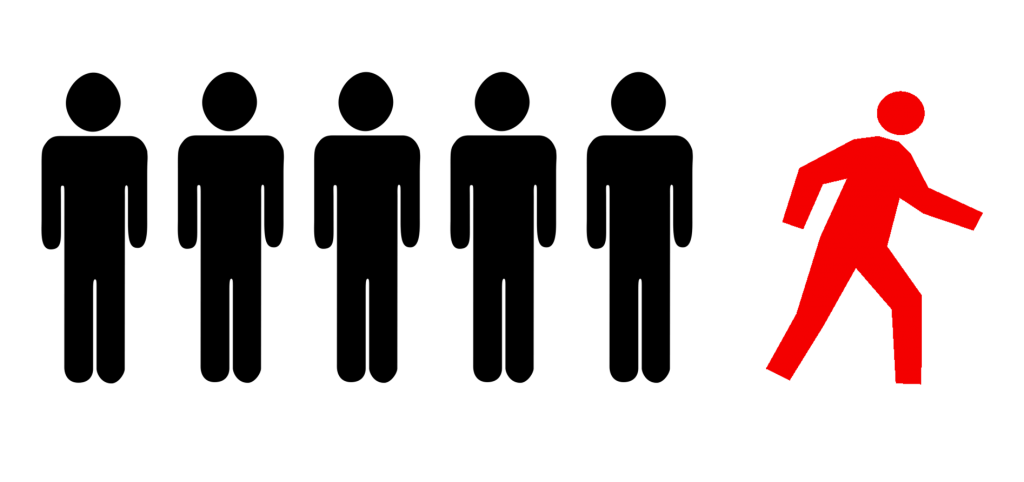
**Customer Churn Analysis**

****

In this blog-post, I will go through the whole process of creating a machine learning model on the famous Telecom customer churn dataset. We will examine customer data from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models.

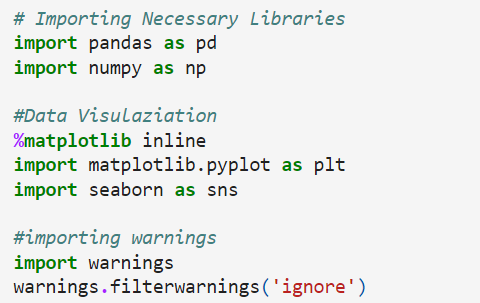
**Problem Statement:**

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 prioritise focused marketing efforts on that subset of their customer base.

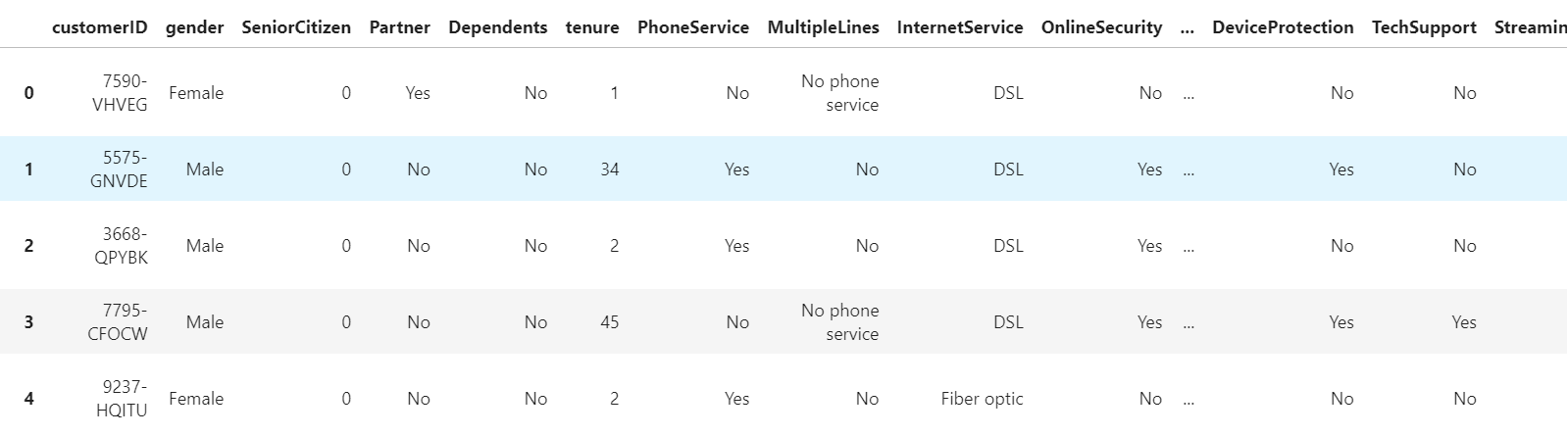
Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low.

**Importing Libraries**

****

# Importing dataset

# 

****

But, dataset contains large number of columns and we can’t see all columns here.

# For displaying all columns we use this code.

# Here we can see all the columns present in the dataset. Here "Churn" is the target variable which contains 2 categories so it will be termed as "Classification problem" where we need to predict the several churn using the classification models.

# Exploratory Data Analysis (EDA)

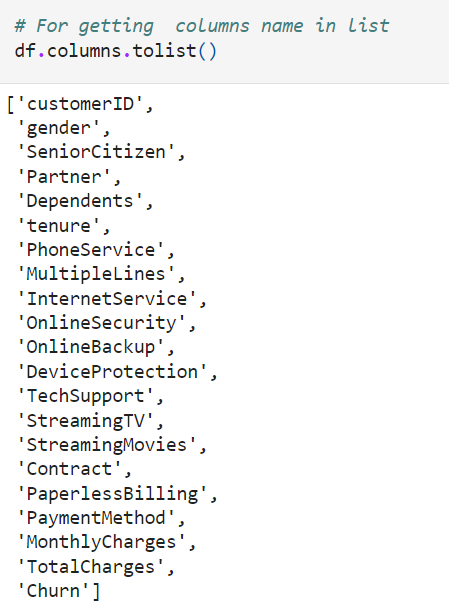
# *Checking the shape of dataset*

# 

# This dataset contains 7043 rows and 21 columns.

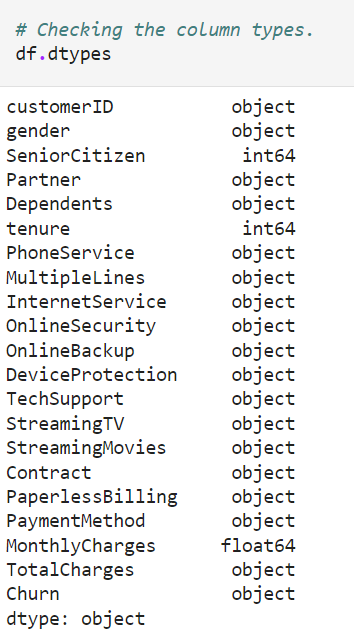
# *Checking the columns*

# We can see all the column names of the dataset with the help of this code.



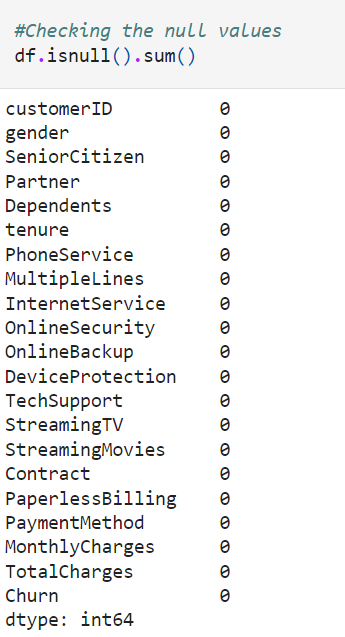
We can see all the columns present in dataset. Here 'Churn' is the target variable and remaining 20 are independent variables.

*Checking data type*



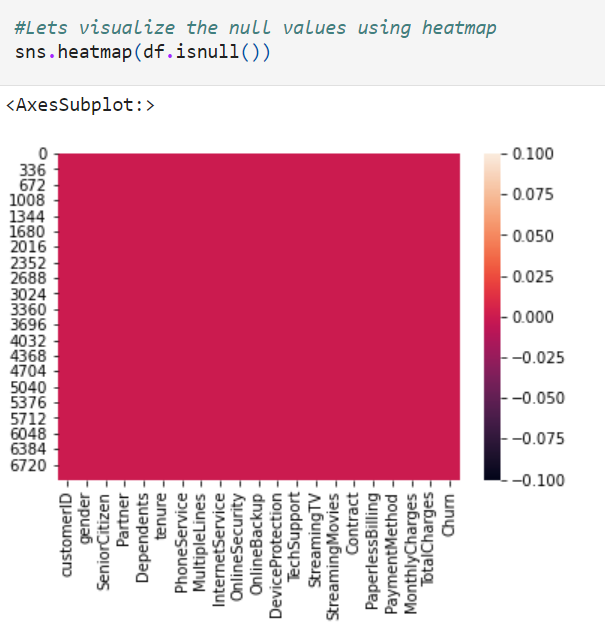
This will show the datatypes of all the columns. This dataset contains ‘object’, ‘int64’ & ‘float64’ datatypes.

*Checking the Null values*

****

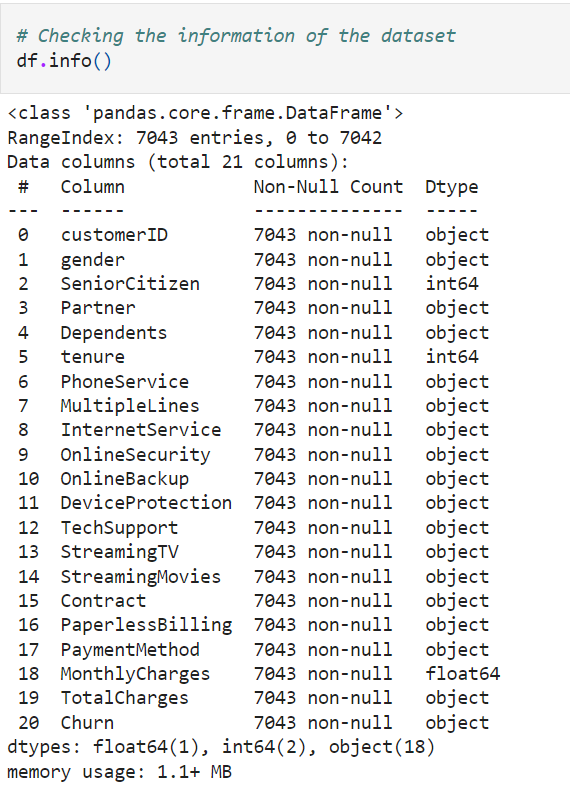
*No null values is present in our dataset.*

*Visualizing the Null values*

**

We can se that there se no white spaces present in heatmap, it means the dataset is free from null values.

*Checking the information of the dataset*

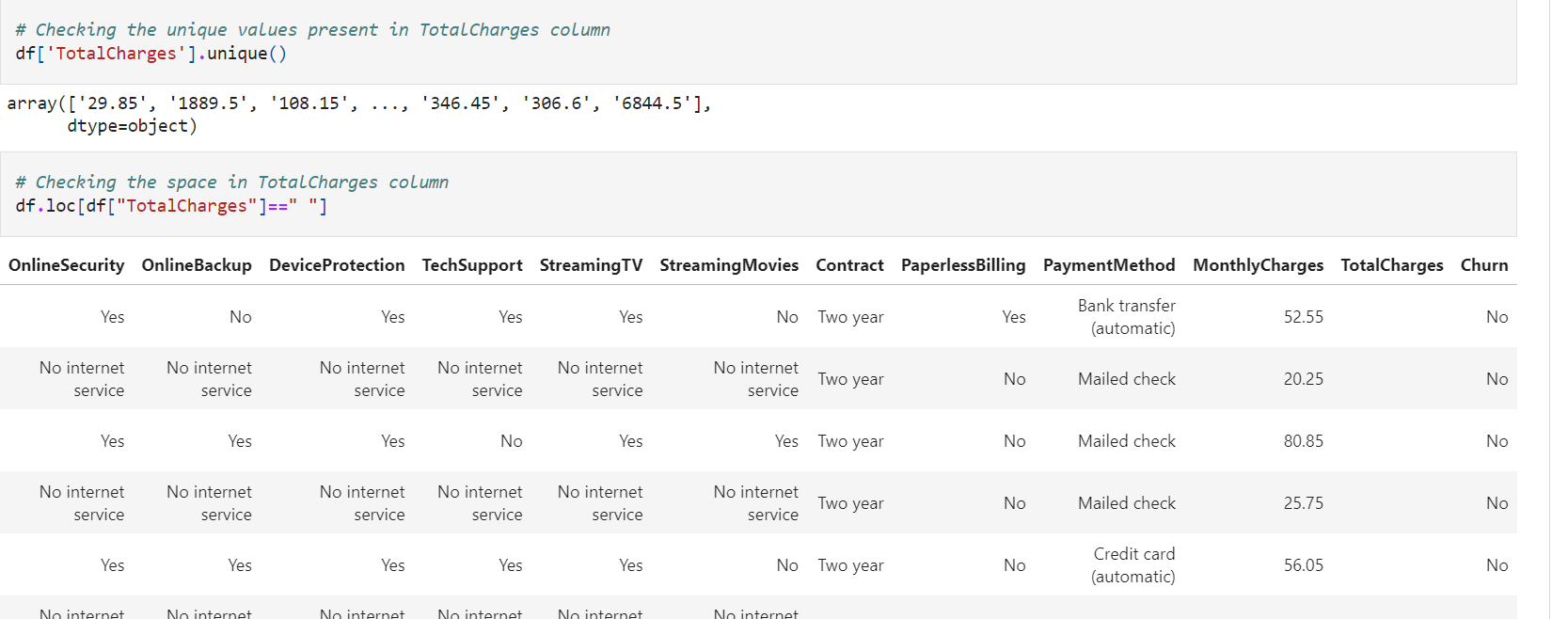
**

This gives the complete information about dataset i.e. columns, null counts and datatype. The dataset contains 7043 rows and we can see that all the columns have 7043 entries, so we can easily observe that no null values is present in the dataset.

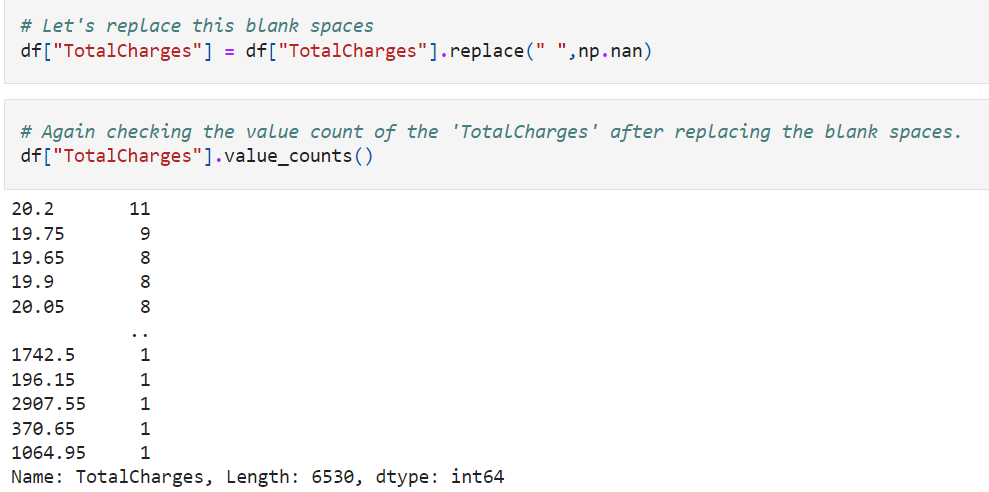
*Checking the value counts of each columns*

**

This shows the value counts of each columns and gives the idea about the unique values.



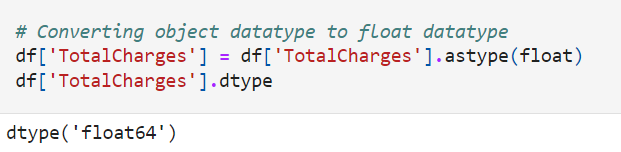
We can see that "TotalCharges" contain blank data. So we can treat this values after checking for unique values.

*Replacing blank spaces with ‘NaN’* 

As we can see that the blank spaces is successfully replaced.

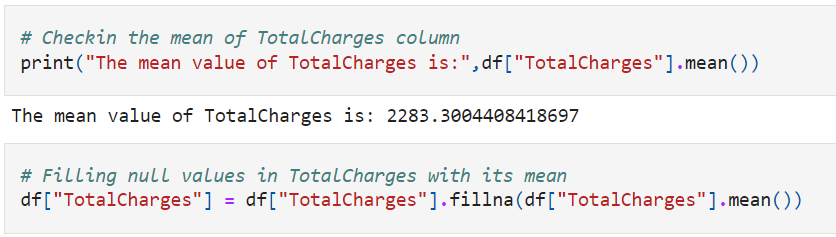
As we can see that 'TotalCharges' contains float values but we are getting its datatype as Object. so we need to convert it to float datatypes.

*Converting datatype from ‘object’ to ‘float’*

**

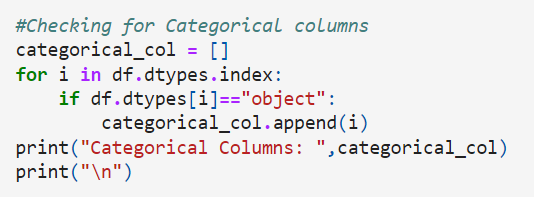
We have converted the datatype of "TotalCharges" from object to float

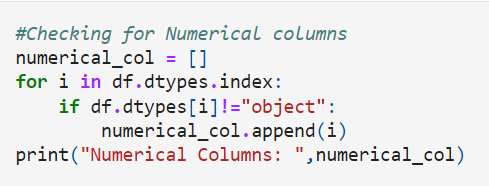
*Again checking the null values and after getting null values present in ‘TotalCharges’ column, we will fill this null values with the mean values.*

**

*Now, again we will check the null values after filling the null values. If there is no null values present in dataset then we will move further.*

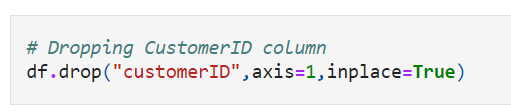
*Now separating the Numerical and Categorical data*

**

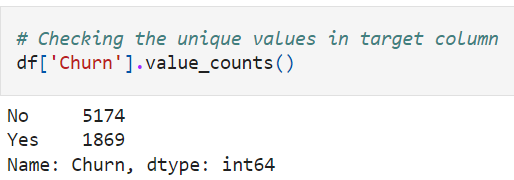
**

*We can get numerical and categorical data separately with this code.*

*As we can see that the customer id is unique for each customer therefore we will replace this column from the dataset.*

**

*Checking the value counts of the target variable*

**

Only two categories is present in target column i.e. 'Yes & No'.

"No" stands for the customers who have not churned and "Yes" stands for the customers who have got churned from the company.

***Description of Dataset***

# 

The summary of the dataset looks perfect since there is no negative/incvalid values present.

**Observations:**

The counts of all the columns are same which means there are no missing values in the dataset.

The mean value is greater than the median(50%) in tenure and TotalCharges columns which means the data is skewed to right in these column.

The data in the column MonthlyCharges have mean value less than median which means the data is skewed to left.

By summarizing the data we can observe there is a huge differences between 75% and max hence there are outliers present in the data.

# Data Visualization

# *Univariate Analysis*

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# *After doing univariate analysis we observe the following:*

1. There are more number of customers who have not churned than churned customer. This leads to class imbalance issue in the data. we will rectify it by using oversampling method in later part.
2. we can observe that the total number of males and females are almost same.
3. Number of senior citizen customers are 1142 while non senior citizen customers are 5901.
4. Customers having partners are 3402 while customers having no partners are 3641.
5. There are 2110 customers having dependents whereas 4933 customers not having dependents.
6. There are 6361 customers having phone services whereas customers who do not have phone services are only 682.
7. The customers having phoneservices from single line have high counts compared to the customers having phone services from multiple lines, also the customers who do not have phone services have covered very less data compared to others.
8. There are 3498 customers who do not have any online security.
9. We can observe that the customers having no internet services have very less online backup counts compared to others.
10. We can see that the customers who do not have internet service, they do not need any device protection.
11. There are 3473 customers who don't need any techsupport.
12. There are 2810 customers who do not use StreamingTV.
13. The customers who do not have Streaming movies are high in count followed by the customers who have Streaming movies services. And the customers who do not have internet services, they have less streaming movies services compared to others.
14. There are 4171 customers prefer paperless billing and 2872 customers do not prefer paperless billing.
15. The customers who prefer Electronic check payment method are very high in numbers whereas average number of customers prefer Mailed Check, bank transfer and Credit card payment method.

# *Plotting numerical columns*

# 

# 

**Observations:**

1. The two columns tenure and MonthlyCharges do not have skewness.
2. TotalCharges is skewed to the right.
3. The data almost looks normal in all the columns except SeniorCitizen.

# *Bivariate Analysis*

# 

# 

# 





















# *After doing Bivariate analysis we observe the following:*

1. Here, the person who is senior citizen and the person who is not senior citizer are having almost equal tenure.
2. Here we can notice the strong linear relation between the features.As the tenure increses, TotalCharges also increases rapidly. If the customers have low tenure services than there is high chance of churn.
3. There is no difference between the Male and Female senior citizen.
4. There is a linear relationship between the MonthlyCharges & TotalCharges. The customer with high monthly charges have high tendency to stop the services since they have high total charges.
5. In the first plot we can see that Male and Female both have equal churn level.

In the second plot we can see the customers without partners have high churn rate compared to the customers with partners.

In the third plot the customers who do not have any dependents have high churn rate compared to the customers who have dependents.

In the last plot we can notice the customers who have phone service have high tendency of getting churned.

1. The customers who have phone services from single line have high churn rate compared to the customers having phone services from multiple lines, also there are very less number of customers who do not have phone services.
2. The ratio of churn is high when the customers prefer Fiber optic internet services compared to other services, may be this type of service is bad and and need to be focused on. And the customers who own DSL service they have very less churn rate.
3. The customers who have no internet service have very less churn rate and the customers who do not have online security services have high tendency to getting churned.
4. It is also same as in the case of online security. It is obvious that the customers having who do not have internet services they do not need any online backup. The customers who do not have online backup services they have high churn rate.
5. The customers who do not own any Device protection have very high churn rate compared to others.
6. Here we can clearly see that the customers who do not have any techsupport then they have high churn ratio
7. The churn rate is nearly same if the customer own StreamingTV or not.
8. The customers who are existing in the company they do not own StreamingMovies in their devices. And the churn rate is low when the customer do not have internet services.
9. The customers who have churned are mostly having month to month contract.
10. The customers who prefer paperless billing they have high churn rate.
11. The customers who prefer Electronic check have high churn rate also the customers who existing in the company uses equal payment method.

# *Multivariate Analysis*

# 

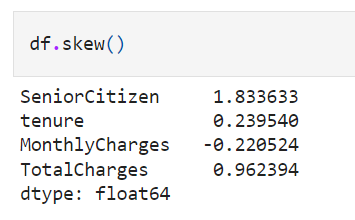
1. The pairplot gives the pairwise relation between the features on the basis of the target "Churn". On the diagonal we can notice the distribution plots.
2. The features tenure and TotalCharges, Monthlycharges and TotalCharges have strong linear relation with each other.
3. There are no outliers in any of the columns but let's plot box plot to identify the outliers.

# Checking for Outliers

# 

# The column Senior Citizen has outliers but it contains categorical data so no need to remove outliers. Apart from this none of the columns have outliers.

# Checking for skewness



The columns SeniorCitizen and TotalCharges have skewness in the data. Since SeniorCitizen is categorical no need to remove skewness. Since TotalCharges is continuous in nature, lets use cube root method to remove skewness.

# Removing skewness

# 

# We have removed the skewness using cube root method.

# 

# We can see the skewness has been reduced in TotalCharges column.

# 

# Encoding categorical columns

# 

# We have converted the categorical columns into numerical columns using Ordinal Encoding method.

# Checking the statical summary of the dataset

# 

# After encoding the categorical columns we can see all column details here. The counts of all the columns are same that means no null values in the dataset. This describe method describe the count, mean, std, min, IQR and max values of all the columns.

# Correlation between target variable and independent variables using heatmap

# 

# 

This heatmap shows the correlation matrix by visulaizing the data. We can observe the relation between feature to feature and feature to label. This heatmap contains both positive and negative correlation

1. There is no much positive correlation between the target and features.
2. The column MonthlyCharges, PaperlessBilling, SeniorCitizen and PaymentMethod have positive correlation with the Label Column "Churn".
3. The label is negatively correlated with Contract, tenure, OnlineSecurity, TechSupport, TotalCharges, DeviceProtection, OnlineBackup, Partner and Dependents.
4. Also the column gender has very correlation with the label, we can drop it if necessary.
5. The column TotalCharges and tenure, Contract and tenure, TotalCharge and MonthlyCharges and many other columns have high correlation with each other.

# Visualizing the correlation between label and features using bar plot

# 

# From the above bar plot we can notice the positive and negative correlation between the features and the target. Here the features gender and PhoneService have very less correlation with the column.

# Separating features and label

# 

# Feature Scaling using Standard Scalarization

# 

# We have scaled the data using Standard Scalarization method to overcome the issue of biasness.

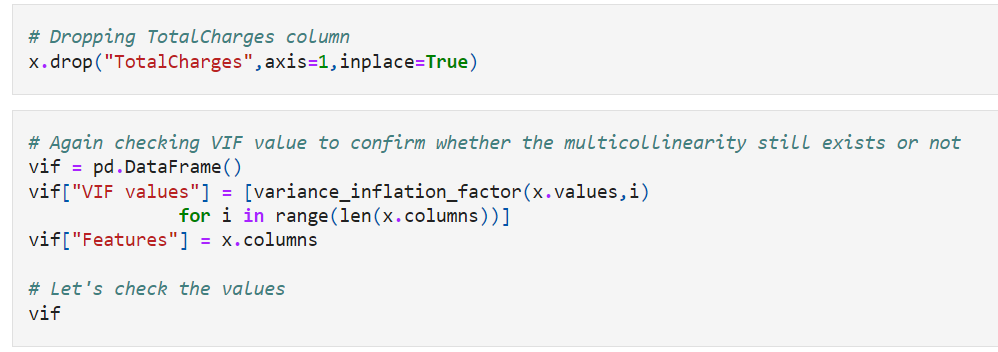
# Checking Variance Inflation Factor(VIF)

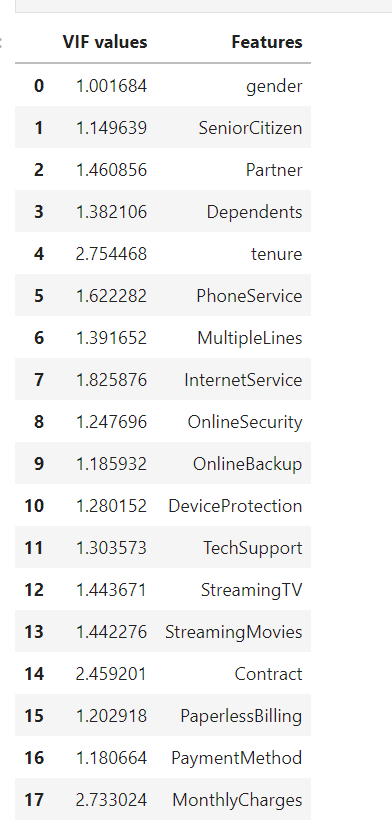
# 



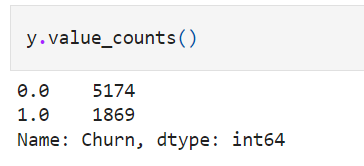
By checking VIF values we can find the features causing multicolineraity problem. Here we can find the feature TotalCharges and tenure have VIF value greater than 10 which means they have high correlation with other features. We will drop one of the column first, if the same issue exist then we will try to remove the column having high VIF.

**Dropping TotalCharges column**

****

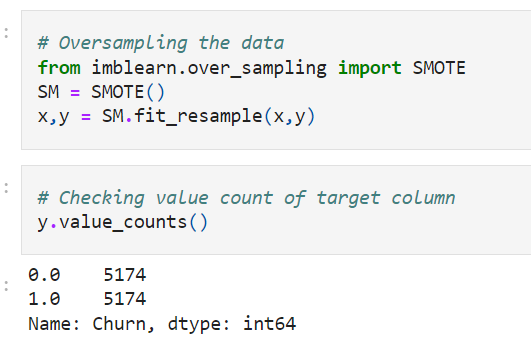


So, we have solved multicollinearity issue. We can now move ahead for model building.



Here we can observe that the data is not balanced, since it is a classification problem we will balance the data using oversampling method.

# Oversampling



Now the data is balanced. Now we can build machine learning classification models.

# Modeling

# *Finding the best random state*

# 

Best accuracy is 0.8640901771336554 at random\_state 143

# Creating train test split

# 

# Classification Algorithms

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# Plotting ROC and compare AUC for all the models used

# 

# 

# 

# Predicting the saved model

# 

# 