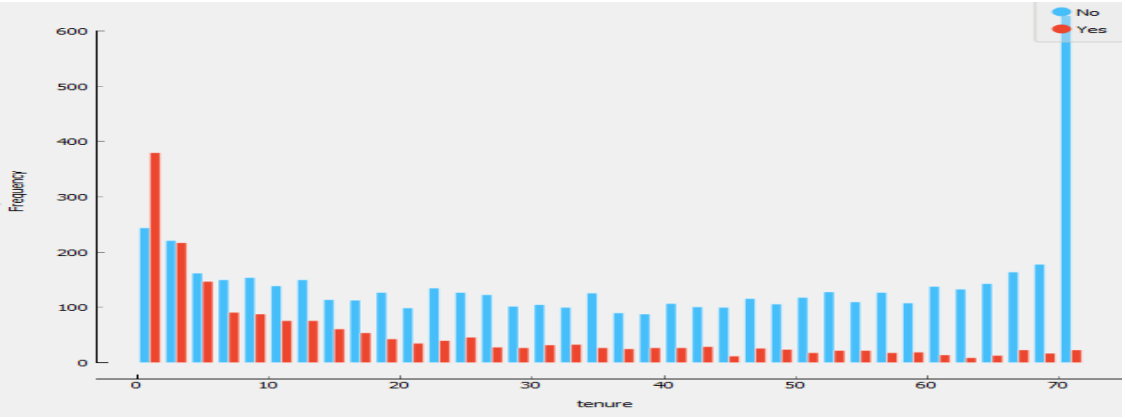
Most of the samples are on paperless billing and pay by electronic check.



Now the distribution of the quantitative variables.

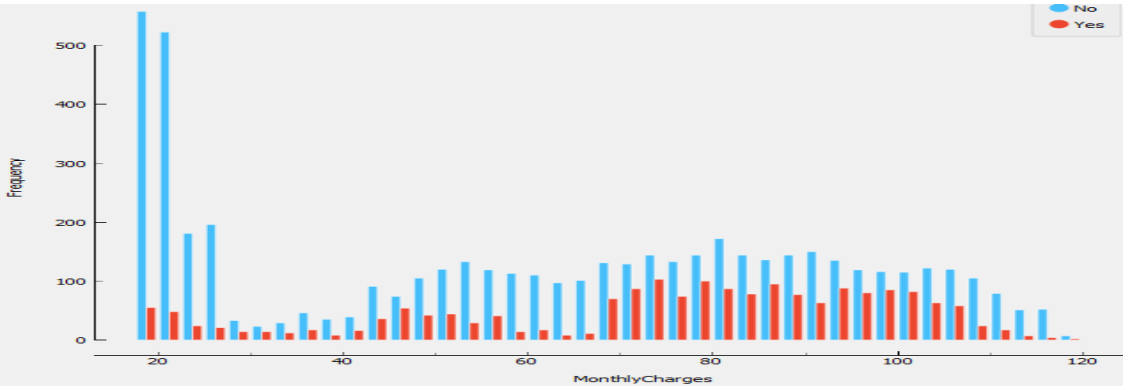
Churn Vs Tenure

From the below histogram we can see that a lot of customers have been with the telecom company for just a month, while quite a many are there for about 72 months. This could be potentially because different customers have been different contracts. Thus based on the contract they are into it could be more/less easier for the customers to stay/leave the telecom company.



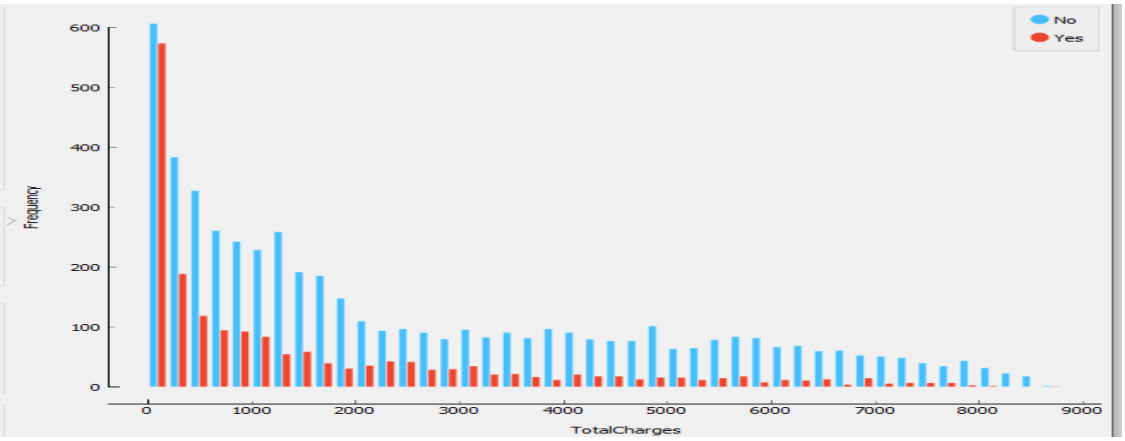
Churn Vs MonthlyCharges

It appears as if the monthly charges variable is roughly normally distributed around $80 per month with a large stack near the lowest rates.

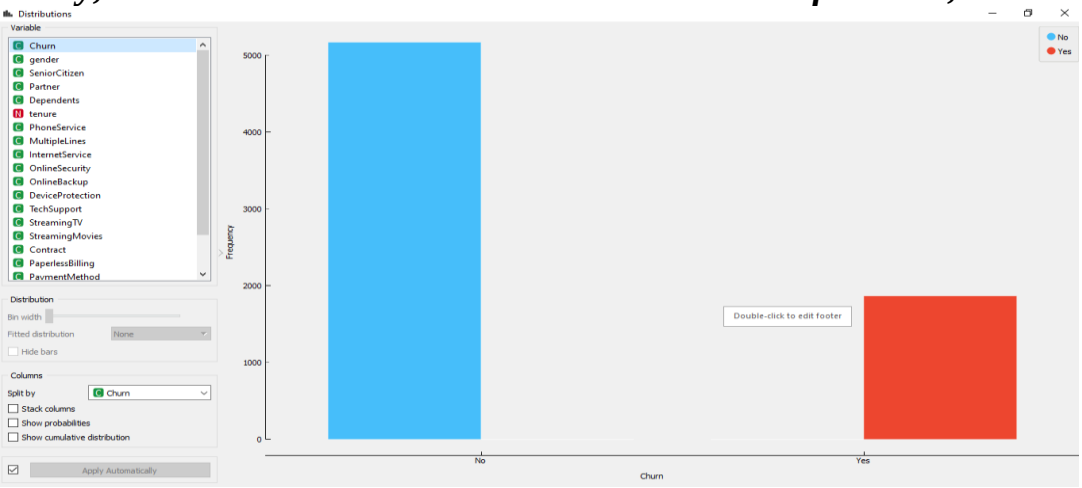


Churn Vs TotalCharges

The total charges variable is positively skewed with a large stack near the lower amounts.



Lastly, we will examine our main outcome variable of interest, churn.

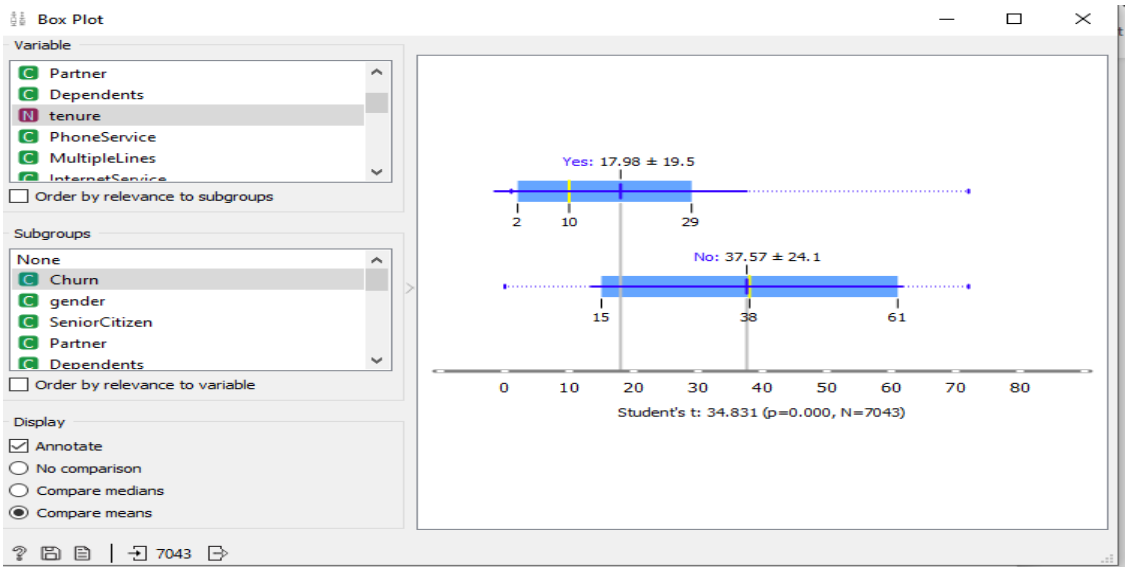


In our data, 74% of the customers do not churn. Clearly the data is skewed as we would expect a large majority of the customers to not churn.

Now let's explore the churn rate by tenure, seniority, contact type, monthly charges and total charges to see how it varies by these variables.

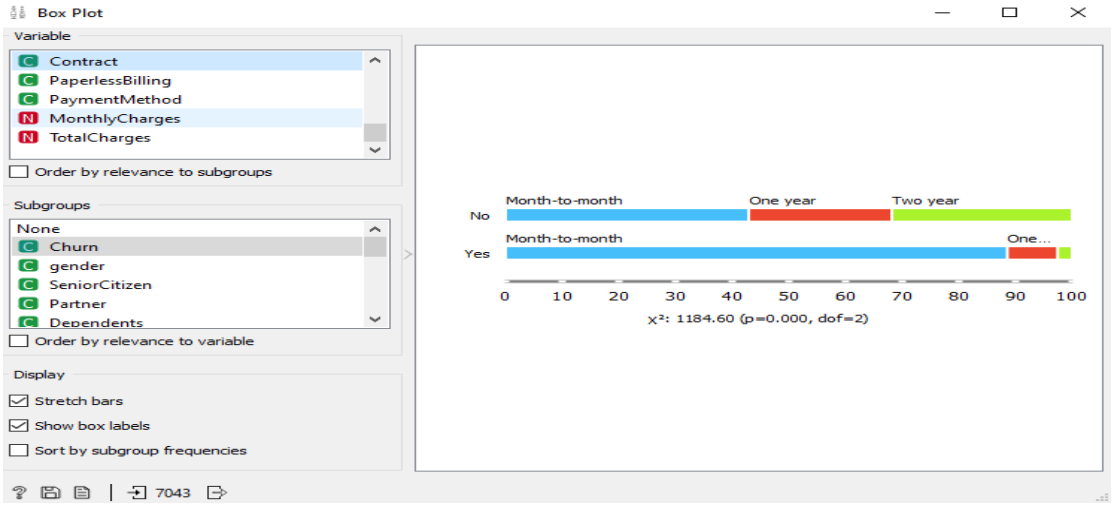
Churn Vs Tenure:

As we can see from the below box plot, the customers who do not churn, they tend to stay for a longer tenure with the telecom company.



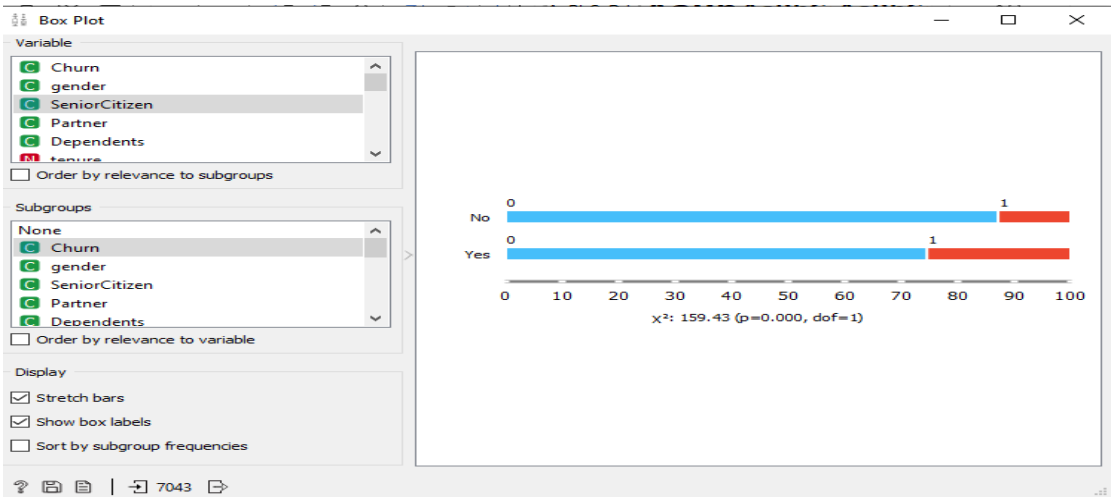
Churn by contract type:

The customer who have a month to month contract have a very high churn rate.



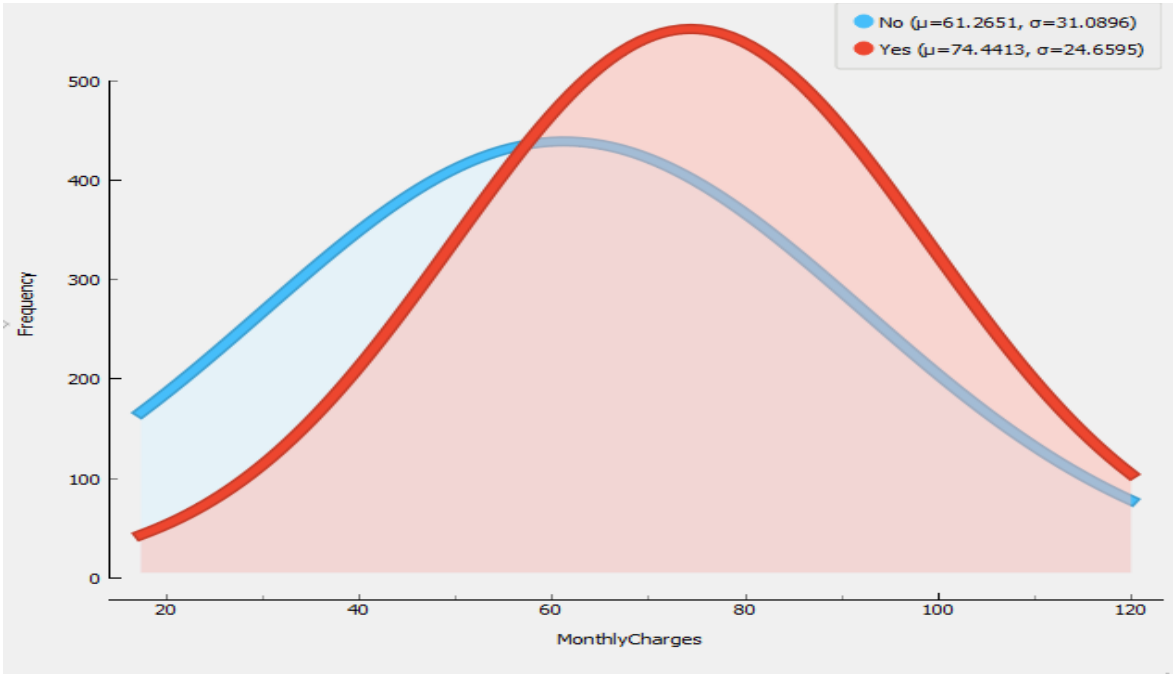
Churn by Seniority:

Senior citizens have almost double the churn rate than younger population.



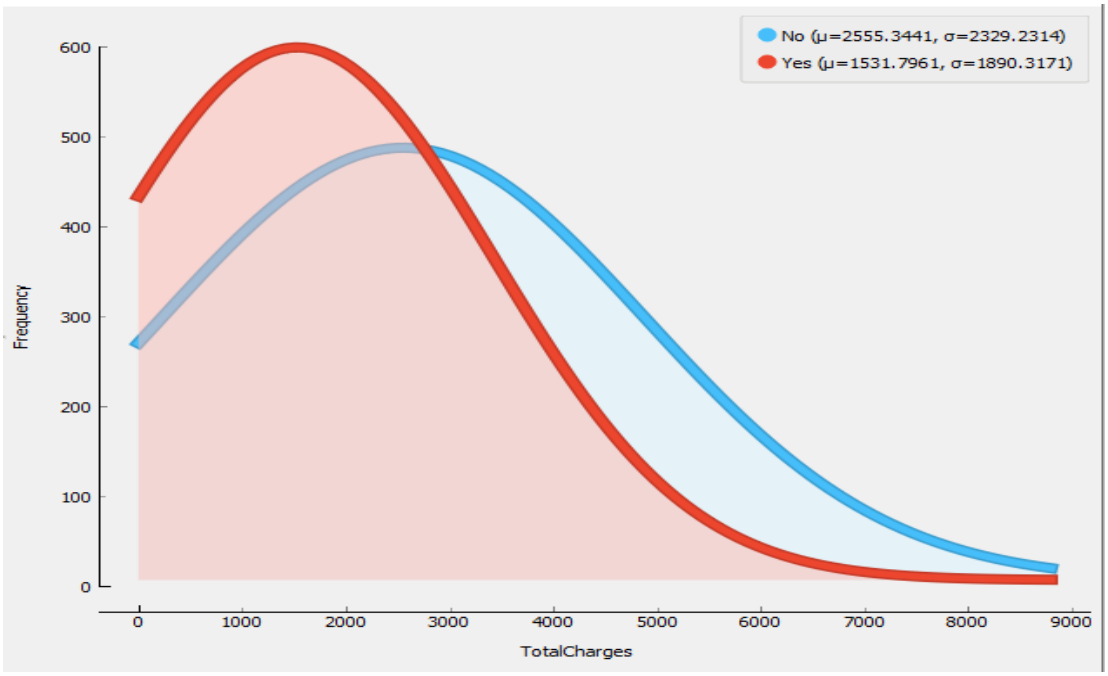
Churn by monthly charges:

Higher percentage of customers churn when the monthly charges are high.



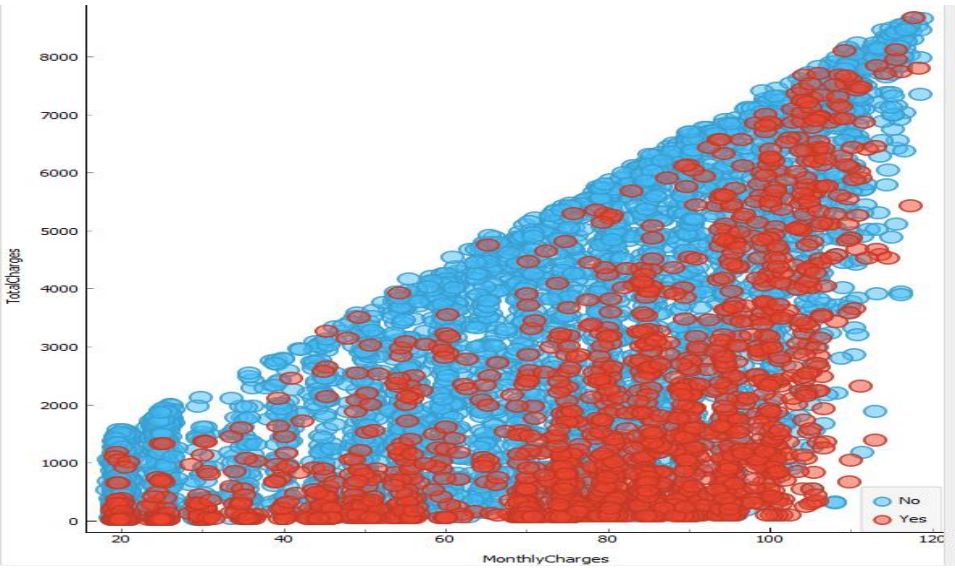
Churn by total charges:

It seems that there is high churn when the total charges are lower.



Relation between monthly and total charges:

We are seeing from the below scatter plot that the total charges increases as the monthly bill for a customer increases.



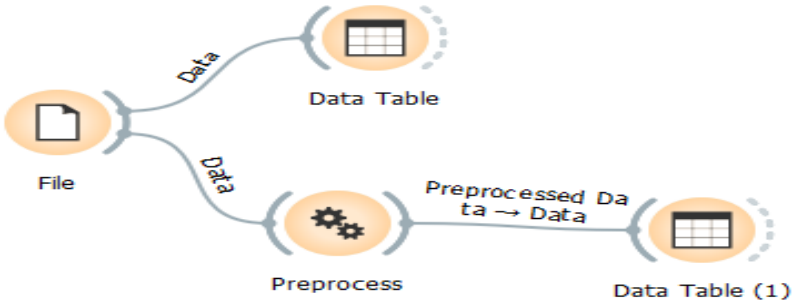
Phase*3:* Data Preparation:

Data Preparation Steps:

Removing the null values (missing values) from the datase.
Converting some binary variables (Yes/No) to 0/1

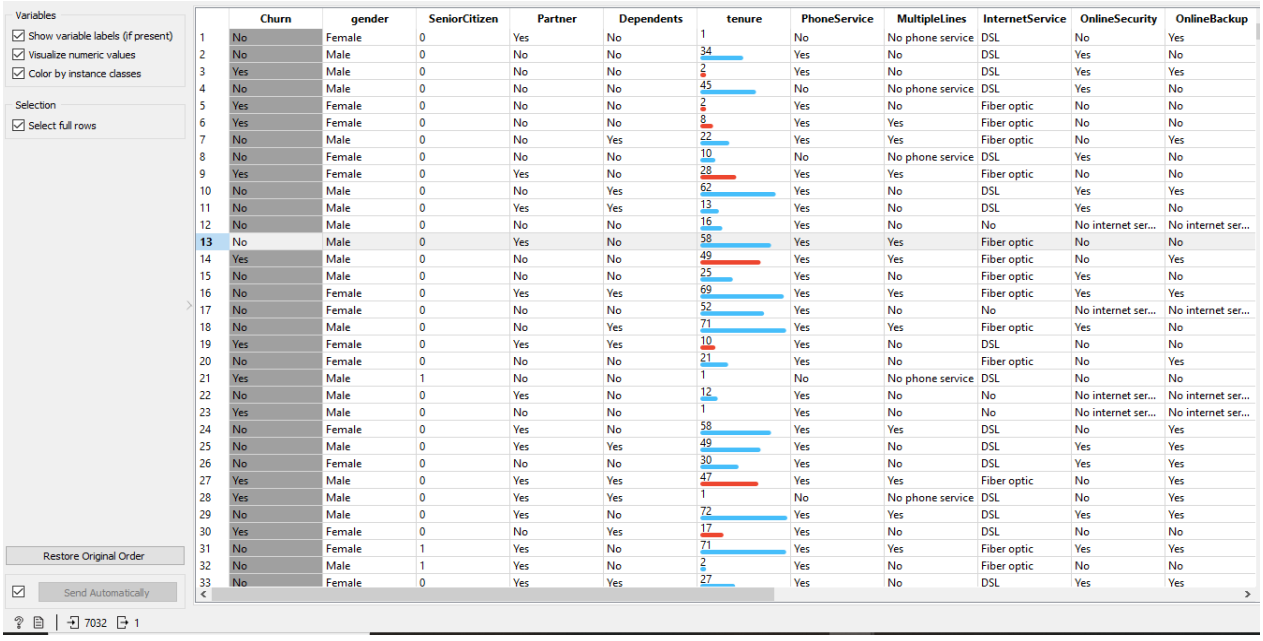
Removing the null values (missing values) from the datase.

We removed the null values from TotalCharges rows. And now the total number of observations are 7032



Data after removing the null values.

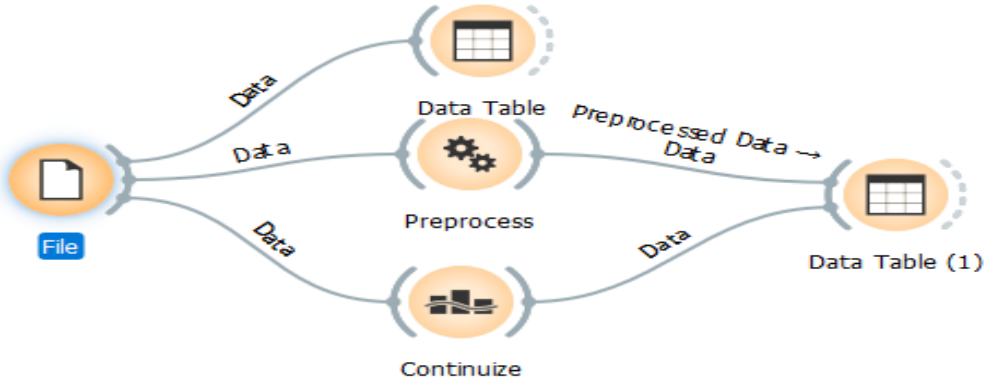
All the null values from the column TotalCharges got removed.



	Churn	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup
1	No	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes
2	No	Male	0	No	No	34	Yes	No	DSL	Yes	No
3	Yes	Male	0	No	No	2	Yes	No	DSL	Yes	Yes
4	No	Male	0	No	No	45	No	No phone service	DSL	Yes	No
5	Yes	Female	0	No	No	2	Yes	No	Fiber optic	No	No
6	Yes	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No
7	No	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes
8	No	Female	0	No	No	10	No	No phone service	DSL	Yes	No
9	Yes	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No
10	No	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes
11	No	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No
12	No	Male	0	No	No	16	Yes	No	No	No internet ser...	No internet ser...
13	No	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No	No
14	Yes	Male	0	No	No	49	Yes	Yes	Fiber optic	No	Yes
15	No	Male	0	No	No	25	Yes	No	Fiber optic	Yes	No
16	No	Female	0	Yes	Yes	69	Yes	Yes	Fiber optic	Yes	Yes
17	No	Female	0	No	No	52	Yes	No	No	No internet ser...	No internet ser...
18	No	Male	0	No	Yes	71	Yes	Yes	Fiber optic	Yes	No
19	Yes	Female	0	Yes	Yes	10	Yes	No	DSL	No	No
20	No	Female	0	No	No	21	Yes	No	Fiber optic	No	Yes
21	Yes	Male	1	No	No	1	No	No phone service	DSL	No	No
22	No	Male	0	Yes	No	12	Yes	No	No	No internet ser...	No internet ser...
23	Yes	Male	0	No	No	1	Yes	No	No	No internet ser...	No internet ser...
24	No	Female	0	Yes	No	58	Yes	Yes	DSL	No	Yes
25	No	Male	0	Yes	Yes	49	Yes	No	DSL	Yes	Yes
26	No	Female	0	No	No	30	Yes	No	DSL	Yes	Yes
27	Yes	Male	0	Yes	Yes	47	Yes	Yes	Fiber optic	No	Yes
28	Yes	Male	0	Yes	Yes	1	No	No phone service	DSL	No	Yes
29	No	Male	0	Yes	No	72	Yes	Yes	DSL	Yes	Yes
30	Yes	Female	0	No	Yes	17	Yes	No	DSL	No	No
31	No	Female	1	Yes	No	71	Yes	Yes	Fiber optic	Yes	Yes
32	No	Male	1	Yes	No	2	Yes	No	Fiber optic	No	No
33	No	Female	0	Yes	Yes	27	Yes	No	DSL	Yes	Yes

Converting some binary variables (Yes/No) to 0/1

Converting the variables like Gender, partner, dependents, phone service, multiple lines, internet security, online security, online backup, device protection, tech support, streaming TV, streaming movies, contract, paperless billing, and payment method into binary form (0/1)



Data Transformation (Data after converting the text to binary form)

Churn	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup
0	0	0	1	0	1	0	1	0	0	2
0	1	0	0	0	34	1	0	0	2	0
1	1	0	0	0	2	1	0	0	2	2
0	1	0	0	0	45	0	1	0	2	0
1	0	0	0	0	2	1	0	1	0	0
1	0	0	0	0	8	1	2	1	0	0
0	1	0	0	1	22	1	2	1	0	2
0	0	0	0	0	10	0	1	0	2	0
1	0	0	1	0	28	1	2	1	0	0
0	1	0	0	1	62	1	0	0	2	2
0	1	0	1	1	13	1	0	0	2	0
0	1	0	0	0	16	1	0	2	1	1
0	1	0	1	0	58	1	2	1	0	0
1	1	0	0	0	49	1	2	1	0	2
0	1	0	0	0	25	1	0	1	2	0
0	0	0	1	1	69	1	2	1	2	2
0	0	0	0	0	52	1	0	2	1	1
0	1	0	0	1	71	1	2	1	2	0
1	0	0	1	1	10	1	0	0	0	0
0	0	0	0	0	21	1	0	1	0	2
1	1	1	0	0	1	0	1	0	0	0

DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges
0	0	0	0	0	1	2	29.85	29.85
2	0	0	0	1	0	3	56.95	1889.50
0	0	0	0	0	1	3	53.85	108.15
2	2	0	0	1	0	0	42.30	1840.75
0	0	0	0	0	1	2	70.70	151.65
2	0	2	2	0	1	2	99.65	820.50
0	0	2	0	0	1	1	89.10	1949.40
0	0	0	0	0	0	3	29.75	301.90
2	2	2	2	0	1	2	104.80	3046.05
0	0	0	0	1	0	0	56.15	3487.95
0	0	0	0	0	1	3	49.95	587.45
1	1	1	1	2	0	1	18.95	326.80
2	0	2	2	1	0	1	100.35	5681.10
2	0	2	2	0	1	0	103.70	5036.30
2	2	2	2	0	1	2	105.50	2686.05
2	2	2	2	2	0	1	113.25	7895.15
1	1	1	1	1	0	3	20.65	1022.95
2	0	2	2	2	0	0	106.70	7382.25
2	2	0	0	0	0	1	55.20	528.35
2	0	0	2	0	1	2	90.05	1862.90
2	0	0	2	0	1	2	39.65	39.65

Phase4: Model Implementation

- ***Logistic Regression***
- ***Random Forest***
- ***Tree***

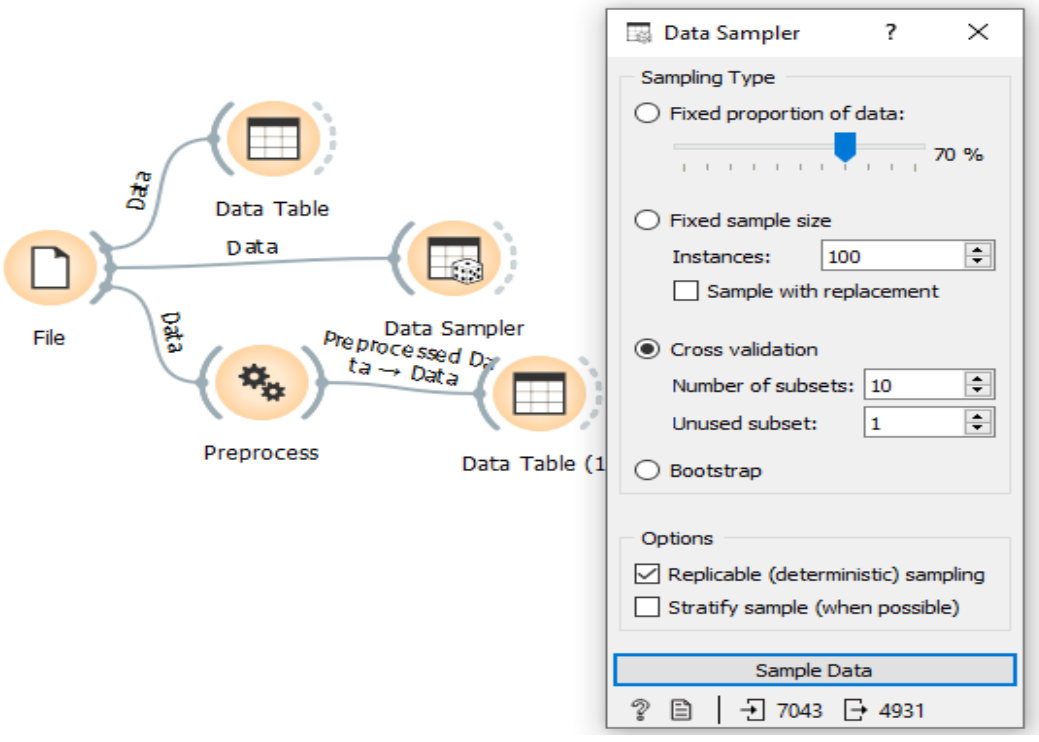
After going through the above EDA (Exploratory Data Analysis) we will develop some predictive models and compare them. We will develop Logistic Regression and Random Forest.

It is important to scale the variables in logistic regression so that all of them are within a range of 0 to 1. Before model implementation we split the data into training and test sets (70% vs 30%)

Cross validation:

For any model in machine learning this considered as a best practice if the model is tested with the independent data-set. Normally any prediction model work on unknown dataset which is also known as training set. It is fit to work with any model in the future. It helps us better use our data and it gives us much more information about our algorithm performance.

Since our data is imbalanced dataset the cross validation process with 10 folds is carried out. Data Sampler widget selects a subset of data instances from an input dataset.



Train Test splitting by using Data Sampler widget.

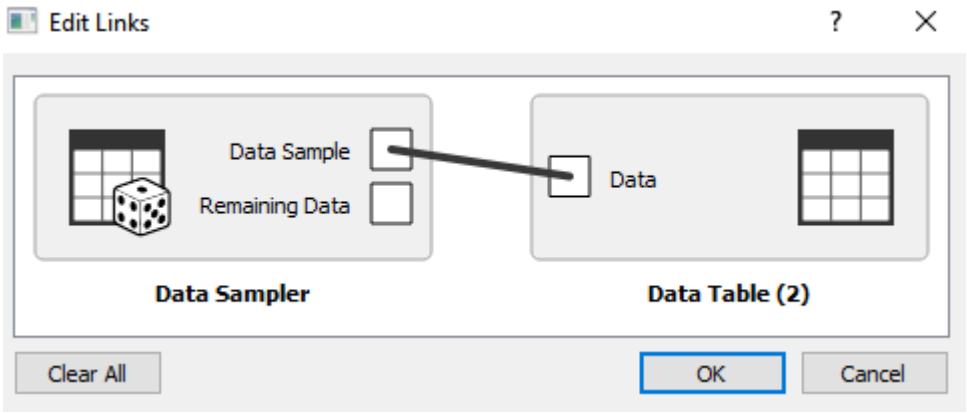
In Data sampler we are dividing the data into 70, 30 ratio and cross validation as 10

Train a training set.

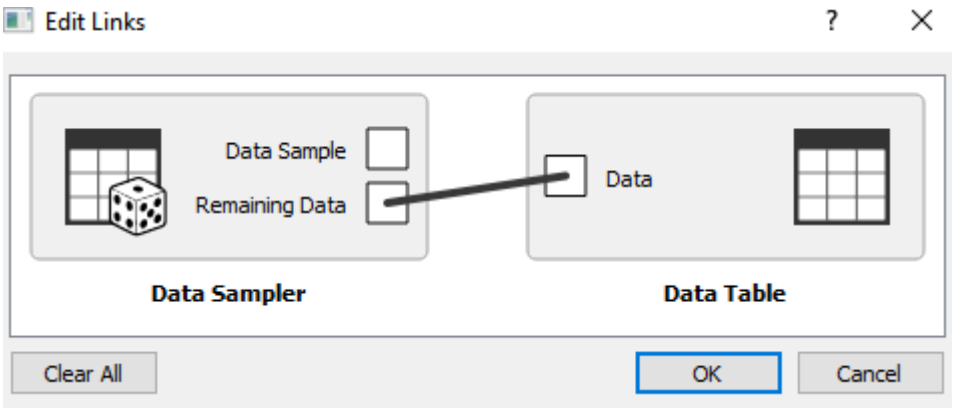
Validating a validation set (eg. using k-fold cross validation)

Testing the model with a test set.

This below figure shows that the 70% of the data which is 4931 records are into training set.



And the remaining data that is 30% (2112) records are into test set.



Logistic Regression:

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Types of logistic regression:

Generally, logistic regression means binary logistic regression having binary target variables, but there can be two more categories of target variables that can be predicted by it. Based on those number of categories, logistic regression can be divided into following types -

Binary or Binomial
In such a kind of classification, a dependent variable will have only two possible types either 1 and 0. For example, these variables may represent success or failure, yes or no, win or loss etc.

Multinomial
In such a kind of classification, dependent variable can have 3 or more possible unordered types or the types having no quantitative significance. For example, these variables may represent “Type A” or “Type B” or “Type C”.

Ordinal
In such a kind of classification, dependent variable can have 3 or more possible ordered types or the types having a quantitative significance. For example, these variables may represent “poor” or “good”, “very good”, “Excellent” and each category can have the scores like 0,1,2,3.

Logistic Regression Assumptions:

Before dividing into the implementation of logistic regression, we must be aware of the following assumptions about the same -

In case of binary logistic regression, the target variables must be binary always and the desired outcome is represented by the factor level 1.

There should not be any multi-collinearity in the model, which means the independent variables must be independent of each other.

We must include meaningful variables in our model.

We should choose a large sample size for logistic regression.

Random Forest:

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however it is mainly used for classification problems.

As we know that a forest is made up of trees and more trees means more robust forest.

Similarly Random forest algorithm creates decision trees on data samples and gets the prediction from each of them and finally selects the best solution by means of voting.

It is assemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Working of Random Forest Algorithm

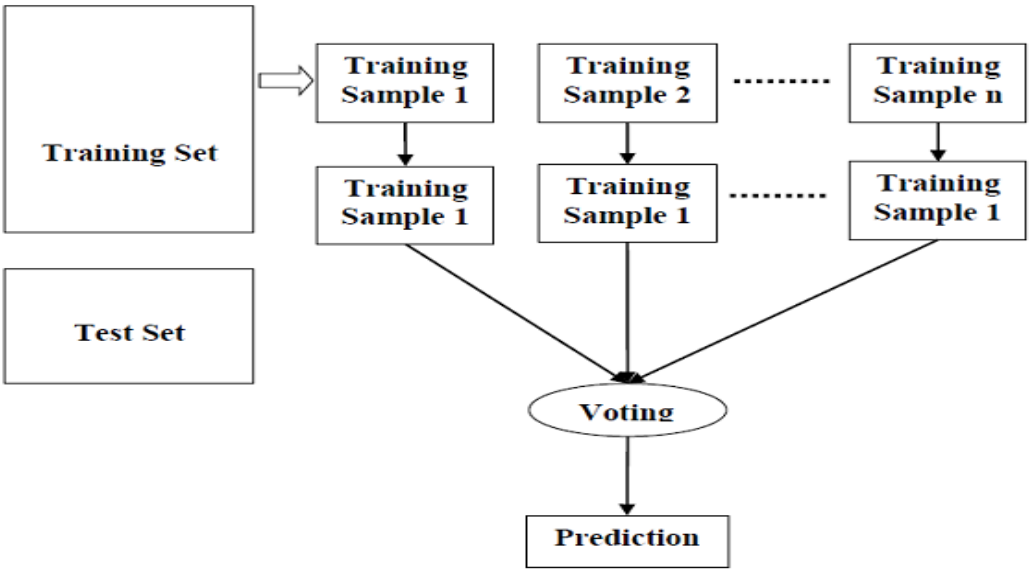
We can understand the working of Random Forest algorithm with the help of following steps

Step1- First, start with the selection of random samples from a given dataset.

Step2- Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

Step3- In this step, voting will be performed for every predicted result.

The following diagram will illustrate its working-



Tree:

Tree is a simple algorithm that splits the data into nodes by class purity. It is a precursor to Random Forest. Tree in Orange is designed in-house and handle both discrete and continuous datasets.

It can also be used for both classification and regression tasks.

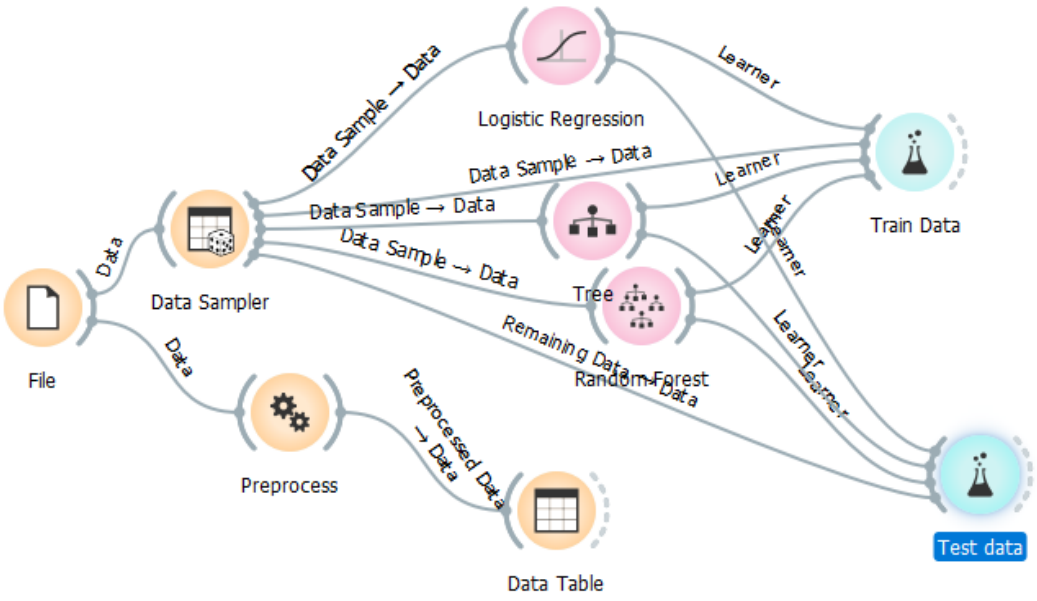
Tree parameters:

- Induce binary tree: build a binary tree (split into two child nodes)
- Min. number of instances in leaves: if checked, the algorithm will never construct a split which would put less than the specified number of training examples into any of the branches.

- Do not split subsets smaller than: forbids the algorithm to split the nodes with less than the given number of instances.
- Limit the maximal tree depth: limits the depth of the classification tree to the specified number of node levels.

Training The Models

Data is getting trained through the models (Logistic Regression, Random Forest, and Tree) and making the predictions for the test data.



Phase 5: Evaluation (Model in terms of accuracy)

<u>Evaluation results for Training Data:</u>

Model	AUC	CA	F1	Precision	Recall
Tree	0.629	0.748	0.500	0.538	0.466
Random Forest	0.820	0.792	0.572	0.643	0.516
Logistic Regression	0.849	0.804	0.608	0.661	0.562

The logistic regression model work better than the other two models random forest and Tree model. For the training set, the accuracies are 0.75 for Tree, 0.79 for Random Forest and 0.80 for Logistic Regression.

Evaluation Results for Test Data:

Model	AUC	CA	F1	Precision	Recall
Tree	0.596	0.735	0.450	0.478	0.425
Random Forest	0.796	0.767	0.484	0.557	0.429
Logistic Regression	0.833	0.790	0.541	0.613	0.484

The logistic regression model work better than the other two models random forest and Tree model. For the test set , the accuracies are 0.75 for Tree, 0.79 for Random Forest and 0.80 for Logistic Regression.

Confusion Matrix:

The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known. Since it shows the errors in the model performance in the form of a matrix, hence also known as an error matrix.

Some features of Confusion matrix are:

For the 2 prediction classes of classifiers, the matrix is of 2*2 table.
The matrix is divided into two dimensions, that are predicted values and actual values along with the total number of predictions.
Predicted values are those values, which are predicted by the model, and actual values are the true values for the given observations.

True Negative: Model has given prediction No, and the real or actual value was also No.

True Positive: The model has predicted yes, and the actual value was also true.

False Negative: The model has predicted no, but the actual value was Yes, it is also called as Type-II error.

False Positive: The model has predicted Yes, but the actual value was No. It is also called a Type-I error.

Accuracy:

The accuracy is the measurement of how many cases the model has correctly identified. It is one of the important parameters to determine the accuracy of the classification problems. It defines how often the model predicts the correct output. It can be calculated as the ratio of the number of correct predictions made by the classifier to all number of predictions made by the classifiers.

Accuracy = TP+TN/TP+FP+FN+TN

Precision:

It can be defined as the number of correct outputs provided by the model or out of all positive classes that have predicted correctly by the model, how many of them were actually true. It can be calculated as:

Precision=TP/TP+FP

Recall:

It is defined as the out of total positive classes, how our model predicted correctly. The recall must be as high as possible.

Recall=TF/TP+FN

F-**measure**:

If two models have low precision and high recall or vice versa, it is difficult to compare these models. So, for this purpose, we can use F-score. This score helps us to evaluate the recall and precision at the same time. The F-score is maximum if the recall is equal to the precision.

The F1 score is the harmonic mean of the precision and recall. The higher F1 score is, the better performance of the model (best value is at 1 — perfect precision and recall)

F-score= 2*Recall*Precision/Recall+Precision.

Sensitivity:

Sensitivity measures the proportion of actual positive cases that are correctly captured by the model

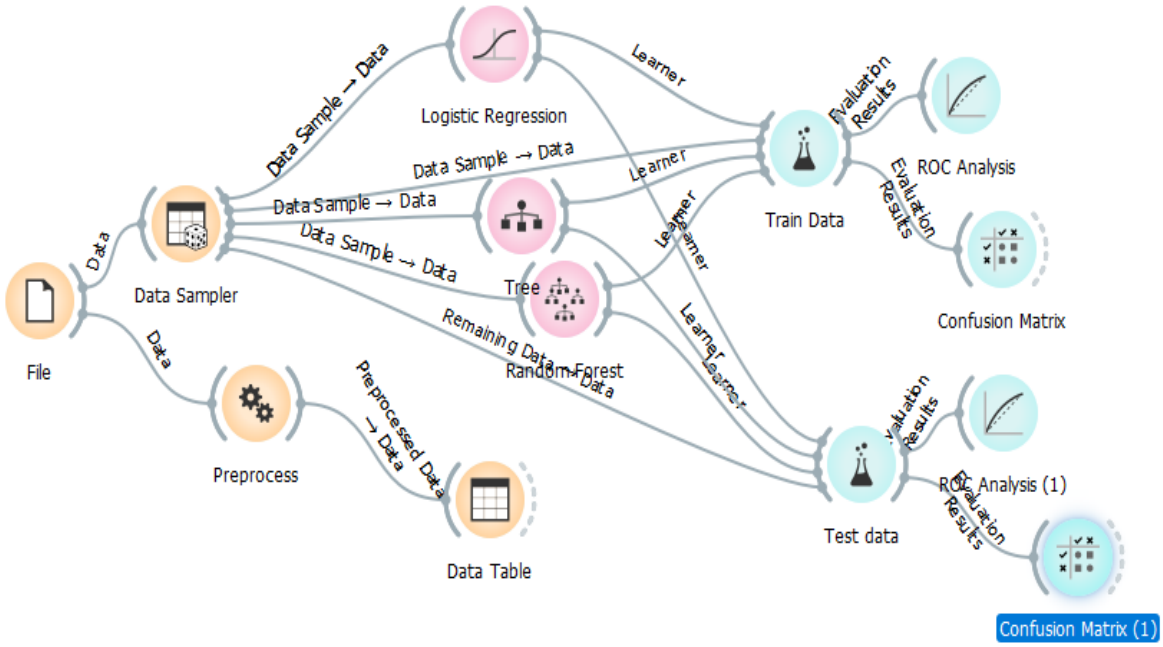
Sensitivity = TP/TP+TN

Specificity:

Specificity measures the ability to detect the actual negative cases correctly.

Specificity=TN/TN+FP

Implementation of confusion matrix and ROC analysis for Train Test Data:



Confusion matric for Logistic Regression of Training Dataset

		Predicted		
		No	**Yes**	**Σ**
Actual	**No**	3217	384	**3601**
	Yes	582	748	**1330**
	Σ	**3799**	**1132**	**4931**

There are actual and predicted values . It has been taken 4931 records for Training set.

True Positive (TP) = 3217 (These are cases in which we predict yes (they churned), and they did churn.

True Negative(TN) = 748 (We predicted no, and they didn't churn).

False Positive(FP) = 384 (We predicted yes, but they didn't actually churn, also known as Type I error)

False Negative(FN) = 582 (We predicted no, but they actually churned, also known as Type II error)

Confusion matric for Logistic Regression of Test Dataset

		Predicted		
		No	**Yes**	**Σ**
Actual	**No**	1408	165	**1573**
	Yes	278	261	**539**
	Σ	**1686**	**426**	**2112**

There are actual and predicted values. It has been taken 2112 records for Training set.

True Positive (TP) = 1408 (These are cases in which we predict yes (they churned), and they did churn.

True Negative(TN) = 261 (We predicted no, and they didn't churn).

False Negative(FP) = 165 (We predicted yes, but they didn't actually churn, also known as Type I error)

False Positive(FN) = 278 (We predicted no, but they actually churned, also known as Type II error)

ROC Curve:

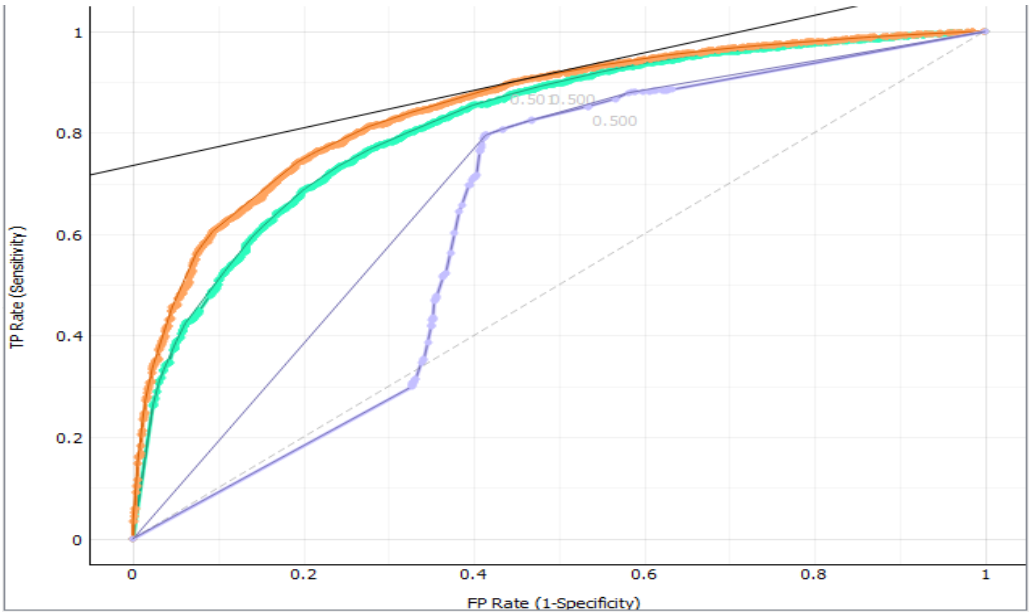
A receiver operating characteristic curve, or ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

ROC curve is a two dimensional graph in which the false positive rate(specificity) is plotted on the x-axis and the true positive rate (1-sensitivity) is plotted no the y-axis.

The ROC curves are useful to visualize and compare the performance of classifier methods.

The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the classifier.

ROC of the training dataset for all the models



This ROC curve is constructed by plotting the sensitivity against 1-specificity for each possible cut-off. The maximum values on x-axis and y-axis is 0 to 1. These are probability. So the area is equal to 1.

The diagonal connects from origin to other outer edge of the square. This diagonal divides the square exactly into two parts that is 0.5 and 0.5

Anything above 0.5 is the value addition that we are doing to this analysis. More the region better the model it is.

The ROC (Receiver Operating Characteristic) Curve is a relationship between True Positive Rate and False Positive Rate. Logistic Regression, which is the red line, has an area of 0.849 under the curve. It means the model has performed better.

The AUC of this model is between 0.5 and 1

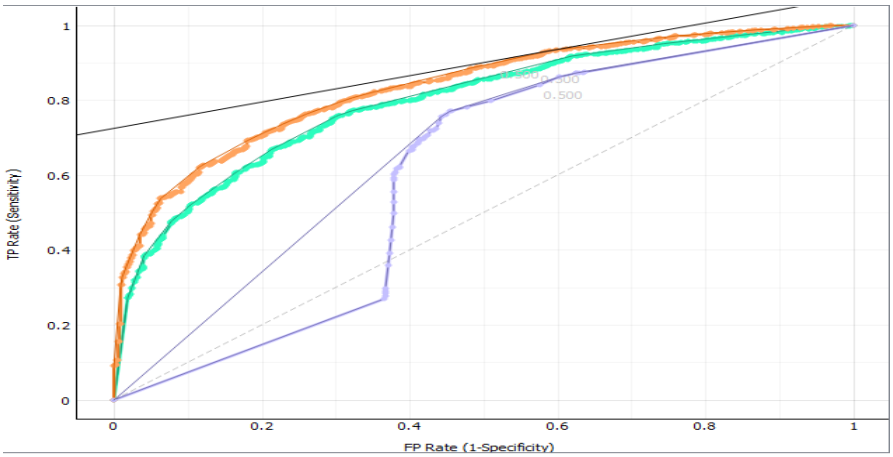
AUC ROC-curve of Logistic Regression = 0.849

AUC ROC-curve of Random Forest = 0.820

AUC ROC-curve of Tree = 0.629

We see that the model Logistic Regression leads to higher AUC and should be preferred over Models Tree and Random Forest.

ROC of the test dataset for all the models



The AUC of this model is between 0.5 and 1

AUC ROC-curve of Logistic Regression = 0.833

AUC ROC-curve of Random Forest = 0.796

AUC ROC-curve of Tree = 0.596

We see that the model Logistic Regression which is (red line) leads to higher AUC and should be preferred over Models Tree and Random Forest.

Conclusion

Telecommunication industry always suffers from a very high churn rates when one industry offers a better plan than the previous there is a high possibility of the customer churning from the present due to a better plan in such a scenario it is very difficult to avoid losses but through prediction we can keep it to a minimal level.

We identified several important churn predictor variables from these models and compared these models on accuracy measures.

Customers with month-to-month contracts are less likely to churn.

Customers with internet service, in particular fiber optic service, are more likely to churn.

Customers who have been with the company longer or have paid more in total are less likely to churn.

The Logistic Regression is found to be the best predictive model with good accuracy compared to other models for the data set taken.

Logistic Regression model helps to identity the probable churn customers and then make the necessary business decisions.