# Telco Customer Churn

**Aim:** To predict the customers who are likely to churn in the next N months & facilitate in taking business actions for reducing the churn.

**Objective:**

In any service providing industry, when a customer decides to stop using the service either by cancelling the subscription or not paying for the service, we call this customer churn.

Churn is defined as how many customers are not using the service for a certain period.

Hence, customer churn is one of the essential metrics that every business must evaluate to grow. The churn rate is calculated by dividing the number of lost customers by the last number of customers. Thus, a company churn rate must be as low as possible, ideally 0%.

But why is it so important to calculate the churn rate? Does it affect the business if you lose around 5% of customers? Yes, the answer is that it costs more to acquire a new customer than retain the existing customers. Retaining the current customers, any company can spend less on operating costs needed to reach new customers.

So, we will use advanced machine learning techniques to predict the potential churners who are about to leave a company’s service and take the necessary steps to prevent it.

This project aims to build a deep learning model that will help predict customers who are likely to churn in the next N months and facilitate in taking business actions for reducing the churn.

**Data Description:**

The available dataset is Telco-Customer-Churn

This dataset has 7043 rows and 21 columns present.

The 21 features of this dataset are as follows:

1. Churn – the target variable, if the customer is churned or not (Yes / No)

2. customerID – The unique identification of every customer

3. gender- If the customer is a male or a female (Female / Male)

4. SeniorCitizen – If the customer is a senior citizen or not (0 / 1)

5. Partner – If the customer has a partner or not (Yes/No)

6. Dependents – If the customer has any dependents (Yes / No)

7. Tenure – The time period(months) the customer has stayed with the company.

8. PhoneService – If the customer has a phone service or not (Yes/No)

9. MultipleLines – If the customer has multiple lines or not (Yes/No/No Phone service)

10. InternetService – If the customer has any internet service or not (DSL/ Fibre optics/ No)

11. OnlineSecurity – If the customer has any online security (Yes/No/No internet service)

12. OnlineBackup – If the customer has any online backup (Yes/No/No internet service)

13. DeviceProtection – If the customer has device protection (Yes/No/No internet service)

14. TechSupport – If the customer has tech support (Yes/ No/ No internet service)

15. StreamingTV – If the customer has any streaming TV (Yes/ No/ No internet service)

16. StreamingMovies – If the customer has streaming movies (Yes/ No/ No internet service)

17. Contract – The customer term period with the company (Month-to-month, One year, Two years)

18. PaperlessBilling – If the customer has paperless billing or not (Yes/ No)

19. PaymentMethod – The payment mode of each customer (Electronic check, mailed

check,Bank transfer, Credit card)

20. MonthlyCharges – The amount that is charged to the customer every month

21. TotalCharges – The total amount charged to the customer

Contents:

1. [Dataset Information](#Dataset_info)
2. [Exploratory Data Analysis (EDA)](#EDA)
3. [Feature Engineering](#FeatureEnginnering)
4. Modeling
5. [Conclusion](#Conclusion)
6. **Dataset Information:**
   * Importing the common libraries such as numpy, pandas, matplotlib, seaborn.
   * Importing and loading the Dataset.
   * Viewing the dataset.
   * Fetching the information about the data, i.e dtype, null values if any.

# Importing the libraries and Loading and viewing the dataset

# 

# Checking information about the datatype

# 

# 

# Changing the Total Charges column from Object to Numeric as it has numeric data

# 

# Checking the Missing values

# 

There are 11 missing data in the TotalCharges field, so getting rid of those rows from the dataset so dropping those null values of Total charges

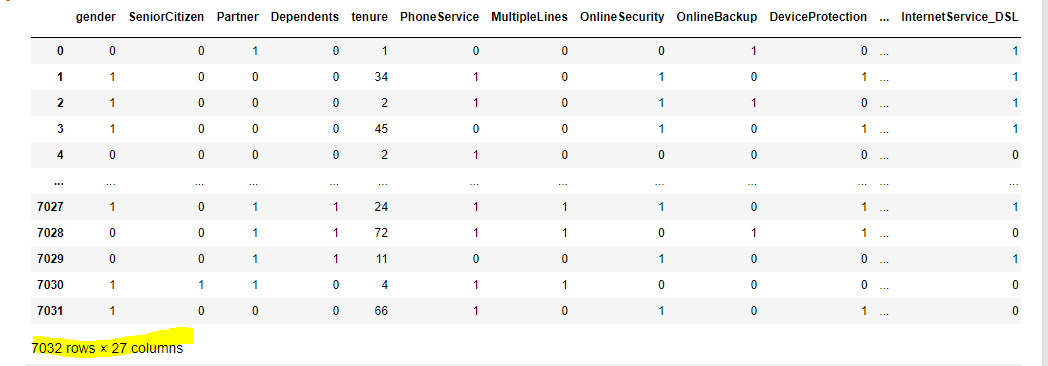
## 

## There are no duplicate values in the dataset.

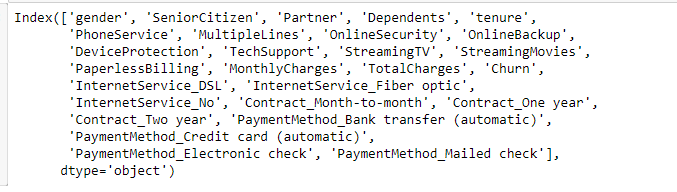
## There are three continuous variables and they are Tenure, MonthlyCharges and TotalCharges. SeniorCitizen is in 'int' form, that can be changed to categorical.

**Exploratory Data Analysis**

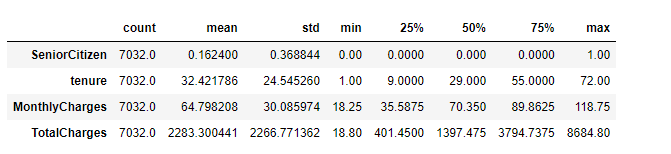
Checking the number of rows and columns in the data:- In this dataset we have 7032 rows and 27 columns.



Columns Name in the Dataset



**Checking the Corealtions**



# 

# Outlier Analysis:- There are no Outliers in the dataset

# 

# We need to take care of the Output i.e. Churn. We need to map the values and cast them to the Columns

# 

# We can see the values are now in 0's and 1's

# Checking the total number of Categorical Features in the Data

# 

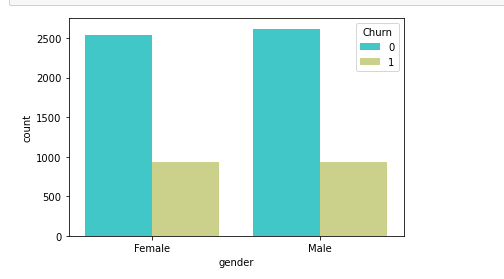
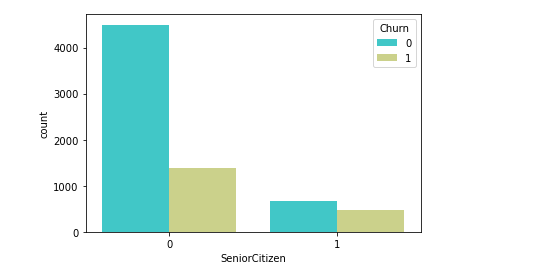
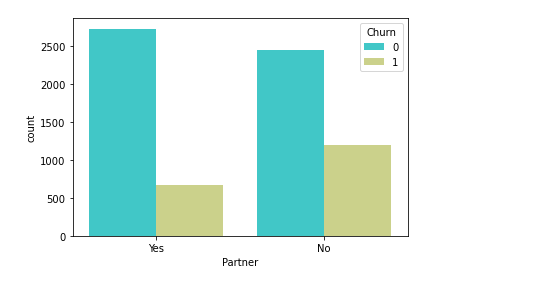
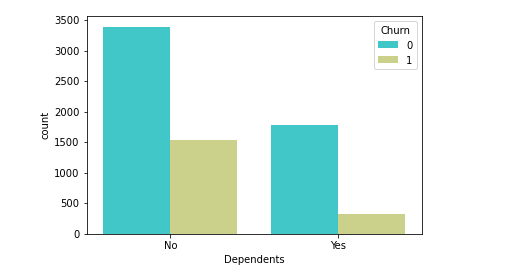
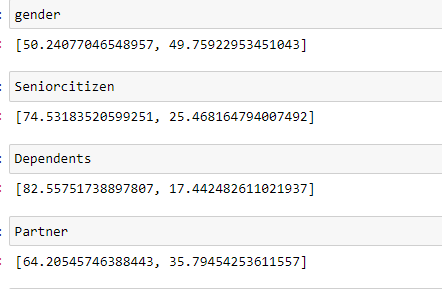
**Customer Information with respect to Churn.**

Analyzing gender

Analyzing Senior Citizen

Analyzing Partner

Analyzing Dependents

1. We can see from the above information that as per the 'Gender' both male and female are almost equal. Female: 50.24 and male:49.75. Hence this doesn’t show much of information.

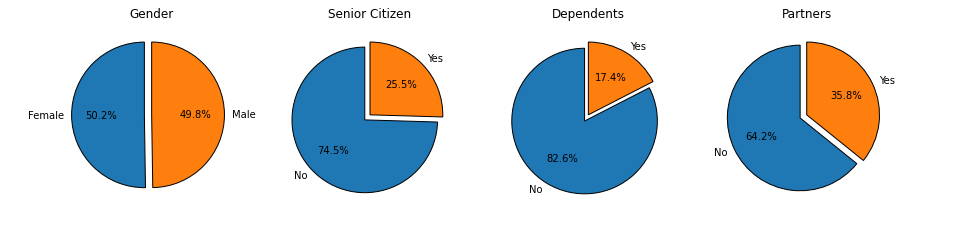
2. The customer who are not Senior Citizens are churning more than the senior Citizens. Non\_senior Citizens:74.53 and Senior Citizens: 25.46

3. The people who have dependents have more chances of staying back than the ones who don't. No Dependents: 82.55 and Dependents:17.44

4. The people who have partners are more likely to churn than the ones who dont have. Partner: 64.20 and No-Partner:35.79

**Customer Information with Relation to Churn**

After using the group by function we have collated the data for the Customers with regards to if the customers will Churn or not. The below Cell and the pie charts show a visualization of the same



Insights from the above Chart

1. Gender: We can see that there are equal chances of both Male and Female to churn as the % is almost the same.

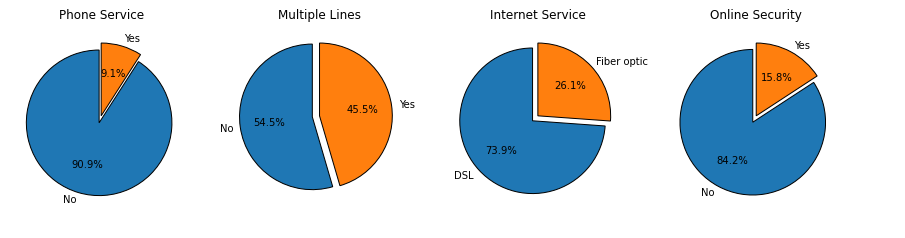
2. Coming to Senior Citizen: It can be seen that the people who are not Senior Citizens are likely to Churn more.

3. As far as Dependents are concerned the people who don’t have any dependents are likely to Churn more as they are at around 82%.

4. The people who are single or who don’t have partners are likely to Churn than the ones who have

# Checking to confirm the impact of the services on the Churn

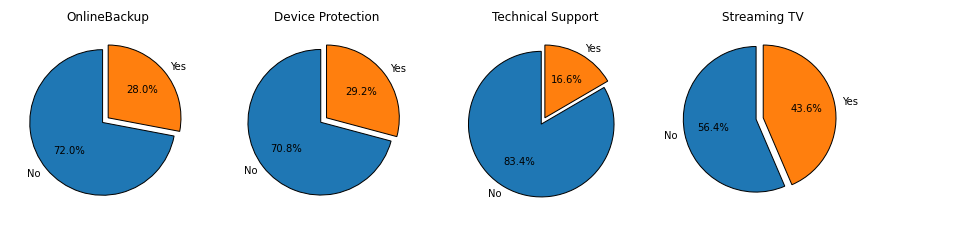
There are various services which are being provided. This visualzation and group by is being used to check as to what is the impact of the services being provided and what is the impact if the services are not being provided.



# Insights from the above Chart

1. We can see that all the services are impacting the customer churn. If the customers are not getting Services then they are likely to churn which is a fact in real world as well

2. Coming to the Internet Services offered we can see that the customers who have DSL are more likely to churn. So the company should work on moving those customers to Fiber Optic.



**Brief Explanation about the Services**

1.Online Backup: This is the backup which is done on the cloud to ensure that there is no loss to the data and if the custoimer is changing the phone then he can get all the data back.

2.Device protection: Protection against theft or loss of device or even Screen damage.

3.Technical Support: The technical support should be available so that if there is any problem in the device or maybe in the network then the customer should not be impacted

4.StreamingTV: Is nothing but Online Streaming of the channels.

# Coming to the Insights from the above Visualization

We can clearly see that if there are no services which are being provided then high number of customers will be churning.

# 

StreamingMovies: Online Streaming of Movies.

We can see that there is not much of a difference in the customer leaving or staying with regards to the Streaming Movies but still we have more customer who will churn if there are no online Streaming Movies

# Payment Information and analysis with respect to Churn

There are columns for Payment related that include Contract,PaymentMethod,Total Charges,Monthly Charges,Paperless Billing for which we will be visualizing the data to understand better.

# Description of the Columns

1.Contract: Whether its a monthly,Yearly or Two-Year and its impact on the Churn

2.PaymentMethod: Various types of payment methods provided and their impact on the churn

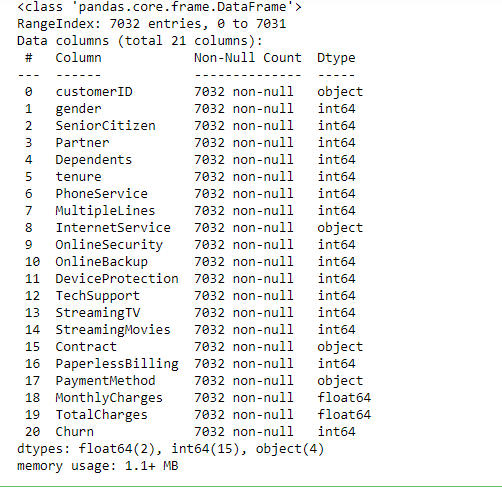
3.Paperless Billing: Whether the bills that the customers are receiving are hard copies or else on mails

4.Total Charges: Total charges that are being charged to the customer

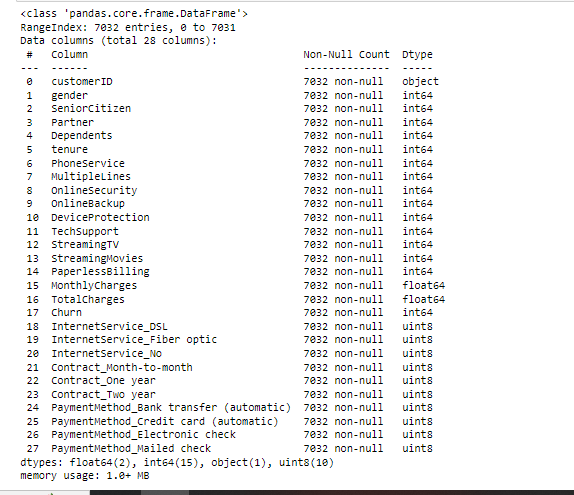
5.Monthly Charges : That are charged to the customer

**Converting all the data to Numeric from Categorical:-**

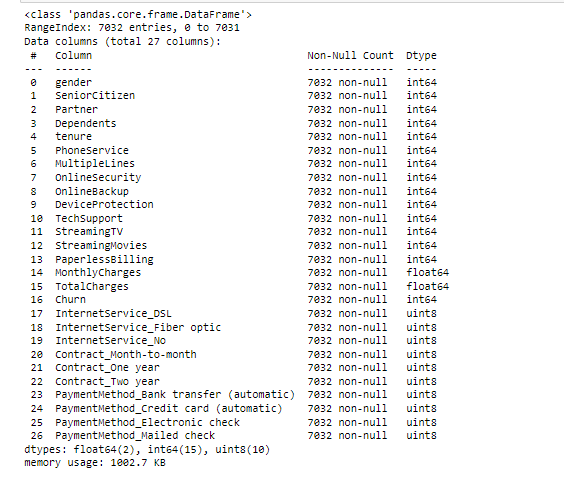
The data types of the categorical columns having Yes and No have been converted to Numeric Column

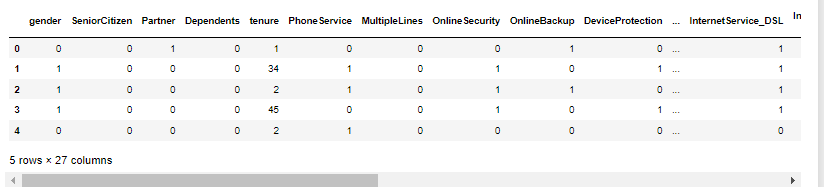


Converting the remaining columns('InternetService','Contract','PaymentMethod’) by using One Hot Encoding

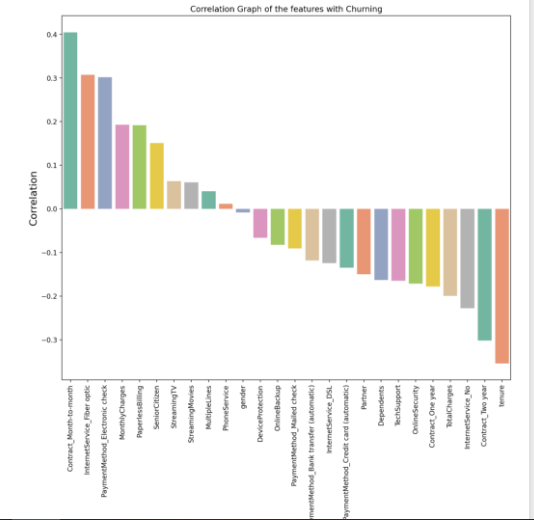


We drop column "customerID", since it won’t' impact anything.





barplot of the correlated features



**Modeling**

Lets first import few libraries like train\_test\_split, roc\_auc\_score, f1\_score, precison\_score, recall\_score., etc

# Imbalance treatment

# Data Balancing using SMOTE :

# 

In problems like this, where the target class ("churn") is imbalanced (there are much more 0's than 1's), we can achieve better results if we conduct some kind of imbalance treatment. Here we will use Smote technique to balance the data.

Synthetic Minority Oversampling Technique (SMOTE) is a statistical technique for increasing the number of cases in your dataset in a balanced way.

# 

# Separating dataset into X and y

## Using the Chi-Square test to check which columns can be removed



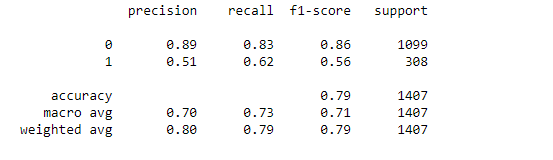
Importing the Logistic Regression Algorithm as this is a classification problem

Fitting the X\_train and y\_train in the Model

Predicting using Model

Evaluating using the Accuracy as this is a classification problem

Import Confusion Matrix





# Running the Model with Random Forest

# 

# 

# Running the Model with Decision Tree

# 

# KNN Classifier

# 

# 

# 

# naive\_bayes

# 

# 

# 

# 

# 

# 

# ****Checking the AUC for all the models:****

# 

# 

# ****A brief Summary of all the models:****

# Logistic Regression: Accuracy 0.83%

Decision Tree: 0.66%

Random Forest: 0.82%

K Nearest Neighbor: 0.79%

The Logistic Regression model (accuracy – 0.83%) gives better accuracy with respect to the others model.