**Customer Churn Analysis using Machine Learning**



Here now we are going to do the customer churn analysis with the help of machine learning algorithms

**Introduction:**

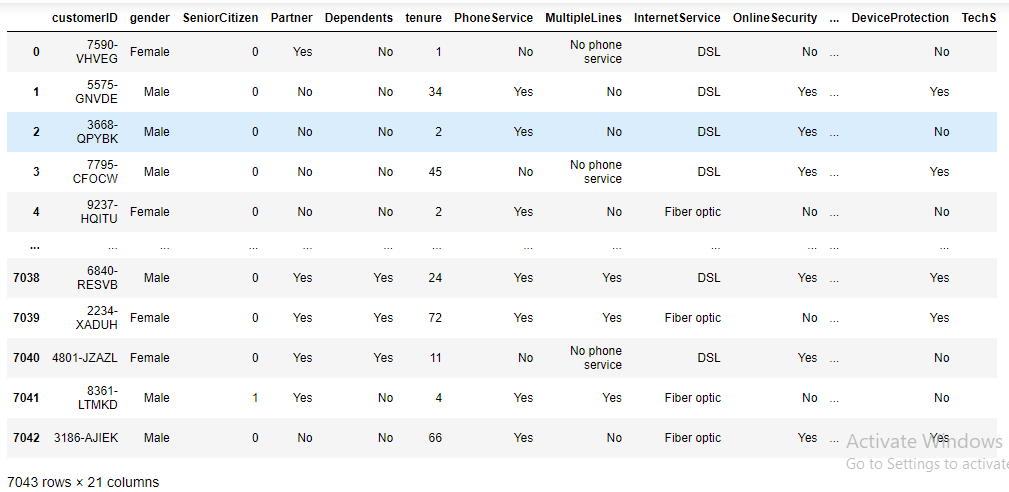
Here we will do the complete customer churn analysis with the help of the ML techniques and ML algorithms build the model and select the best model to get the best accuracy score at the end and also save the model

There are some of the steps to be followed

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

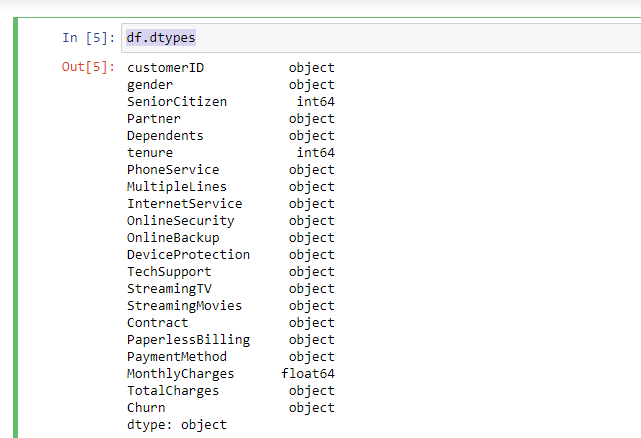
**2. Data Analysis**



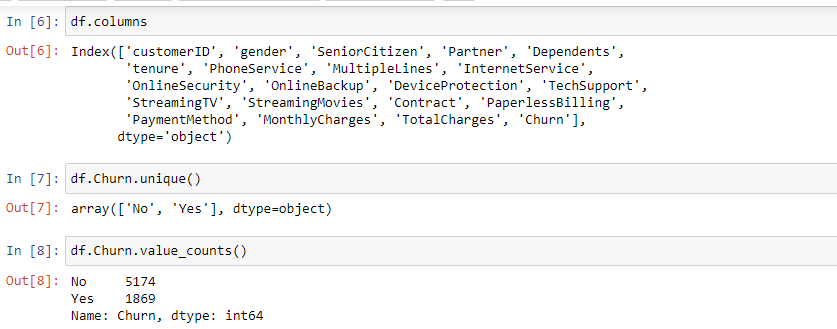
Above is the data set given here we can see that there are 7043 rows and 21 columns in the data set

**3.Exploratory Data Analysis**

We will see the types in the dataset,columns, unique values and value counts



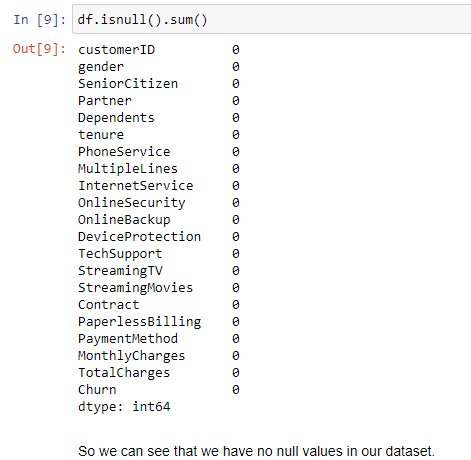
The above are the types in the data set.



The above we can see the columns unique values and value counts

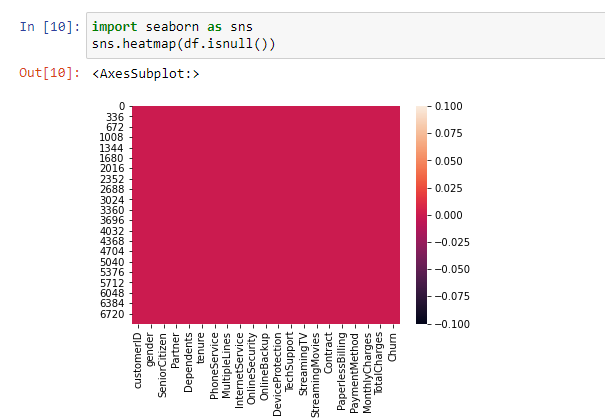
**Observation:-**By looking the dataset we can easily see that it's Logistic Regression problem and Churn is our dependent variable target.

**Checking missing values using isnull function**

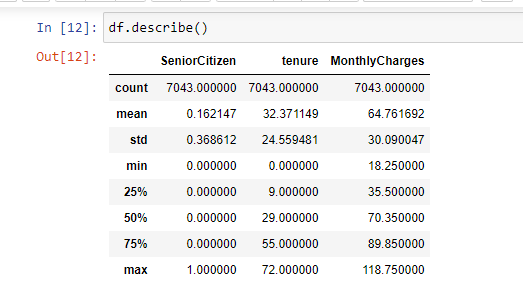
****

Here we can clearly see that there are no any null values in the given dataset.

I have also plotted an heat map for the same.



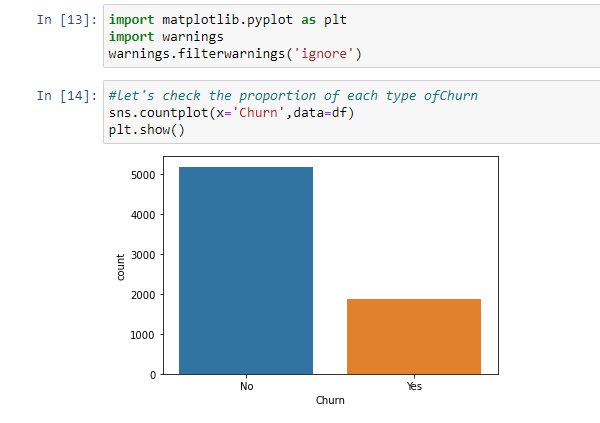
**Describing the data set**



**Key Observations**

* There is a difference in mean and median.
* There is a small difference in max and 75% percentile which shows that many outliers present in many columns
* There is a large difference in max and 75% percentile in monthly charges column which shows that outlier are present in it.

**Now we will check the imbalance in the Dataset**

****

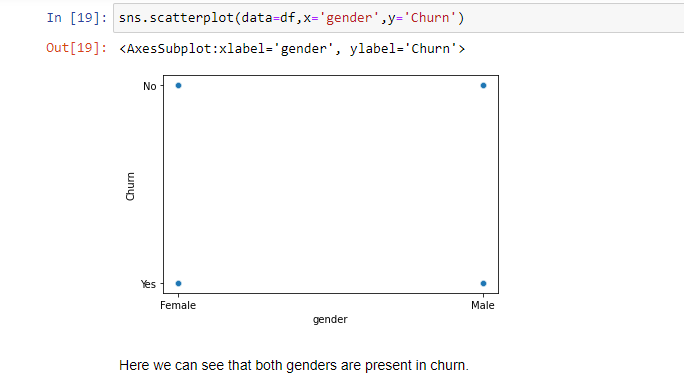
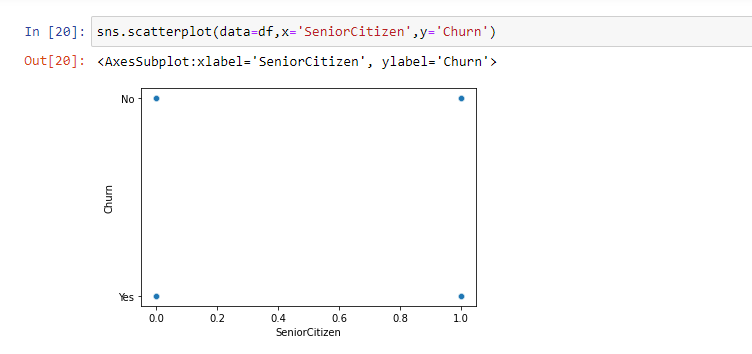
**Here we can see that there is imbalance in the data set**

**3. EDA concluding remarks**

We have checked the missing value in the dataset already

We will do the Bivariate Anaylis hence we will do it with scatterplot

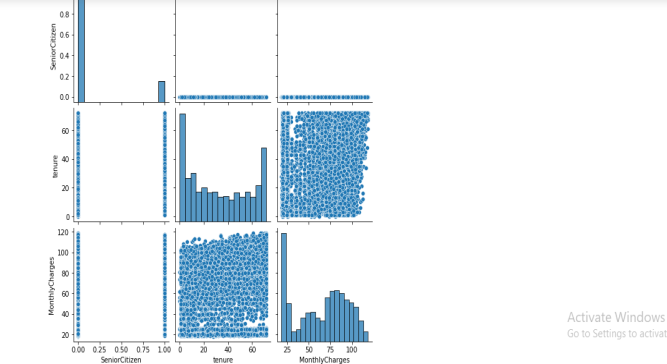
this will also show the relation of each column with the target column



Thus I have plotted and observed relation of each column with the target column.

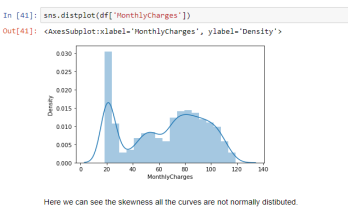
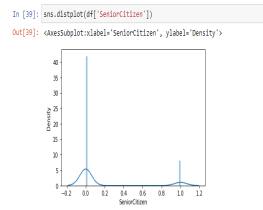
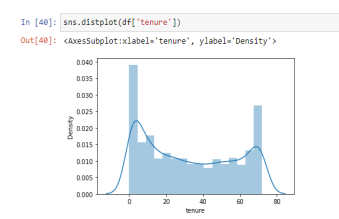
**#Multivariate analysis**

**Here we are not able to find the co-relation as the target column is in the string format**

****

The above plot shows the relation with each column.

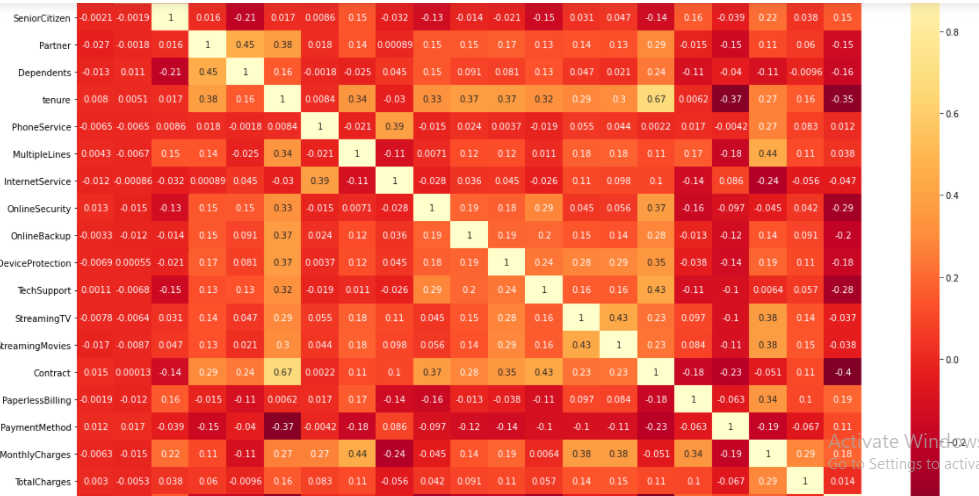
**#Checking the skewness**

****

We can clearly see that in tenure and senior citizen column the skewness is equally distributed but in monthly charges it is not equally distributed

Now checking the co-relation

I have plotted the heat map to check the co-relation



There are certain key observations done after plotting of the the

heatmap

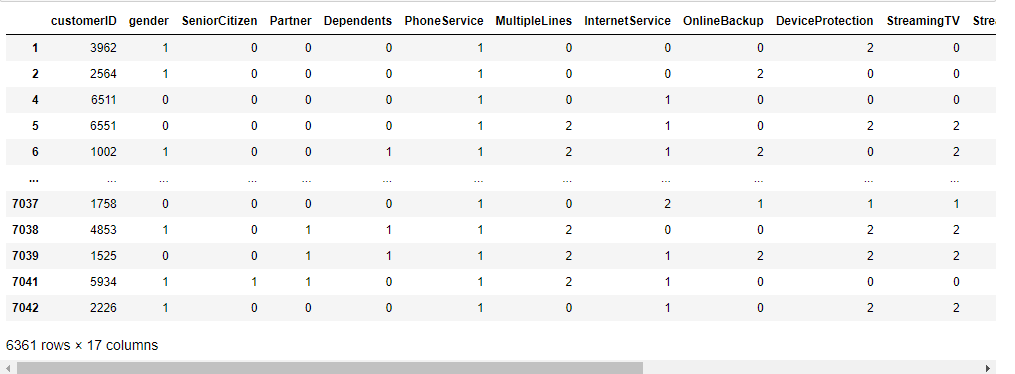
1 Light shades are positively correlated.

2 Here we can see that dark columns liketenure,contract, tech support and online security are highly negatively correlated.

3 Other columns are slightly positively correlated like senior citizen ,mutiplelines,streaming movies total charges.

4 Some columns are positively correlated like paperless billing and monthly charges.

Dropped the data which was not necessary

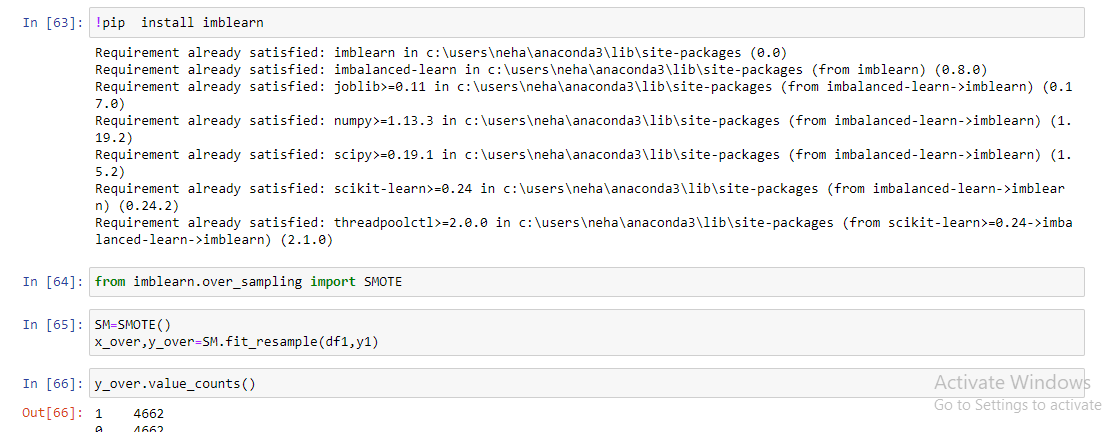
****

Now we will seprate the the indipendant variable and target variable

then importing the power transform from sklearn preprocessing.

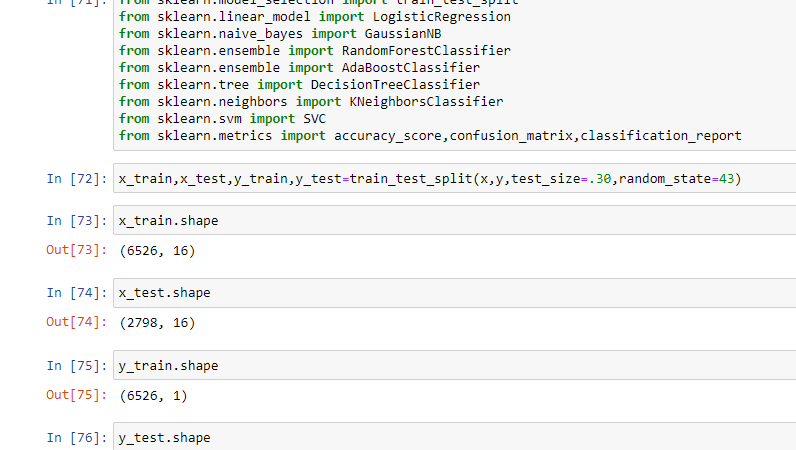
Then we will handle the class imbalance problem by using oversampling the minority class

First install imblearn then from imblearn .oversampling import smote below I am sharing the pic of it

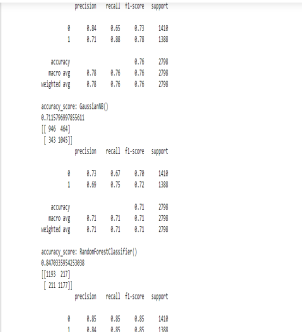
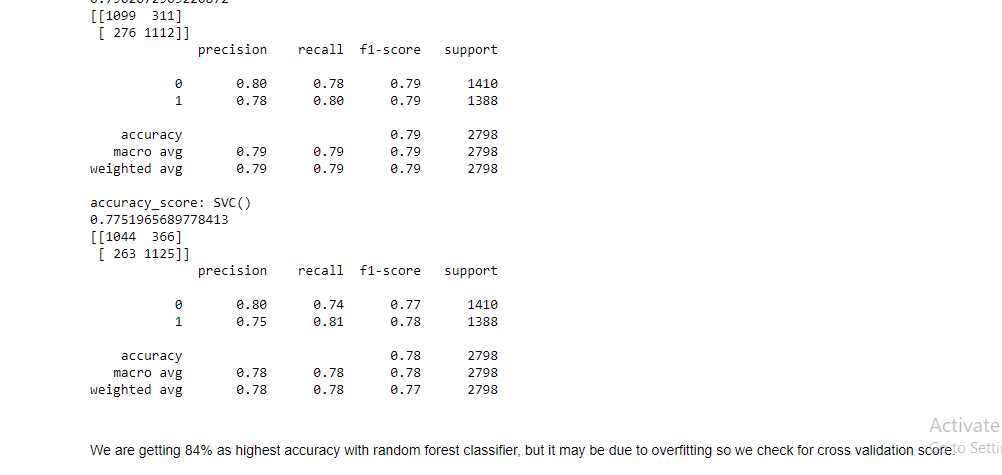
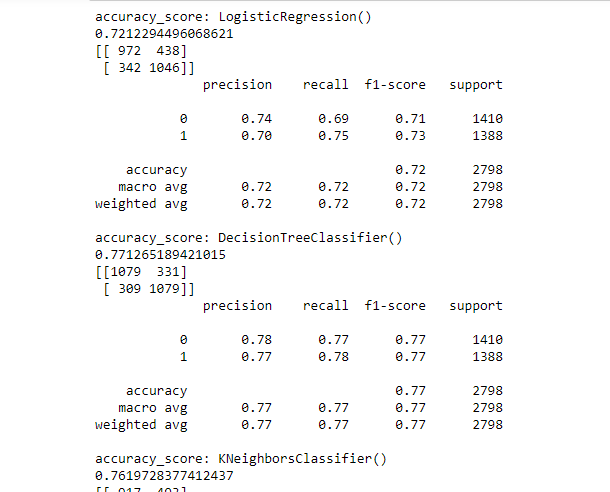


Now creating train\_test\_split

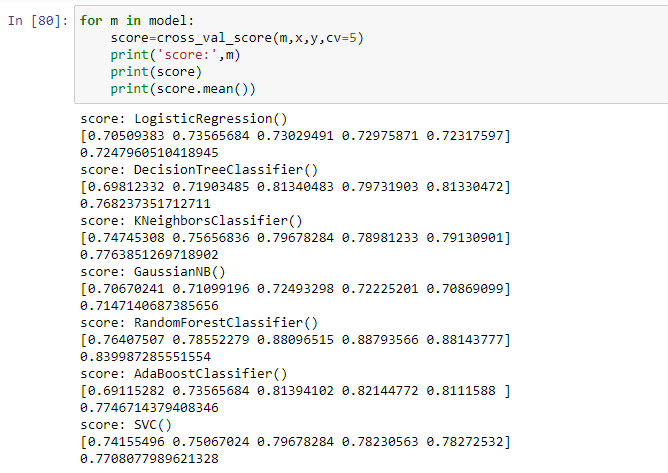
Importing all the necessary library whichever is required



Now we will finally see the best model using the algorithm using multiple models

Here we are getting the score of 84% with random forest classifier Hence now we will go for cross validation now



The above we can see that we have done the cross validation and it has the output above as we can see that random forest classifier gives us the more accuracy score Hence now we can good to go for hyper parameter tuning



The above pic is of we have done hyper parameter tuning and implemented the parameters in our model.

and for hyper parameter tuning we have imported the necessary library

**from sklearn.model\_selection import GridSearchCV**

**#Conclusion remarks**

Here it is clearly seen that after testing on the multiple models we can see that we are getting the highest accuracy score with RFC.

Even after cross\_validation we are getting the maximum accuracy score with RFC(Random Forest classifier)

Hence this will be the best model to be used and implemented