**Customer Churning Prediction**

In this article, I will walk through the entire journey of building a machine learning model that will help a business to decide whether it is able to retain its customer or there is a churn. There will be certain features which will help to decide the factors which help in customer churning like- gender, tech support, monthly charges, broadband availability as well as various features used by a certain customer. The data collected and observed through these features will help us to get our desired output.

**What is Customer Churning:**

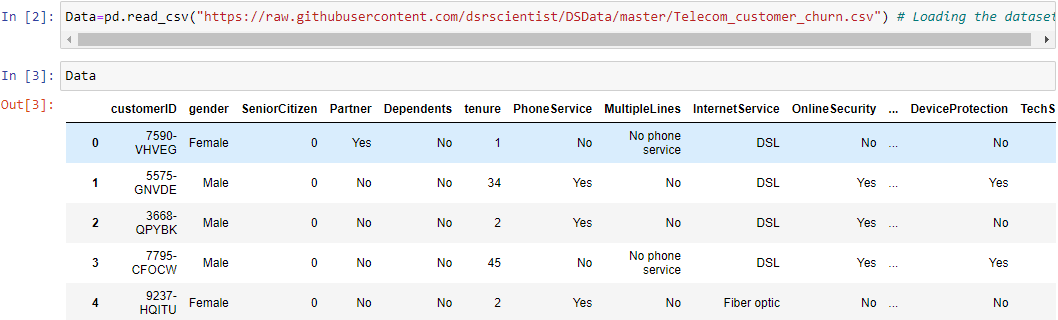
Customer churn is the total number of customers who stop using a company’s service/facilities due to various reasons in a certain period of time. The company fails to retain these customers and as a result it affects a company’s profit and goodwill. Losing old customers for a period of time not only harms a company’s brand image but also gaining new customers is more expensive than upgrading an old customer. Therefore companies should adopt all possible measures and invest in every opportunity areas which are being responsible for losing a customer over a period of time.

The following model is built keeping in mind to recognize and improve the areas which are leading to customer churning in a telecommunication sector and will predict whether a customer will be churned or retained.

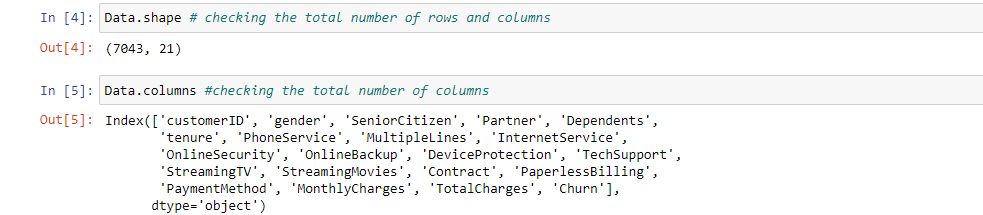
**Importing Libraries**



**Loading the data:**

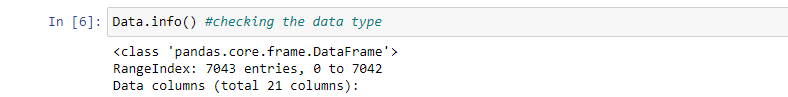


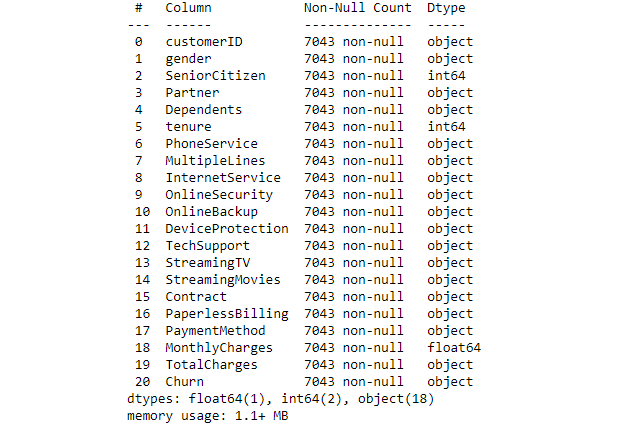
**EDA/Data Analysis:**



The above data gave us 7041 unique rows and 21 columns along with the various column names

Let us check the various data types used in the dataset:





Therefore out of 20 columns we have 2 integer datatype, 1 float and the rest 17 as object datatype ( which will be encoded later on as machine does not understand object data type)

Lets now check if there are any null values in the data:



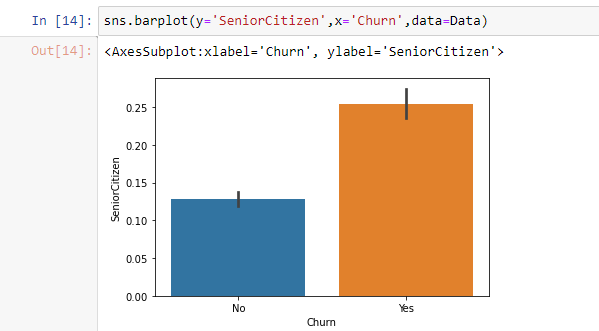
Data has no null values.

We can see customer ID is an unique number which helps as a customer identification feature and hence will not impact in determining whether a customer will be churned or not. Therefore let us drop the column:

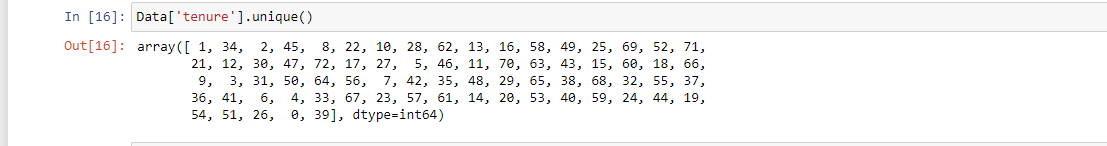
**EDA and Visualization of various features/columns**

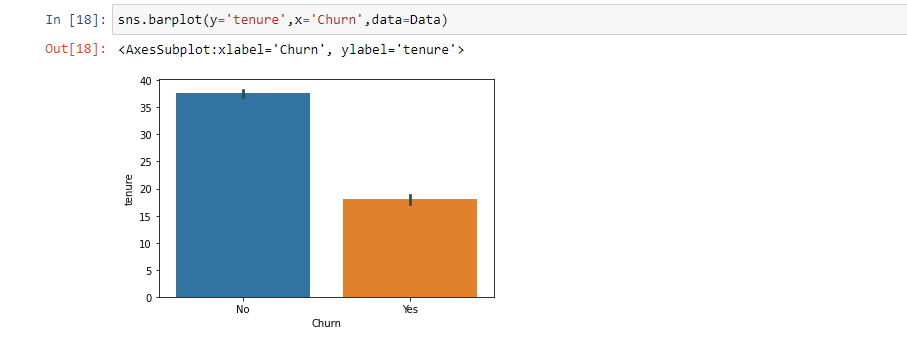


0 stands as young customers and 1 refers to senior citizen. Looking at the above data we can be sure that majority of the customers are non-senior citizens

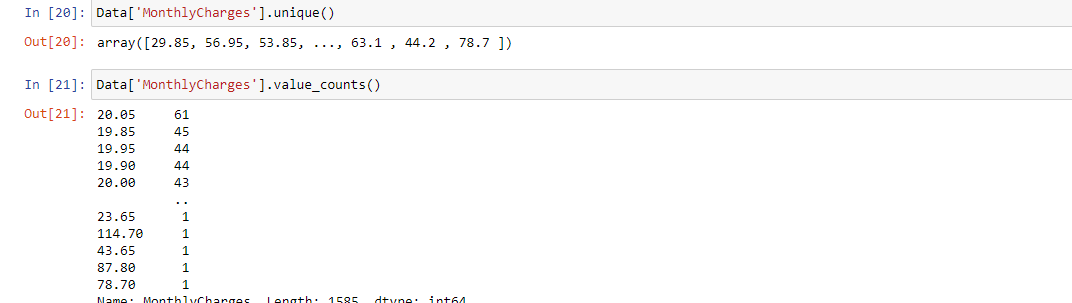


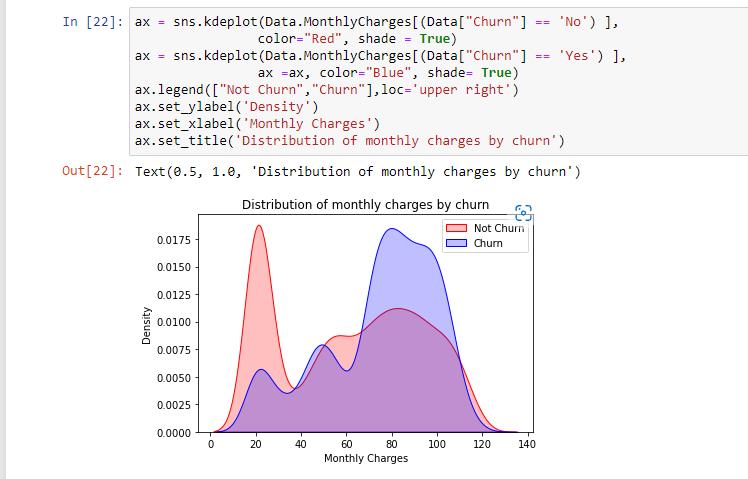
We can also see that majority of the churning takes place from senior citizens than the young customers.



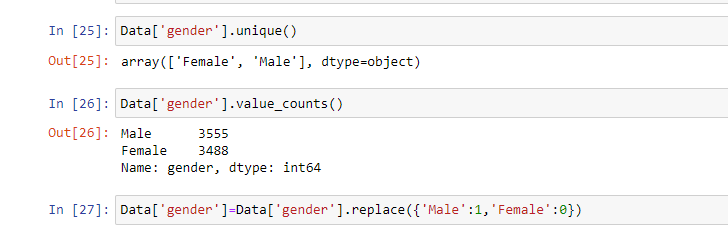


The above numbers in the unique value of ‘tenure’ are in months. The highest month data available is when a customer has been retained till 72 months. However looking at the above pattern it can be said churning occurs majorly when a customer is in their 12-15 months ( 1 to 1.5 years)



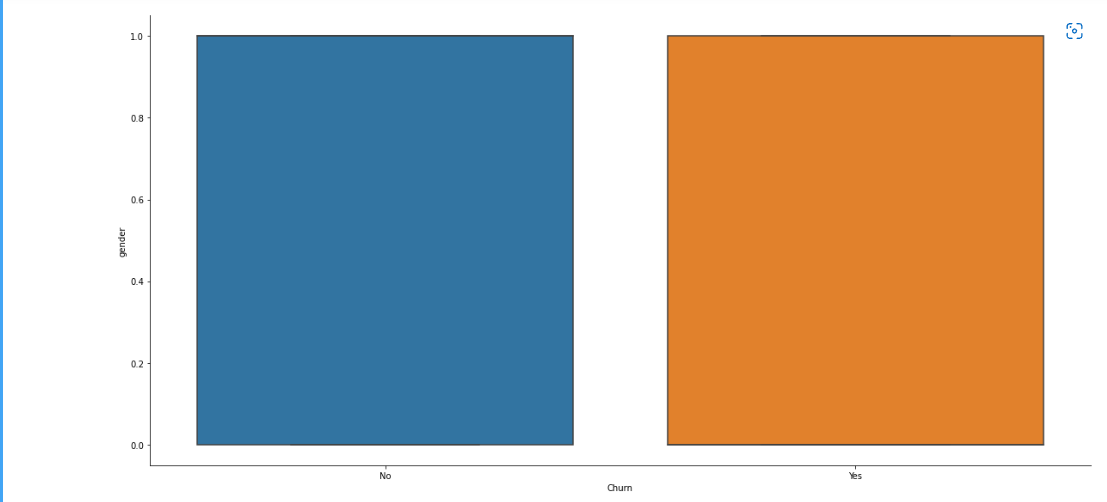


Next we have is the monthly charge rates. Interestingly customers paying lowest monthly charge has the least possibility to be churned than customers paying higher monthly rates specially between ( 60-100)

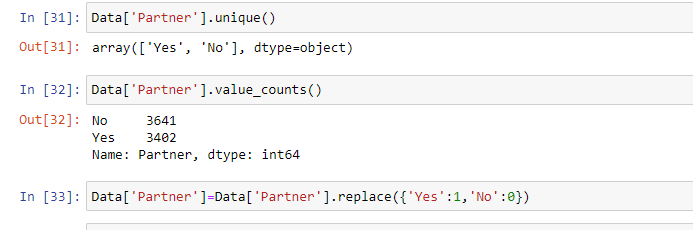


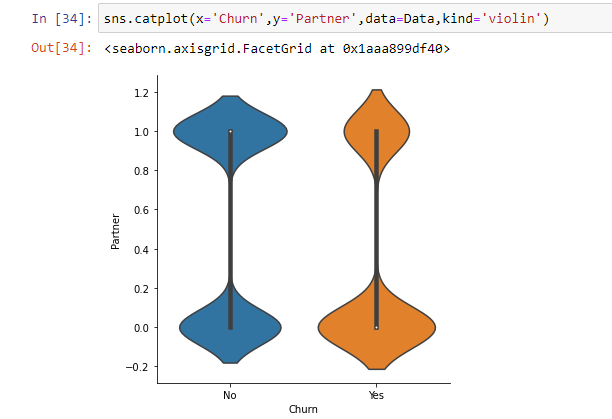
Checking the gender of the customers and it can be said that data is equally split between male and female. We have also replaced the Male value as 1 and Female as 0 here.





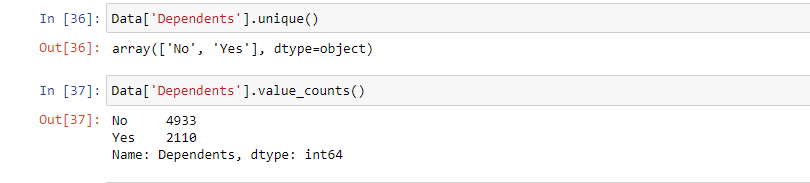
Let us now check how churning affects with people having partners:

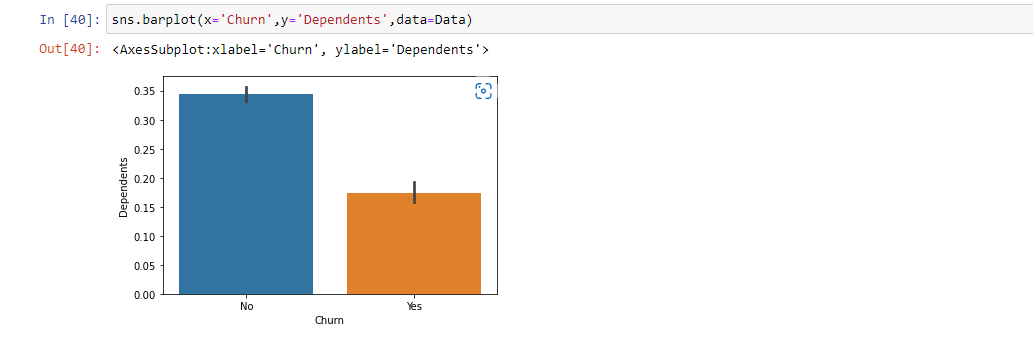




There is not much difference when it comes to a customer who has a partner or is single.

Let us now see that how churning is affected by customers who are dependent on someone for the service as well as non-dependents.

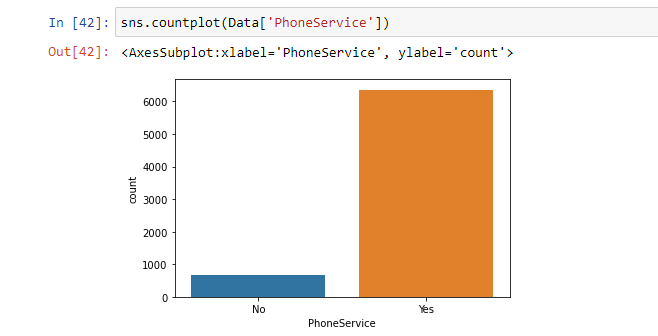




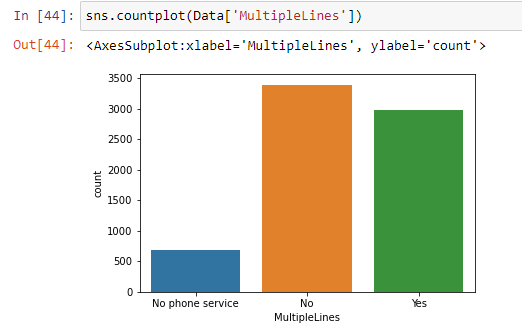
Though the data for dependent customers is much less than the data for non dependent customers. However there is a majority churning of those customers who are usually dependent on someone rather than non dependent.

Let us now understand few of the features that are available to the customers

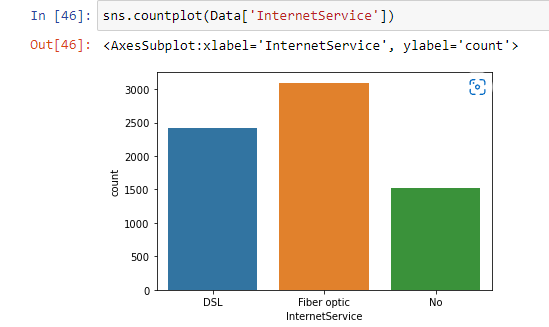
**Phone Service**



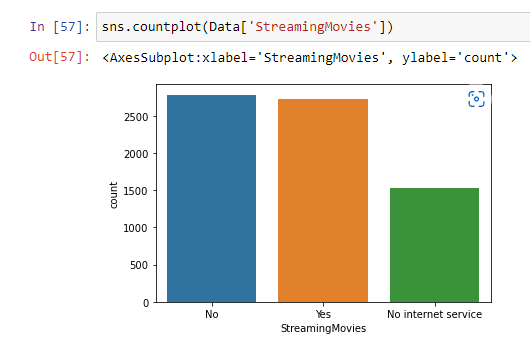
**Multiple Lines:**

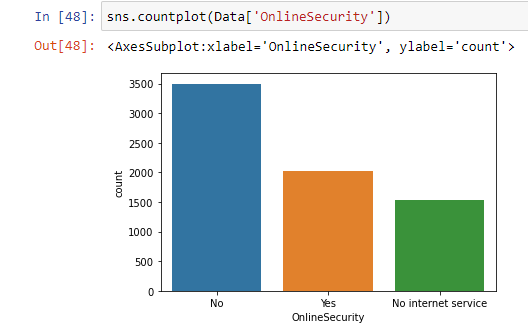


**Internet Service:**

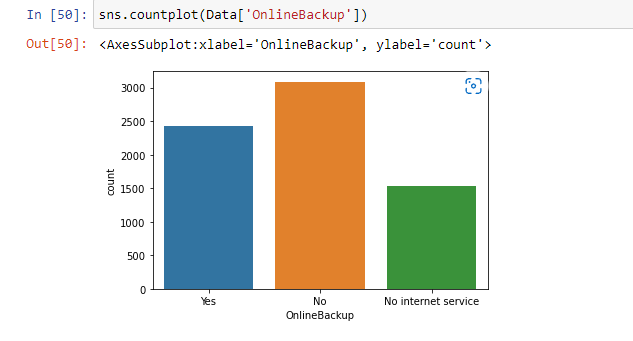


**Streaming Movies**

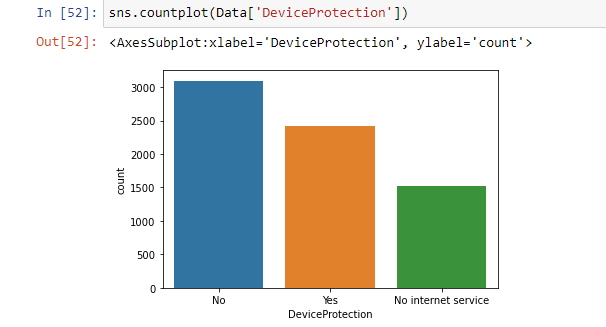




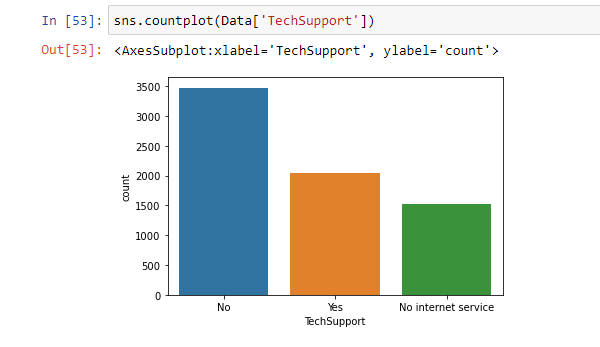
**Online Backup:**



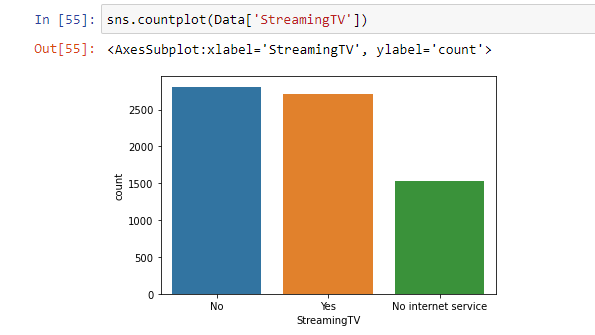
**Device Protection:**



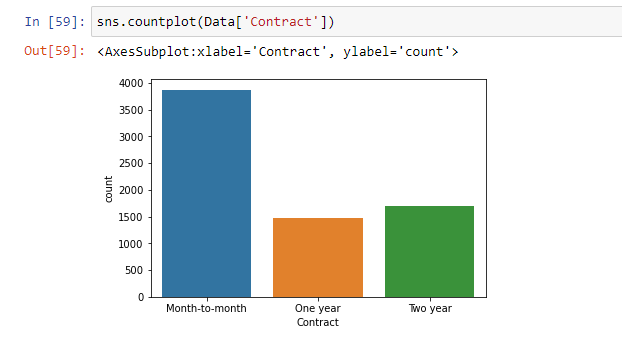
**Tech Support:**



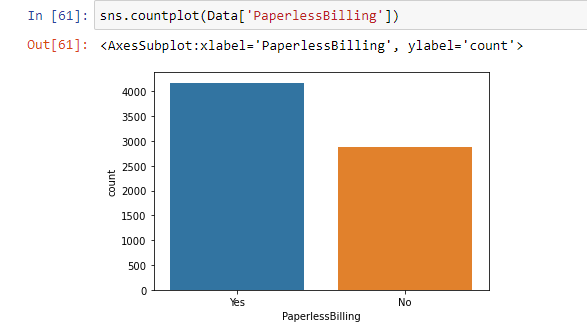
**Streaming TV:**



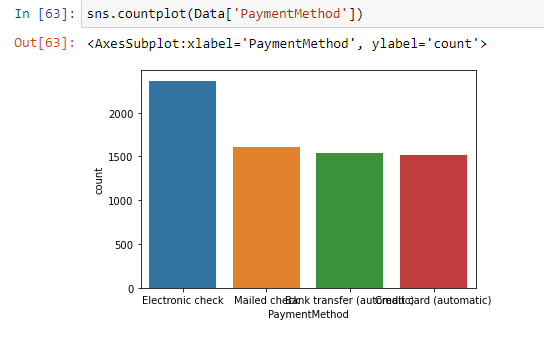
**Contract:**



**Paperless Billing:**



**Payment Method:**



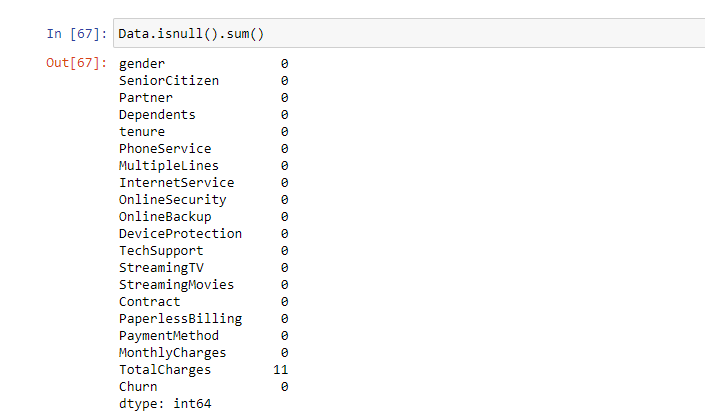
**Observation:**

Majority of the customers use a single telephone line with fibre optic internet connection which has streaming movies and TVs. Majority of the customers opts for month to month payment contract which has less charge with a low rate of churning with paperless billing and electronic payment facilities.

**However the disadvantage here is majority of the customers does not have a device protection, online backup, tech support and online security. These issues will contribute a greater extent in customer churning. Hence companies should adopt more methods or plans to come up with a solution for the above issues.**

Total charges should be an integer or a float value but it shows as an object data type, hence let us convert that and check if there are any null values:





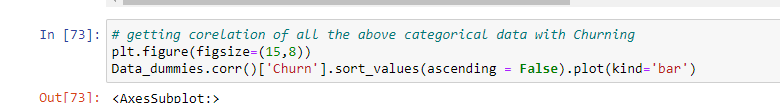
We have 11 null values which can be dropped off:

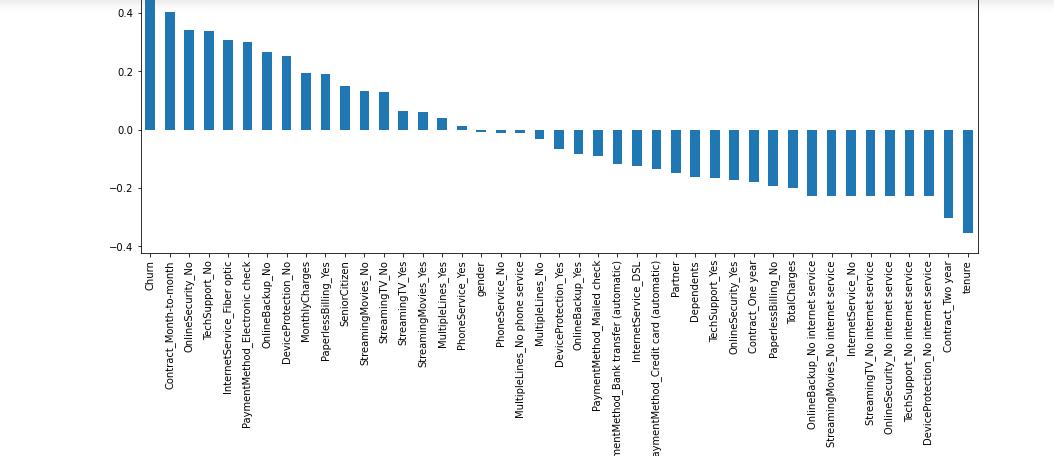


Let us now encode our target variable and then find the dummy values for object data type which we can use as a correlation with the target variable to understand which feature has a positive or a negative relation with Churning.



Let us get the correlation:





**Observations:**

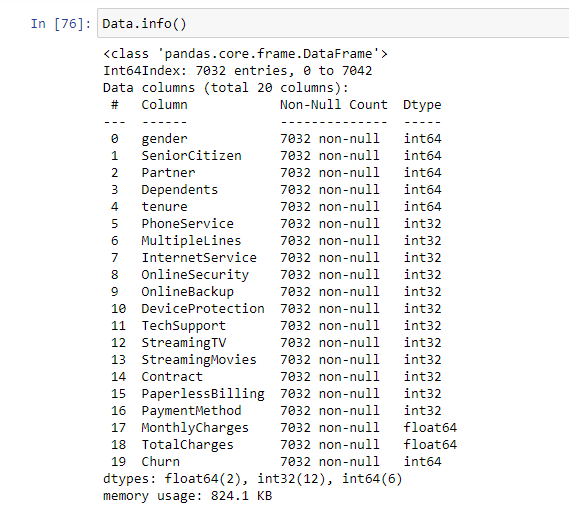
It is observed that- churning rate will be negatively affected with people having two year contract( as seen above churning majorly happens within 1.5 year of contract), and no internet service, whether they are dependent or if they have a partner.

Churning will be positively affected with month to month contract ( as people have opted for that plan more), if there is no online security or tech support and interestingly optic internet fibre connection ( which is the fastest in internet speed) as well as no online backup or device protection. To some extent it is also affected by senior citizen ( as majority churning has been observed for senior citizen)

**Encoding the object datatype:**



**Checking the datatype:**



All object data has now been transformed.

However while converting the target data type we found that the class is imbalanced. That is the rate of churning was much less than the customers who have not been churned. This can make the model biased. Therefore data has been upsampled.

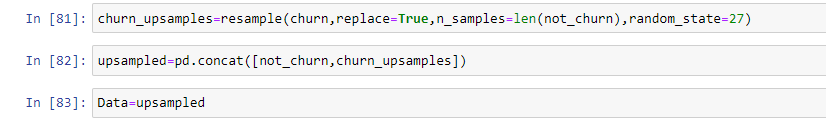
Importing the libraries:



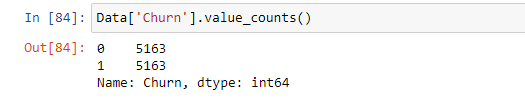
Initializing the Churn and Not Churned data:



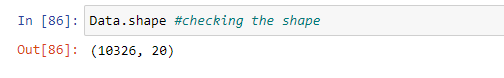
Now upsampling:



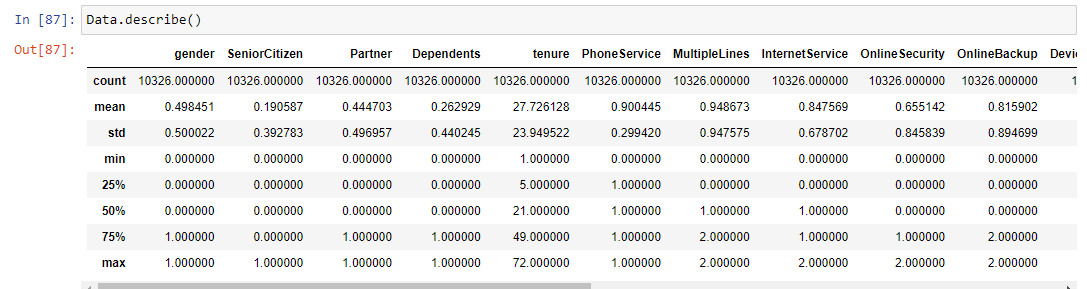
Checking the new values for churn:



Class is now balanced.



Checking the new shape as after upsampling we have 10,326 rows.



Checking data statistics

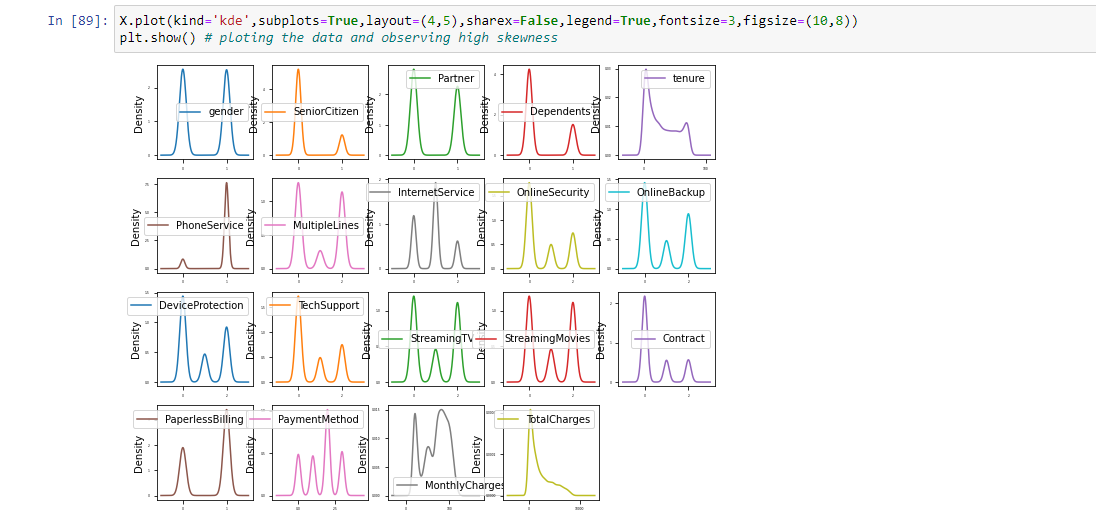
**Seperating Target and Feature Variable:**

Now finally separating the data as target variable and feature variable as X and Y:

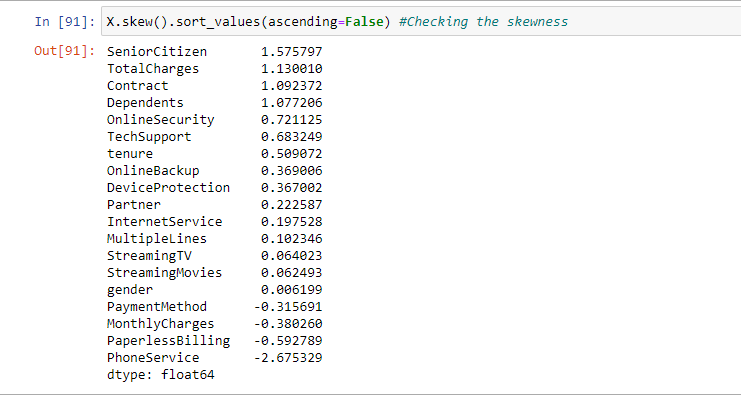


**Data Skewness**

Let us check if the data is skewed :

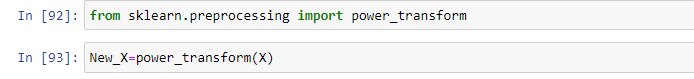


Lets check the data skewness for every column:

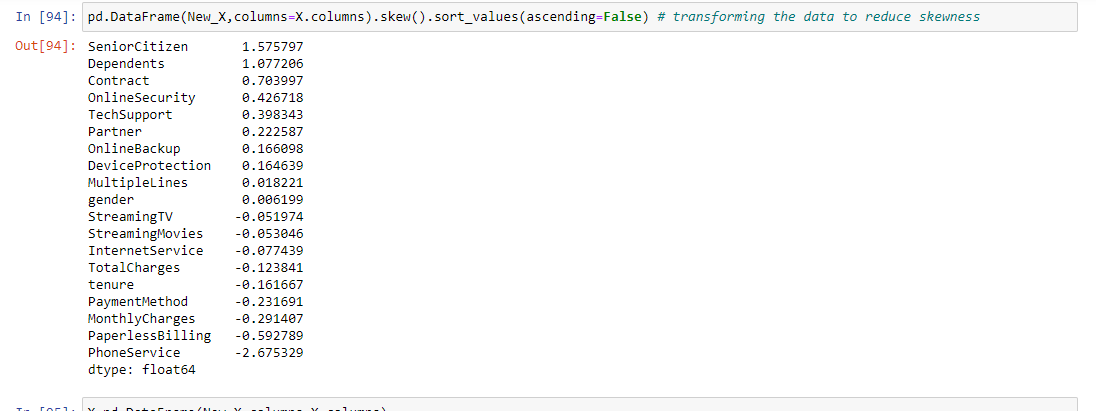


Senior Citizen, Total Charges, Contract, Dependents and Phone service is highly skewed.

As data needs to be normally distributed, hence will try to lessen the skewness by importing power transformation.



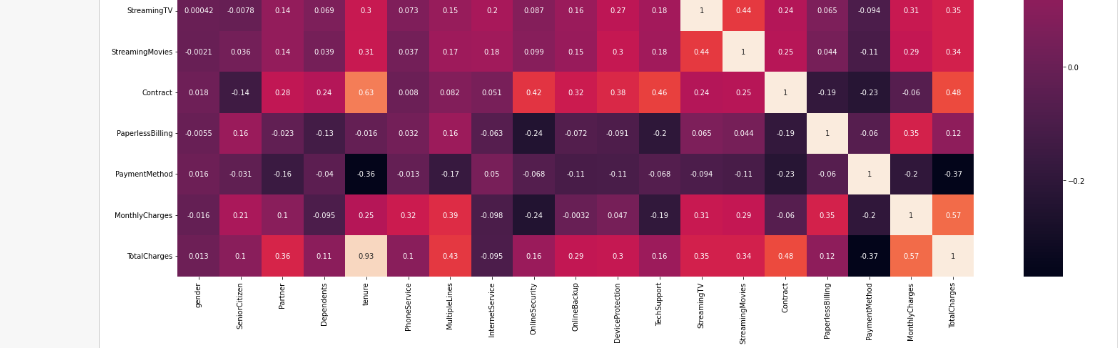
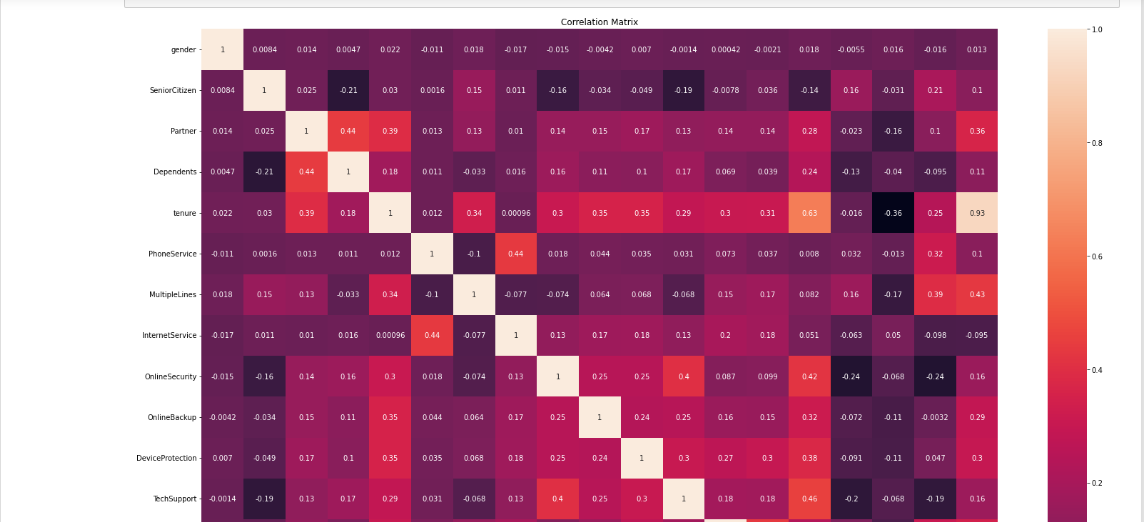
After performing power transformation, checking the skewness:



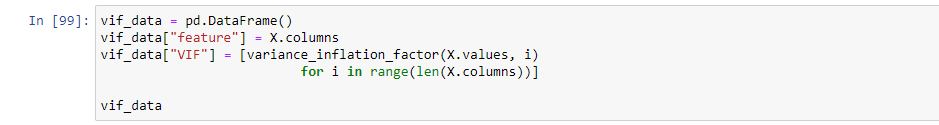
**Checking Multicollinearity Issue**

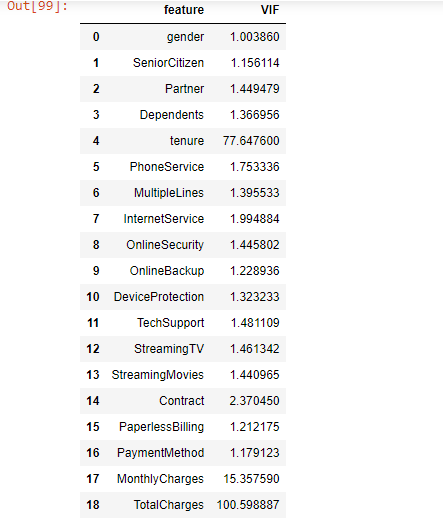
Now checking if there is an issue of multicollinearity between the target variable 1) By Heat Map and 2) VIF.

Heat Map:

From the heat map it can be said- Tenurity is corelated with Total Charges, therefore checking with VIF features

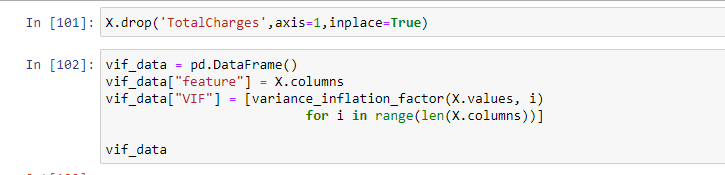
VIF:

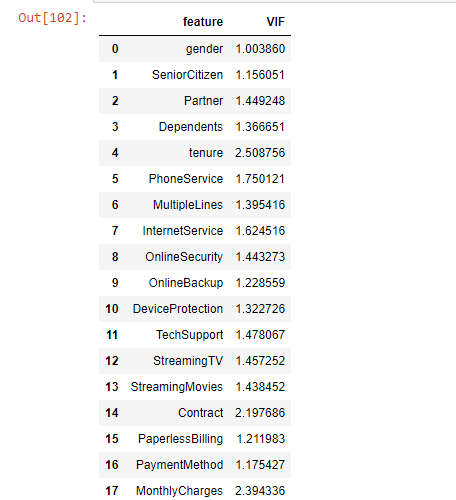




As per the conclusions drawn above, increase in monthly charge results in an increase in total charge. Churn is more impacted with an increase in monthly charge than an increase in total charge. The VIF for total charges is the highest.

Therefore based on the above conclusions, Total Charges column is dropped off.



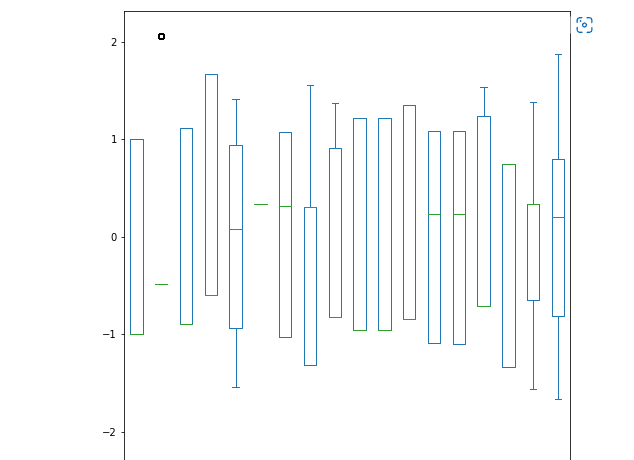


VIF has now been controlled and there is no issue of multicollinearity.

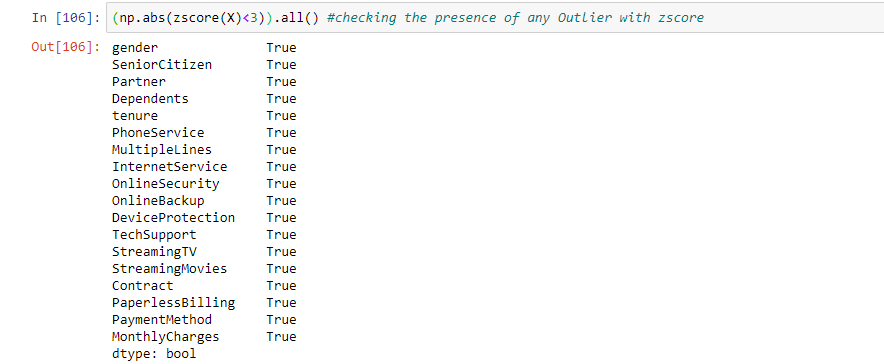
**Checking Outliers**

Now lets check if there is any Outliers present in the data.





By importing z score lets check if the score of all feature variable is within 3. If data is above 3, Outlier will be identified.



This shows there are no Outliers are all data is within zscore 3.

**Scaling the data:**

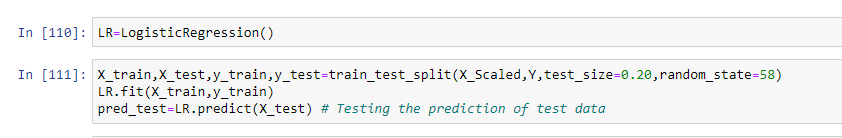


After dropping all the null values, encoding data, fixing the skewness, removing multicollinearity, checking for Outliers and scaling the data. We will now train various models to check for the accuracy score.

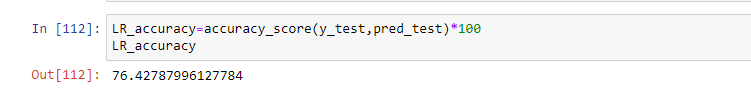
**Model 1: Logistic Regression:**

*Logistic Regression is a supervised machine learning which helps to predict binary outcome, i.e yes/no or 0 or 1. It is performed on categorical dataset and helps to resolve classification problems.*

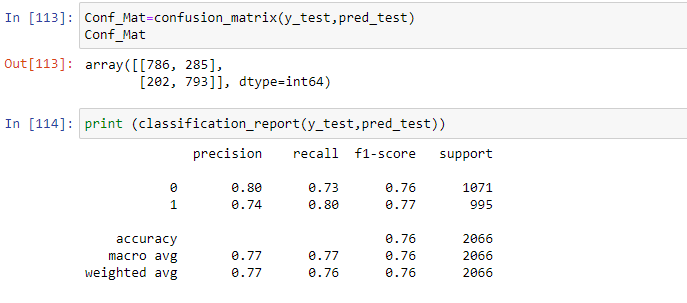
**Training the data:**



**Checking accuracy:**



**Checking confusion matrix and Classification Report:**

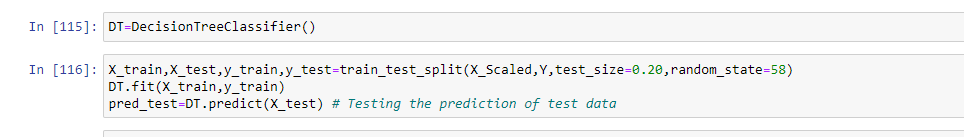


**Model had 786 values as true positives, 285 as false negative(Type 1 error), 202 as false negative(Type 2 error), 793 as true negative.**

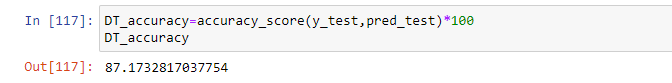
**Model 2: Decision Tree:**

*Decision Tree is a supervised machine learning technique which performs an action based on certain condition and has the graphical representation of a tree. The root node is the main factor on which decision is taken and the root node is further extended into leaf nodes. The nodes are decided based on parameters like Ginny indexing and entropy.*

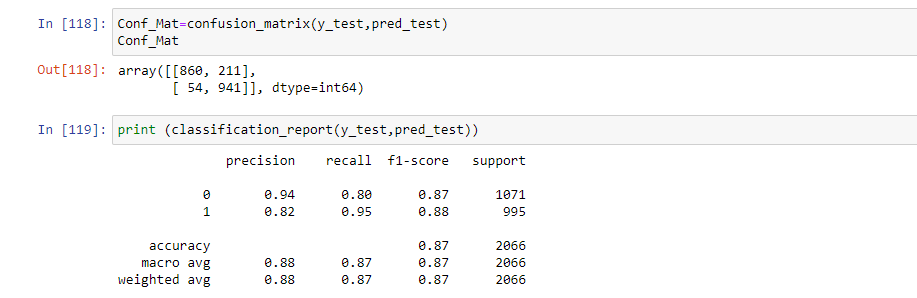
**Training the data:**



**Checking accuracy:**



**Checking confusion matrix and Classification Report:**

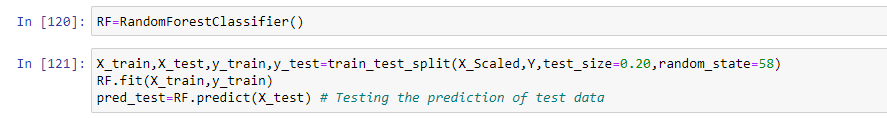


**Model had 860 values as true positives, 211 as false negative(Type 1 error), 54 as false negative(Type 2 error), 941 as true negative.**

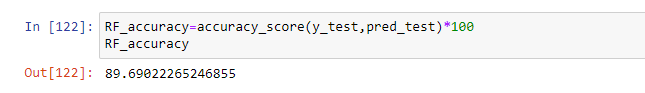
**Model 3: Random Forest Classifier:**

*Random Forest Classifier is an ensemble machine learning technique. It creates multiple decision trees and takes the majority vote. It uses random feature selection and is more accurate than an individual tree*

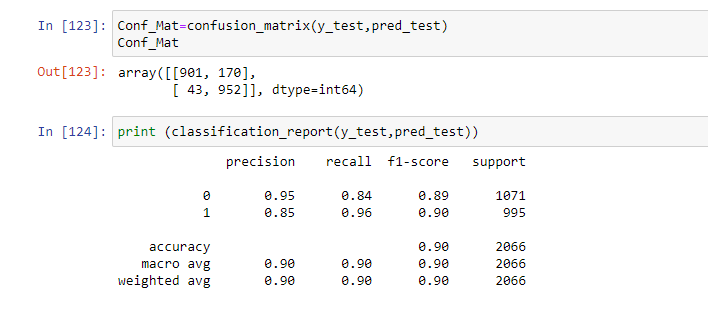
**Training the data:**



**Checking accuracy:**



**Checking confusion matrix and Classification Report:**

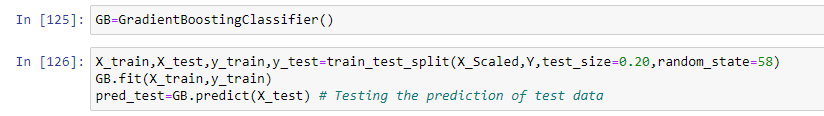


**Model had 901 values as true positives, 170 as false negative(Type 1 error), 43 as false negative(Type 2 error), 952 as true negative.**

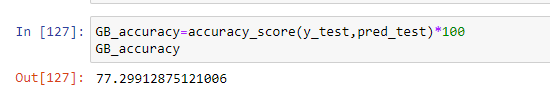
**Model 4: Gradient Boosting:**

*Gradient Boosting combines various week machine learning models and creates a strong predictive model. It reduces overfitting of the model by using regularization technique and it can be used on regression and classification dataset.*

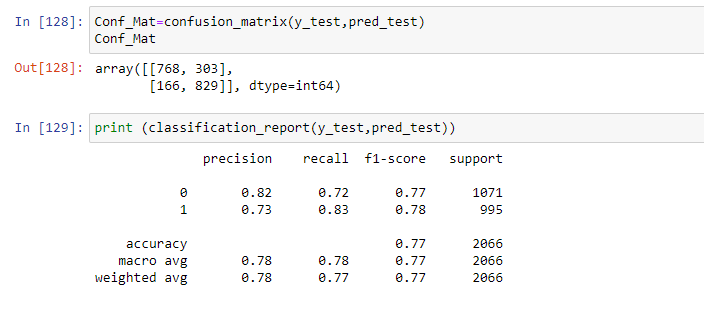
**Training the dataset:**



**Checking accuracy:**



**Checking confusion matrix and Classification Report:**

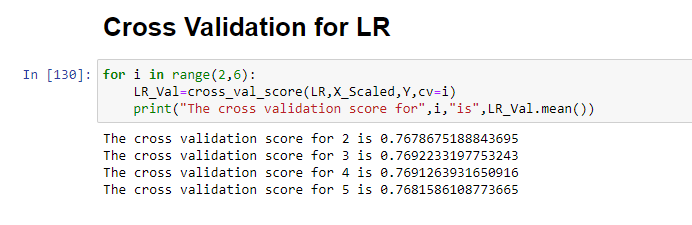


**Model had 768 values as true positives, 303 as false negative(Type 1 error), 166 as false negative(Type 2 error), 829 as true negative.**

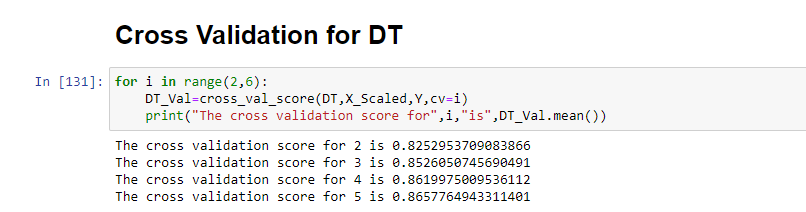
**Therefore based on just accuracy score, Random Forest Classifier is the appropriate model for the dataset as it has the highest accuracy score.**

**However we now need to check if the model is overfitted, hence we will do some cross validation for every model.**

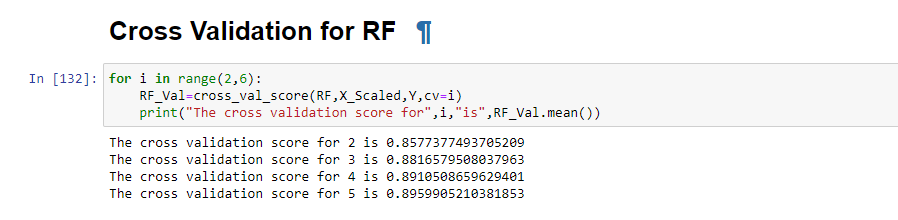
**Cross Validation for Logistic Regression:**



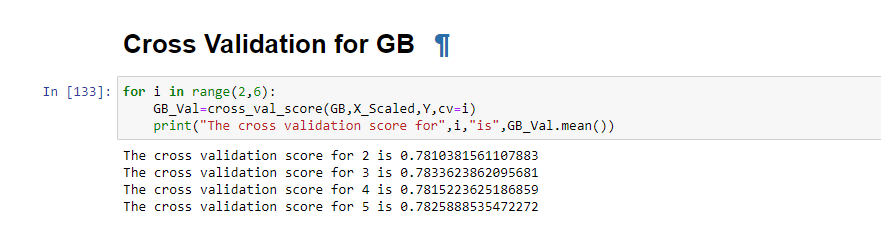
**Cross Validation for Decision Tree:**



**Cross Validation for Random Forest Classifier:**



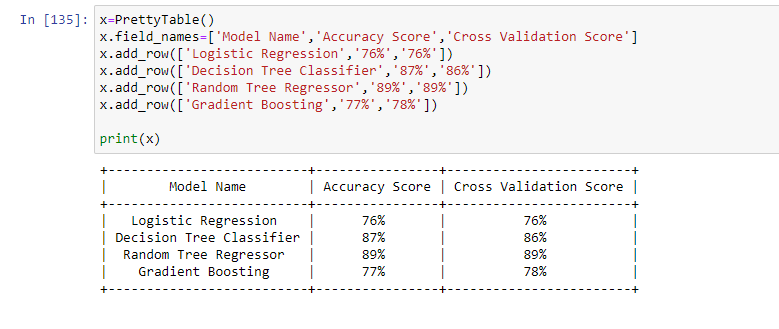
**Cross Validation for Gradient Boosting:**



**From the cross validation score, it can be seen none of the models are overfitted as Logistic Regression had: 76% Decision Tree had 86% Random Forest Classifier had 89% and Gradient Boosting had 78%.**

**Based on cross validation score as well we have Random Forest Classifier has the best model for the dataset.**

**Let us understand this better with a view from Pretty Table:**

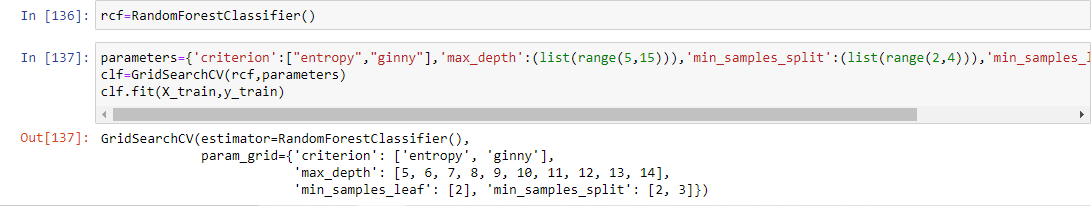


**Therefore, based on accuracy value, confusion matrix and cross validation score- Random Forest Classifier is the appropriate and chosen model for this dataset.**

**Let us try further by doing some Hyper Tuning Parameter to the best fitted model.**

**Hypertuning Parameter**

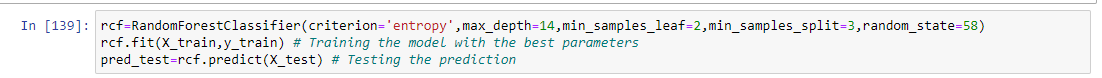
Let us again initialise the Random Forest Classifier and give some parameters to it. We will train the model based on the newly provided parameters and with the help of Grid Search CV we will get the best parameter for the model:



Now that model is trained, we will get the best parameter by using best\_params\_



Therefore, as per the provided parameters this is the best estimators as per Grid Search CV, let us see if it increases the accuracy score further in Random Forest Classifier:

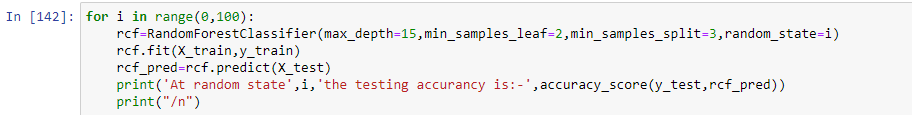


We will first train the model with the best parameters

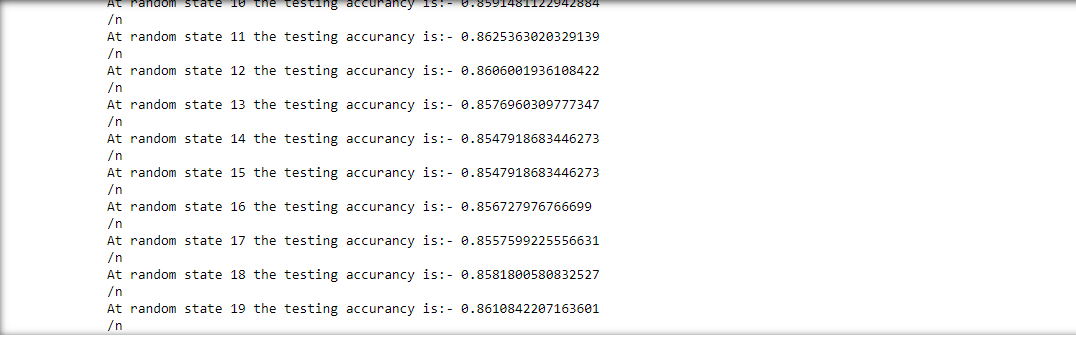
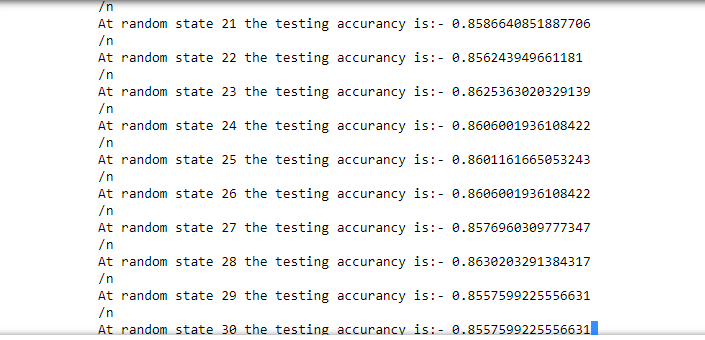


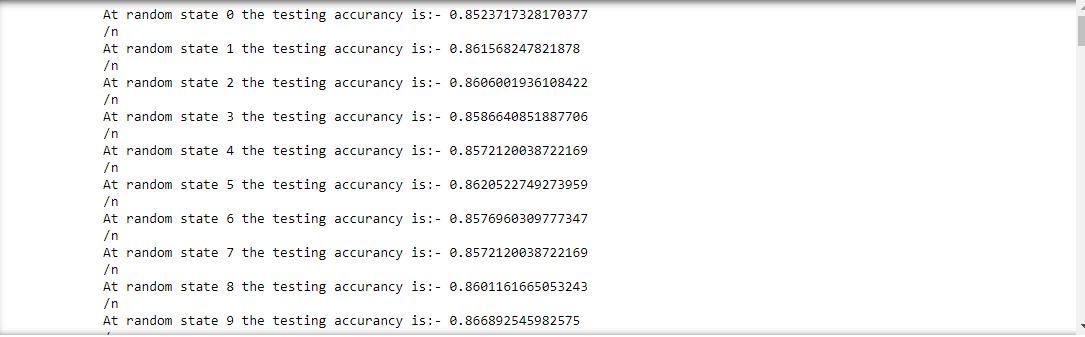
*We can see that accuracy got reduced.*

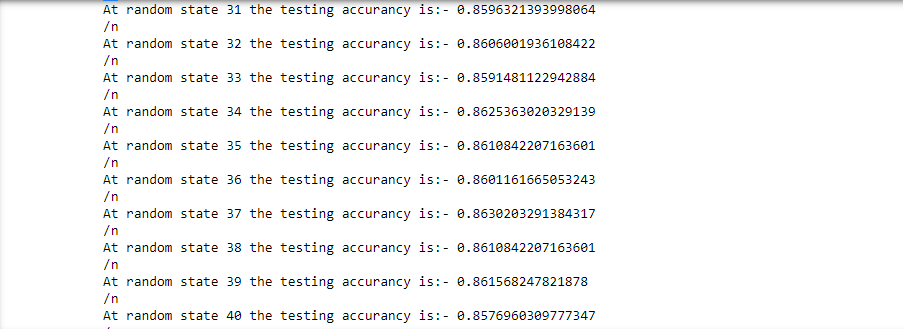
*Therefore, we can try with some more parameters to check if the accuracy score increases:*

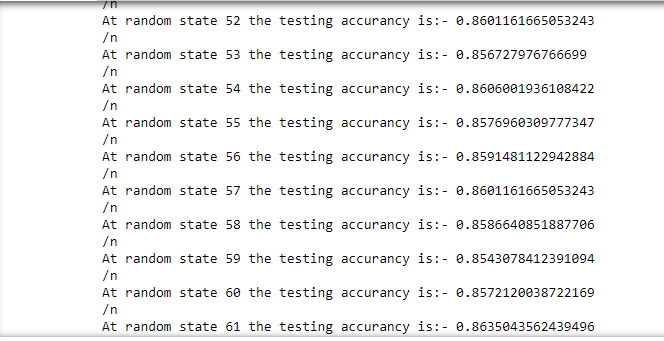
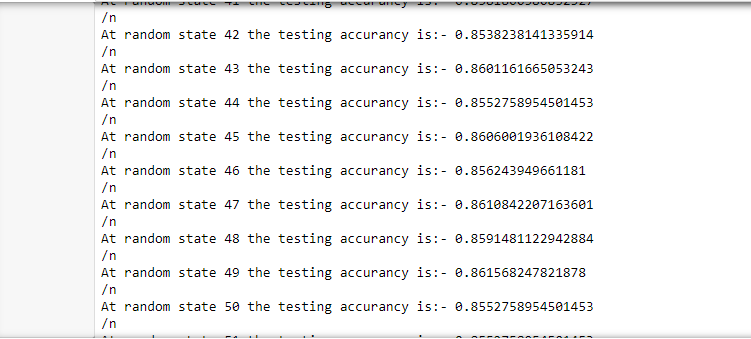


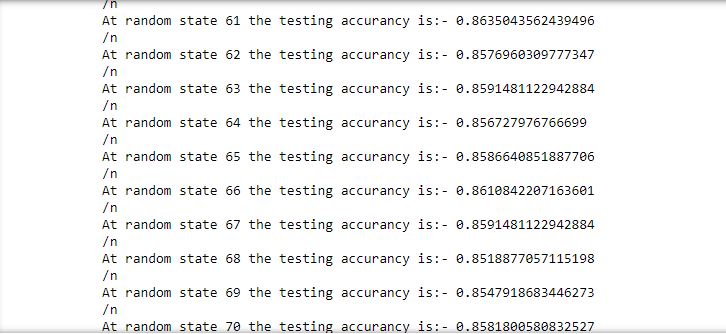
Here we have taken a different parameter. Ginni estimator is used instead of entropy and the depth has been increased to 15. Minimum sample split and leaf has been kept same. However, the random state has been taken from 0 to 100 using a for loop.

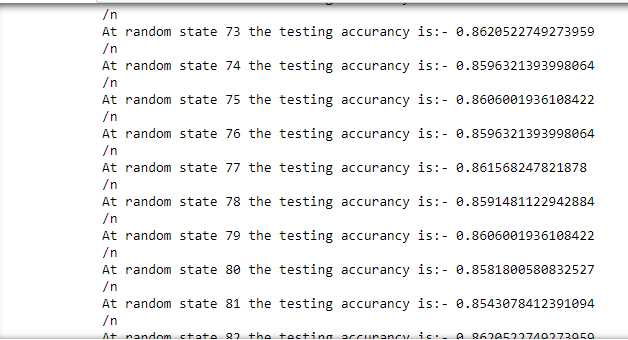


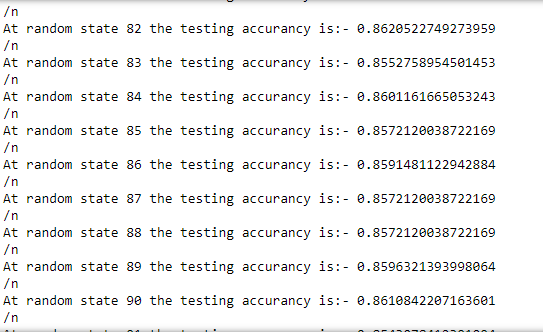


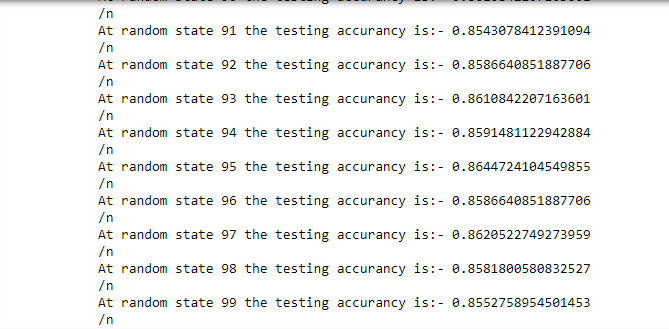












*Apparently it shows even after hyper tuning the accuracy score did not increase. Therefore, it will be better to go back to the initial score of Random Forest Classifier which fetched a higher accuracy score of 89.69%*

**Concluding Remarks:**

Identifying customer churning not only helps a company to increase its revenue and profit but also it gives an in-depth insight as to what are the highlights and lowlights which a company needs to work on. We started with basic data exploration, identified few customer patterns and features via matplotlib and visualization which further helped us to identify the factors or features responsible for Customer Churning. We later on removed skewness by power transformation and checked multicollinearity issues via Heat map and VIF. There was no Outliers present in the data so we scaled all the feature variables and trained the data into various models. Based on the accuracy score, confusion matrix and cross validation- Random Forest Classifier has been chosen as the best fit model for the Customer Churning Dataset.