Customer Churn Analysis

# Problem Definition:

Customer churn is when a company’s customers stop doing business with that company. Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales budgets to gain additional customers. Existing customers will often have a higher volume of service consumption and can generate additional customer referrals.

Customer retention can be achieved with good customer service and products. But the most effective way for a company to prevent attrition of customers is to truly know them. The vast volumes of data collected about customers can be used to build churn prediction models. Knowing who is most likely to defect means that a company can prioritize focused marketing efforts on that subset of their customer base.

Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low.

You will examine customer data from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models.

# 

# 

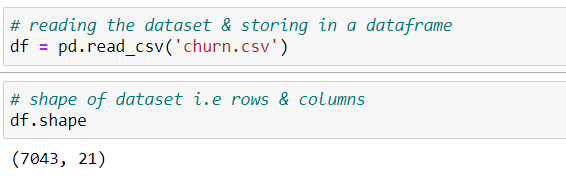
# Data Analysis:

In Machine Learning, Data Analysis is the process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information by informing conclusions and supporting decision making*.*

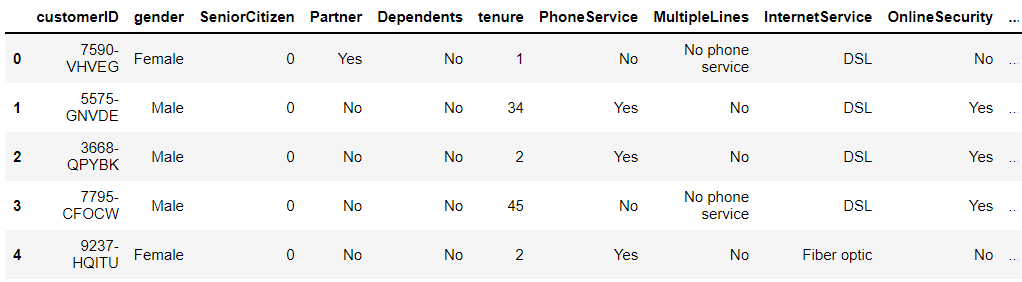
*In the Data Analysis phase we understand the data values lie in different features.As Churn analysis helps you identify pain points throughout the entire customer journey. Understanding those pain points then opens up avenues to improve your products, services, and communication.*

*In the Data Analysis phase we discover the shape and the data types of different columns/features.*

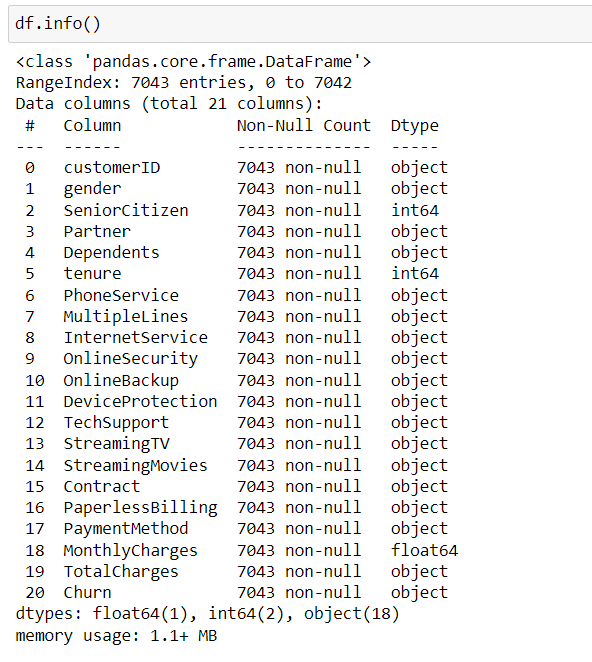
*In Below code snippet I have just stored dataset in a dataframe & also checking the shape of dataset.*

**

*Dataset:-*

**

*Here we can also identify different columns, data types , non-null values etc.*

**

*Here we can also check for null values , constant , quasi-constant columns and also find the correlation between independent features*

*and dependent features. I have discussed them in the Data Preprocessing section.*

# EDA Concluding Remark:-

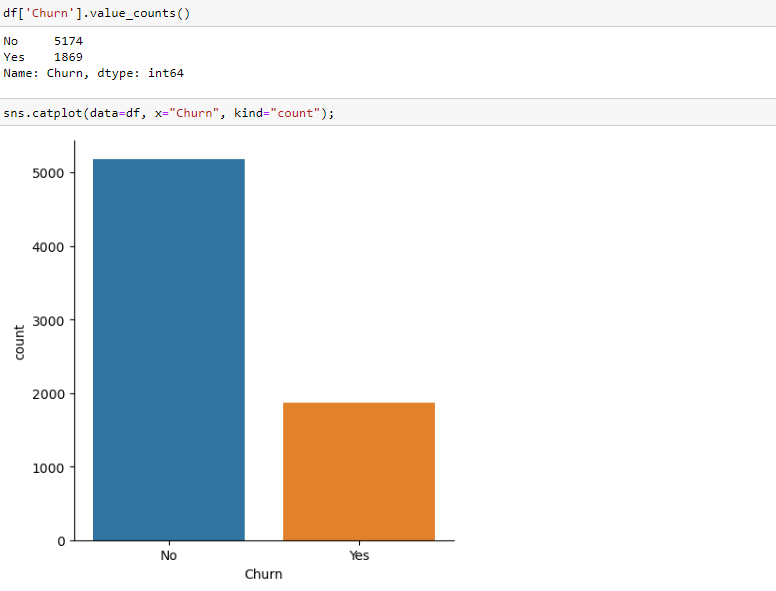
**Exploratory data analysis** is the process of analysing the main characteristics of a data set, typically using visualisation techniques and summary statistics. The goal is to comprehend the data, identify patterns and anomalies, and validate assumptions before proceeding with further analysis.

EDA makes it simple to comprehend the structure of a dataset, making data modelling easier. The primary goal of EDA is to make data ‘clean’ implying that it should be devoid of redundancies. It aids in identifying incorrect data points so that they may be readily removed and the data cleaned. Furthermore, it aids us in comprehending the relationship between the variables, providing us with a broader view of the data and allowing us to expand on it by leveraging the relationship between the variables. It also aids in the evaluation of the dataset’s statistical measurements.

## **Here, some of the visualization from the project notebook.**

## **Distribution of Churn**

Let’s first understand our target variable Churn .

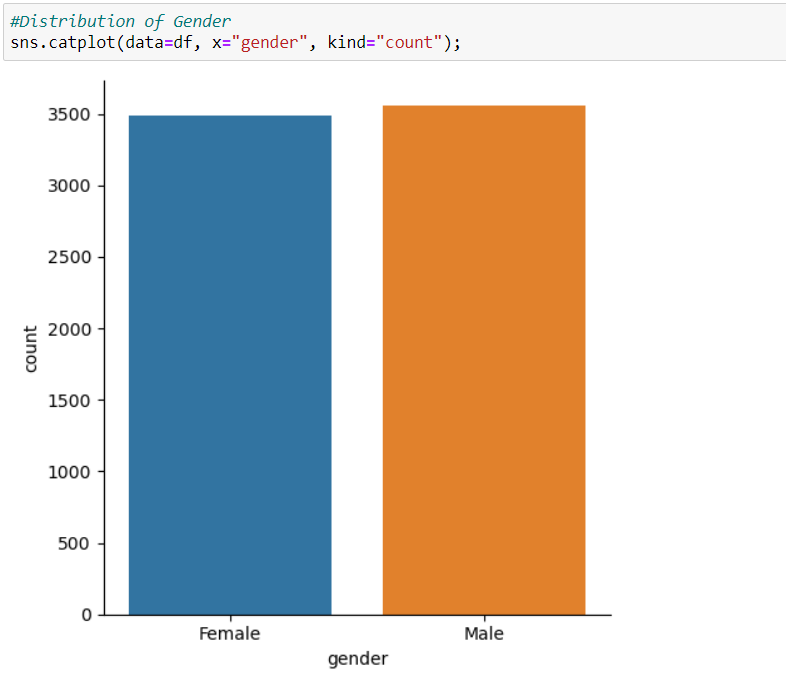


When we use code value count() to determine the number of values in each category, we find that customers with churn status No have 5174 and those with Yes have 1869.

Our goal for this module is to visualise the data as well as see the numbers; we will plot our results using the code below

.

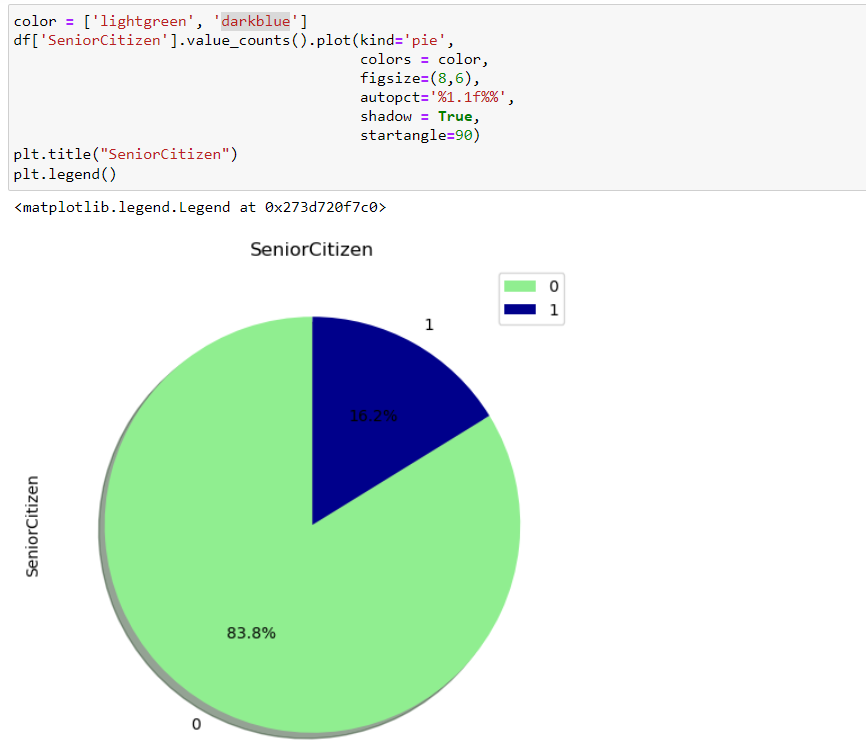
Output: we have written a code for a simple vertical bar graph. This is how the graph looks like.



We see that the proportion for the Yes category is very less as compared to No.

**Distribution of SeniorCitizen**

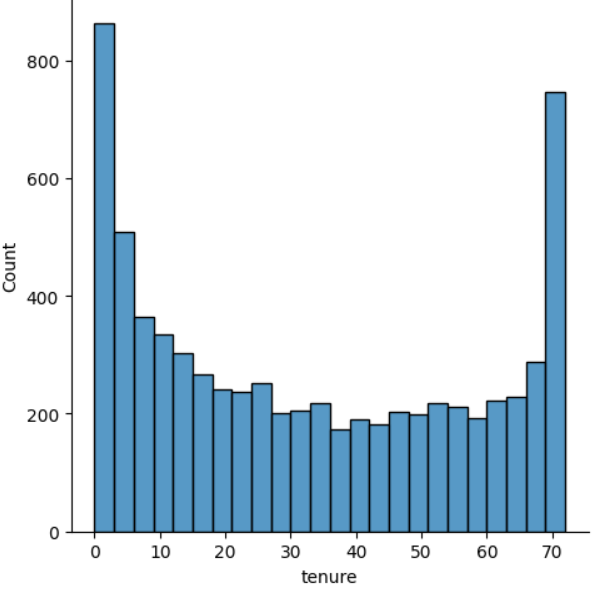
Let’s look at the percentage of senior citizens in each of these dataframes.



Here, we see that 16.2% customers were seniorcitizen, while the remaining 83.8% were not.

## **Distribution of Tenure**

For our analysis, we employ the column tenure. Only customers who have left the company will have their tenure period displayed.

****

# Pre Processing Pipeline :-

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

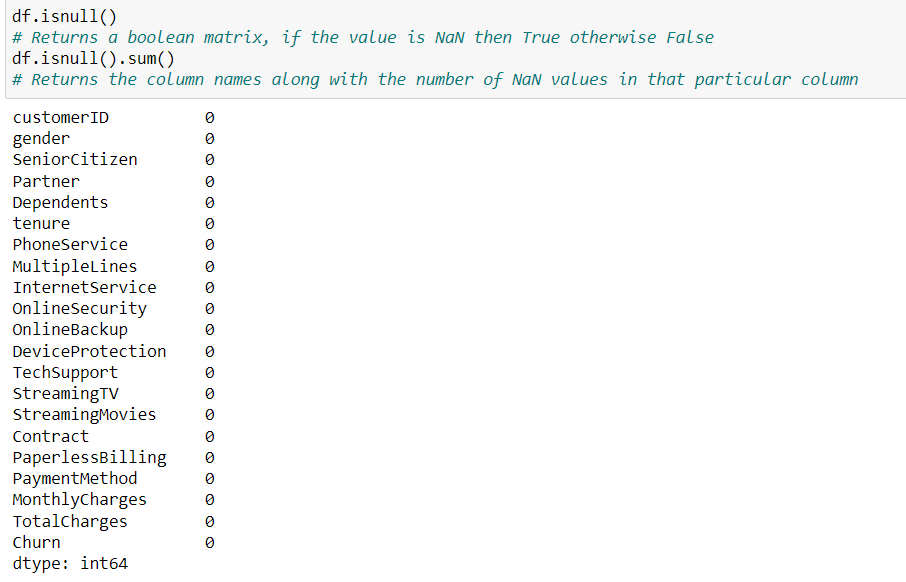
In Data Preprocessing phase we have some common steps to follow:-

1. Handling Null Values
2. Standardization
3. Handling Categorical Variables
4. One-Hot Encoding
5. Multicollinearity

# **Handling Null Values —**

In any real-world dataset, there are always few null values. It doesn’t really matter whether it is a regression, classification or any other kind of problem, no model can handle these NULL or NaN values on its own so we need to intervene.

Checking for null values:-

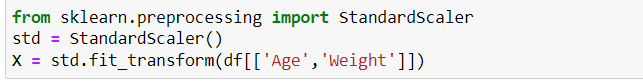


There are various ways for us to handle this problem. The easiest way to solve this problem is by dropping the rows or columns that contain null values.

# Standardization:-

It is another integral preprocessing step. In Standardization, we transform our values such that the mean of the values is 0 and the standard deviation is 1.

Consider the above data frame, here we have 2 numerical values: **Age** and **Weight**. They are not on the same scale as Age is in years and Weight is in Kg and since Weight is more likely to be greater than Age; therefore, our model will give more weightage to Weight, which is not the ideal scenario as Age is also an integral factor here. In order to avoid this issue, we perform Standardization.



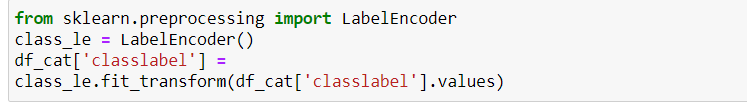
# Handling Categorical Variables:-

Handling categorical variables is another integral aspect of Machine Learning. Categorical variables are basically the variables that are discrete and not continuous. Ex — color of an item is a discrete variable whereas its price is a continuous variable.

Categorical variables are further divided into 2 types —

* **Ordinal categorical variables** — These variables can be ordered. Ex — Size of a T-shirt. We can say that M<L<XL.
* **Nominal categorical variables** — These variables can’t be ordered. Ex — Color of a T-shirt. We can’t say that Blue<Green as it doesn’t make any sense to compare the colors as they don’t have any relationship.

We can use different encoding technique for categorical variable one way is using One Label Encoding.



# **One-Hot Encoding —**

## So in One-Hot Encoding what we essentially do is that we create ’n’ columns where n is the number of unique values that the nominal variable can take.

## Ex — Here if color can take Blue,Green and White then we will just create three new columns namely — color\_blue,color\_green and color\_white and if the color is green then the values of color\_blue and color\_white column will be 0 and value of color\_green column will be 1 .

## So out of the n columns, only one column can have value = 1 and the rest all will have value = 0.

# **Multicollinearity and its impact —**

Multicollinearity occurs in our dataset when we have features that are strongly dependent on each other. Ex- In this case we have features -

color\_blue,color\_green and color\_white which are all dependent on each other and it can impact our model.

# 

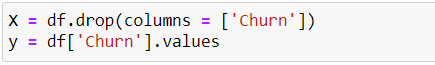
# Building Machine Learning Models:-

Building ml model is one of the crucial step as all the previous steps depend upon this step.

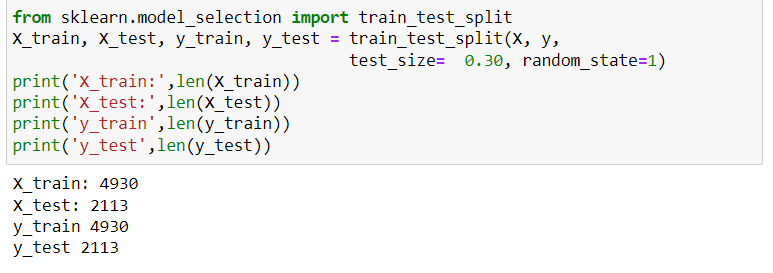
We need to create generalized & robust model.

We can try out different ml model and choose one which is giving better accuracy.

First, we create a variable X to store the dataset’s independent attributes. In addition, we define a variable y to hold only the target variable.



Then, from the sklearn.model\_selection package, we can use the train\_test\_split function to generate both the training and testing sets.



Some important factors that come into play when choosing the right machine learning algorithm include:

1.Level of accuracy needed

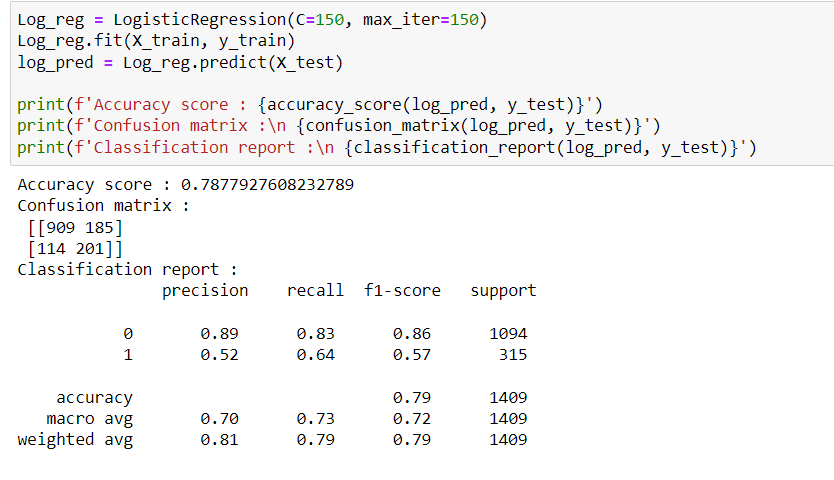
2.The time required to train the model(s).

3.The number of Features in your dataset

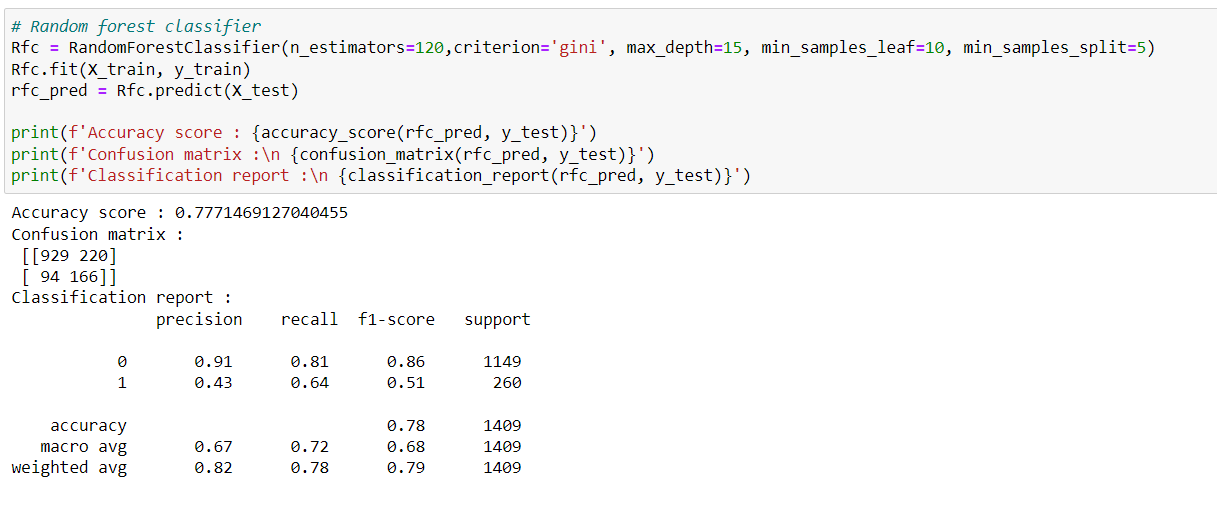
4.Linearity of your data

We can choose different ml algorithm and choose one which one is performing best.

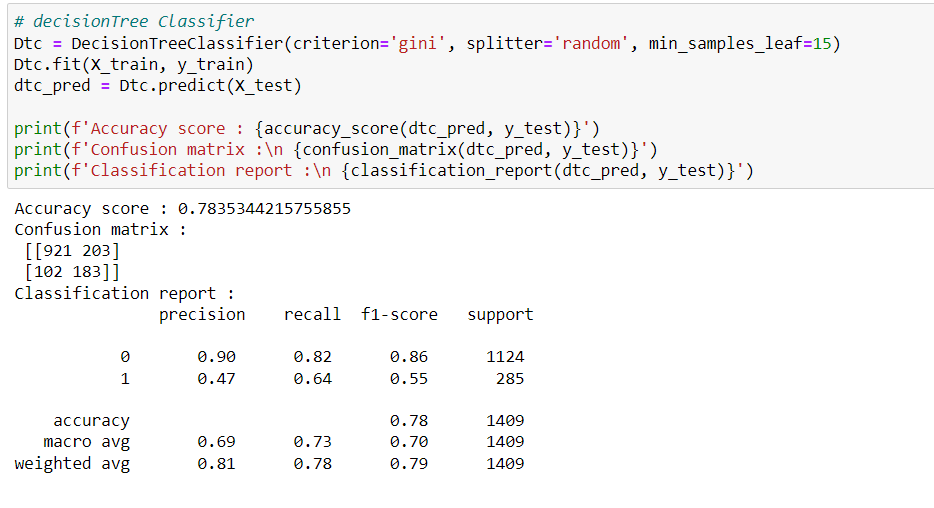
**Logistic Regression**

****

**Random Forest**

****

**Decision Tree**

****

We can also try out other algorithms and which one is performing best.

# Concluding remarks:-

In this post, i have walked through a complete end-to-end machine learning project using the customer Churn Analysis dataset. We started by cleaning the data and analyzing it with visualization. Then, to be able to build a machine learning model, we transformed the categorical data into numeric variables (feature engineering). After transforming the data, we tried 3 different machine learning algorithms using default parameters. Finally, we tuned the hyperparameters of the Decision Tree.

## 