Sonali Warkhade

DATA tRAINED

Customer Churn Analysis

In this article, I will go through the whole process of creating a machine learning model on Customer Churn Analysis dataset. It provides information on the customer retention which can be achieved with good customer service and products.

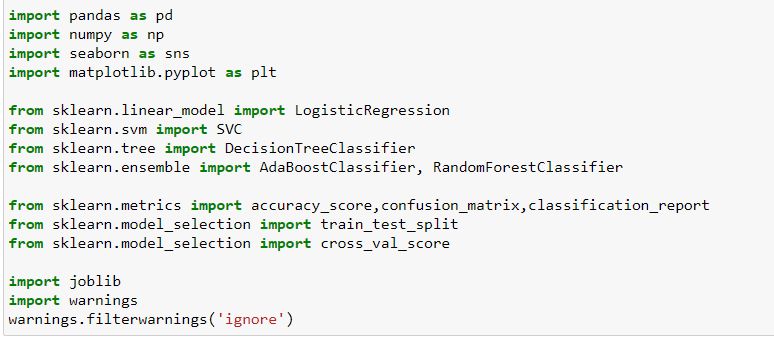
Customer Churn Analysis

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.

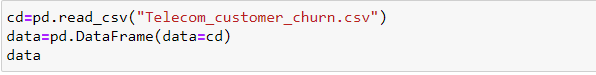
Preventing customer churn is critically important to the telecommunication sector, as the barriers to entry for switching services are so low.

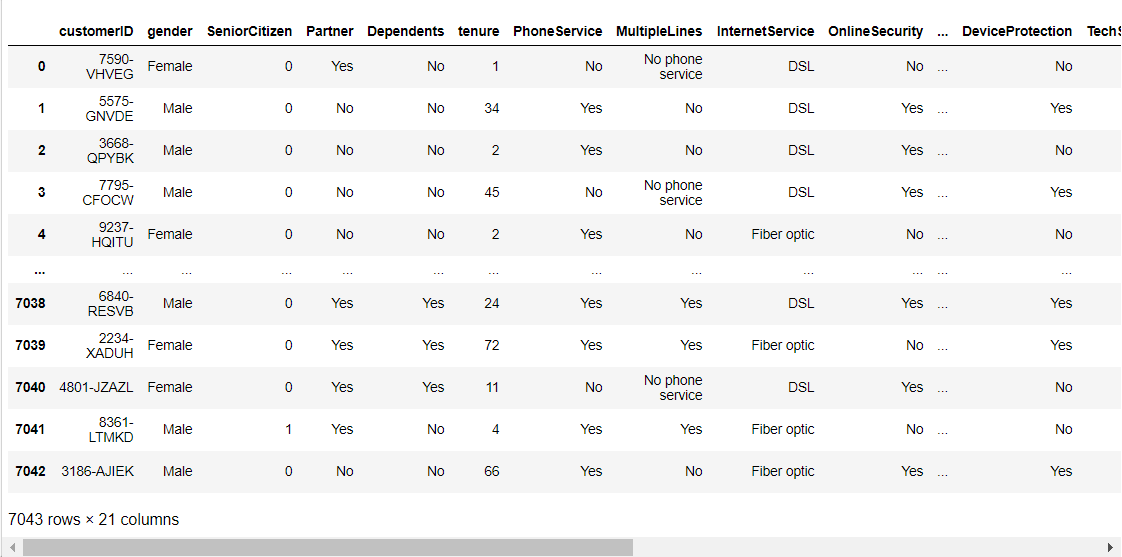
You will examine customer data from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models.

**Imported Libraries:**

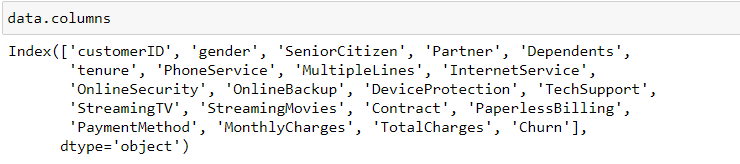


**We will import the csv file for analysis now:**

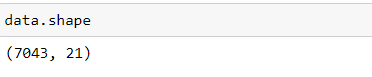


This will give you the overview of data:

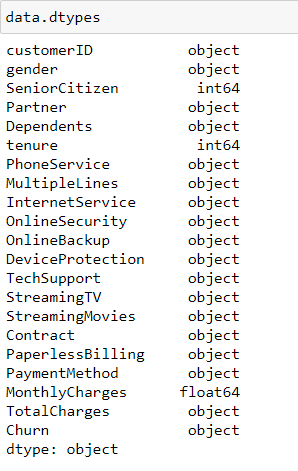
Then I checked for list of all column as we are unable see all column in the above overview:



Above is the list of all columns of our dataset. Let’s check for the number of rows and columns i.e., shape of dataset.



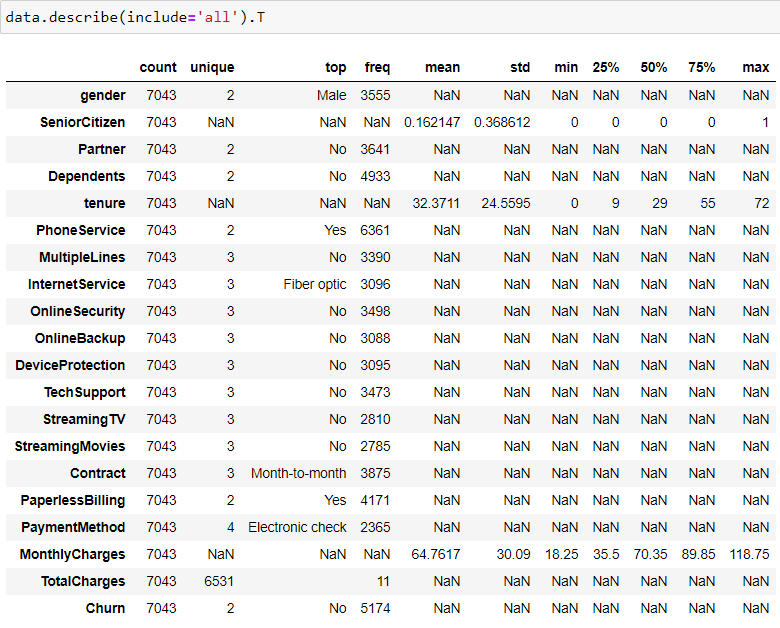
We have 7043 rows and 21 columns. Then I checked for datatypes of all the columns.



We can see maximum columns are of object type, we have 2 columns with integer type and 1 float type. Customer ID is required for further analysis hence I have dropped that column.



Now let's check the describe function for our dataset. As our dataset has maximum object type columns hence, I have used the include parameter. Also, I am reviewing the data in tabular format.

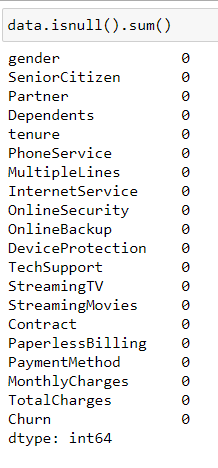


Above output tells us details of all columns.

* Gender column has 2 unique values male and female and with 3555 freq we have maximum male customers.
* In the Senior Citizen column also, we have 2 values yes and no however datatype of this column in integer.
* In the Partner column also has 2 unique values yes and no, 3641 times we have value as ’No’.
* Dependent’s column also has 2 unique values Yes and no, 4933 time we have ‘No’.
* Tenure is the numeric column which has Mean=32.3711, Std=24.5595 min value=0+ and max value=72.
* Again, Phone Service has 2 unique values as Yes and no with 6361 freq we have Yes option.
* In Multiple Lines Column we have 3 unique values and No is with freq 3390 times.
* In the Internet Service column also, we have 3 unique values, Fiber optic is with freq of 3096.
* Online Security column has 3 unique values and No is with freq 3498 times.
* In the Online Backup column also has 3 unique values and no has freq 3088 times.
* Device Protection column also has 3 unique values and no has freq 3095 times.
* In Tech Support column also has 3 unique values and no has freq 3473 times.
* Streaming TV column also has 3 unique values and no has freq 2810 times.
* Streaming Movies column also has 3 unique values and no has freq 2785 times.
* Contract column also has 3 unique values and Moth-to -month has freq 3875 times.
* Paperless Billing column also have 2 unique values and Yes has freq 4171 times
* Payment Method column also has 4 unique values and Electronic check has freq 2365 times.
* Monthly Charges is numerical type hence Mean=64.7617, Std=30.09, min value=18.25, max\_value=118.75.
* Total Charges has 6531 unique values top= Blank(nothing) with freq=11.
* Churn has 2 unique values and no has freq 5174 times.

Total Charges column actually contains numerical values however its datatype is object. Also, the top frequency of that column is on blank value.

Let’s check for the null values first then change the datatype.

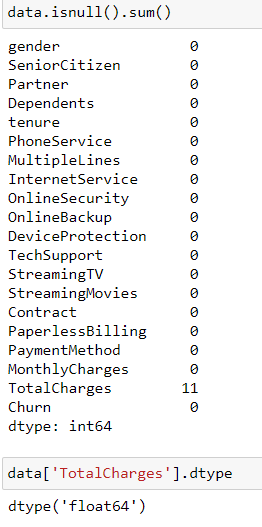


Above output shows that our database does not contain any null values; however, describe function details speaks to the fact that we have some black values in Total Charges.

Let’s first change the datatype to numerical.



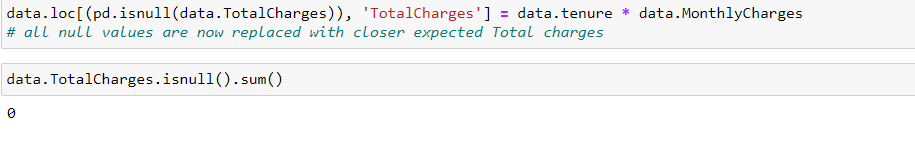
Above code changed the datatype to numeric. I have used parameter errors=coerce will change all error values to Nan (where function is unable to change the value to number then that will be replaced with Nan). Now let’s check for the null values again. Also, we will check the datatype of Total Charges Column.



As we can see, we have 11 rows with total charges value as null. Now we have some options like dropping these columns, the second option is replacing it with mean however as per my observation we can replace the null value by doing some calculation, we might not get the exact value but we can get closest value rather than dropping the row or replacing it with mean. We can see total charges column has some relationship with tenure and monthly charges

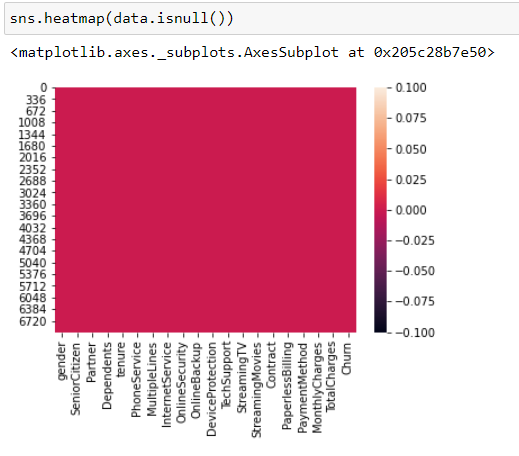
i.e., tenure \* monthly charges + some charges = total charge.

I am unaware of some charges but we get somewhat closer to the expected value.

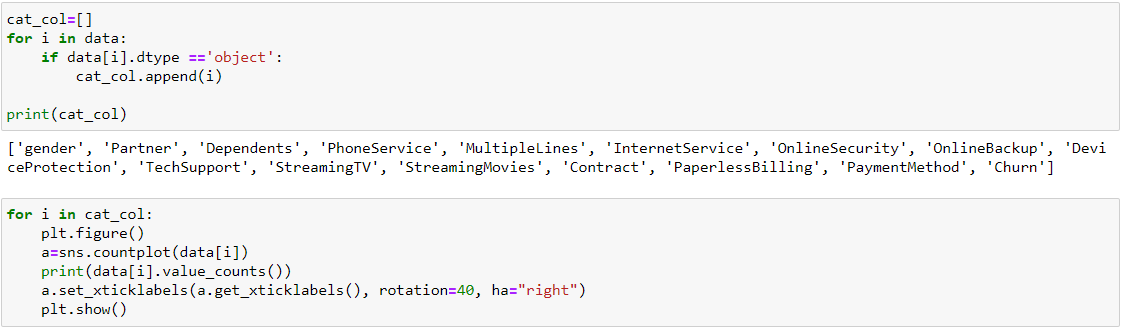


Now we don’t have any null values. Let’s see the same thing with the help of heatmap:

**Visualization:**

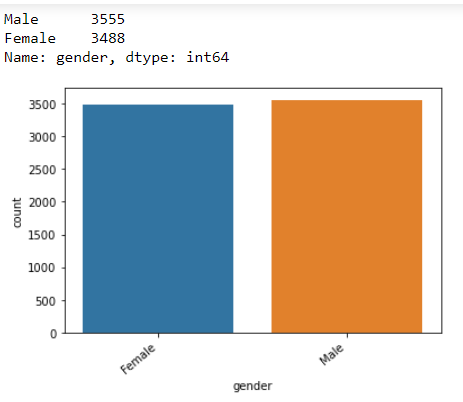


Then I have created to list of all object type column for visualization. Then with the help **countplot** we will compare all value of all the columns.



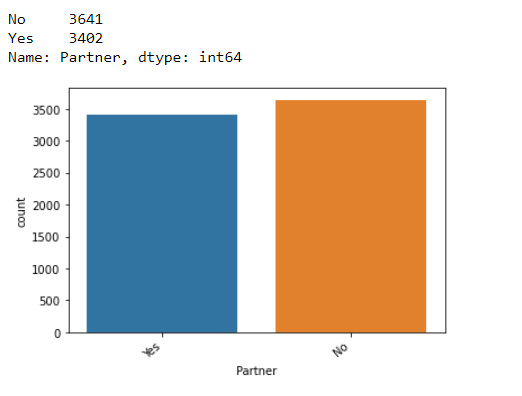
We can see the list of all object type columns now let’s see the plots for all columns.

Gender:



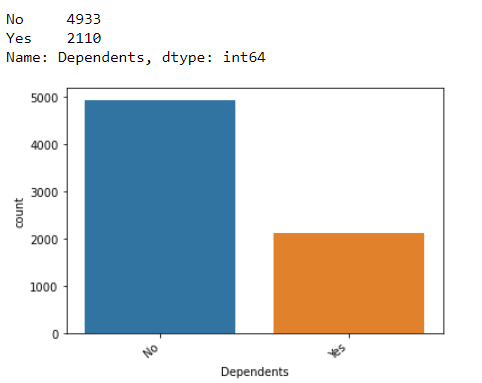
We have 3555 male and 3488 female customers.

Partner:



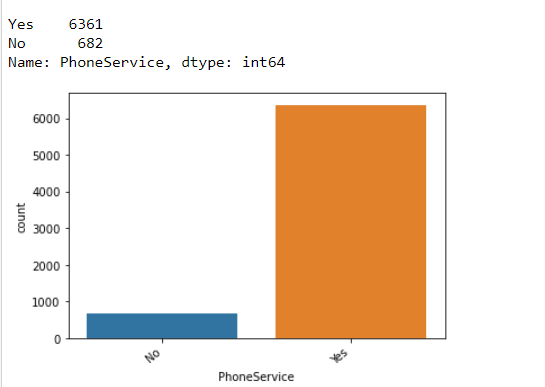
3641 customers are not having any partner and 3402 have partnerss.

Dependents:



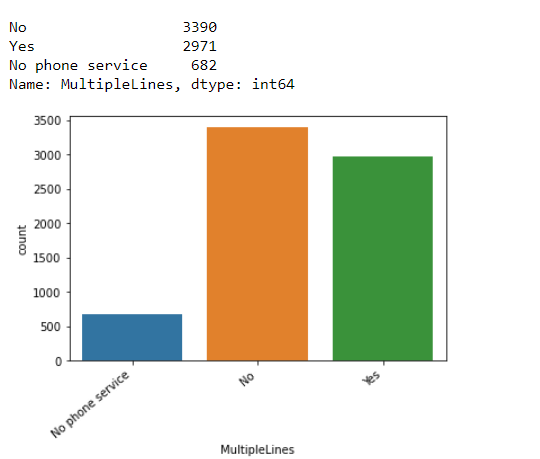
4933 customers are with no dependents and 2110 are with dependents.

Phone Service:



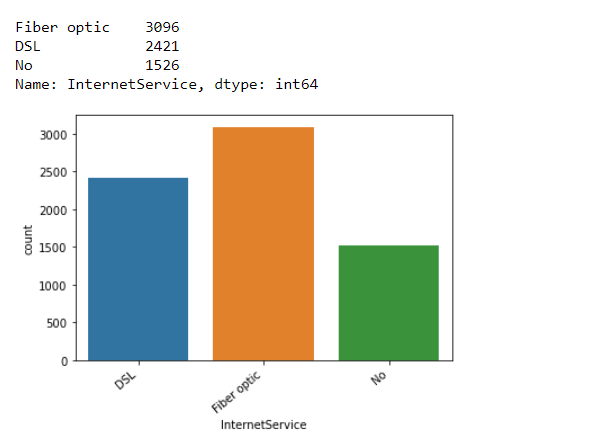
Customers who have phone service are 6361 and 682 does not have any phone service.

Multiple Lines:



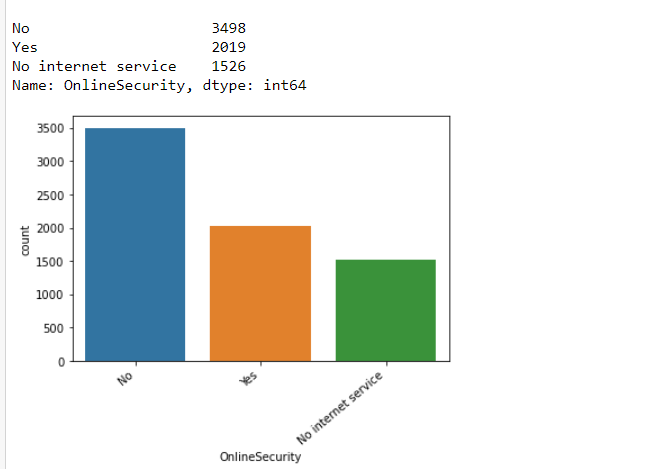
Maximum customer does not have multiple lines then 2971 have multiple lines and then 682 does not have phone service.

Internet Service:



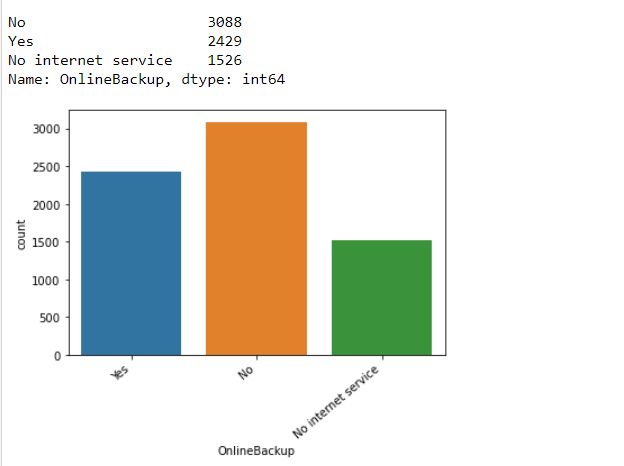
We have 2 types of internet service DSL, Fiber optic, 3096 customers are using Fiber optic and 2421 are using DSL, 1526 are not using internet service.

Online Security:



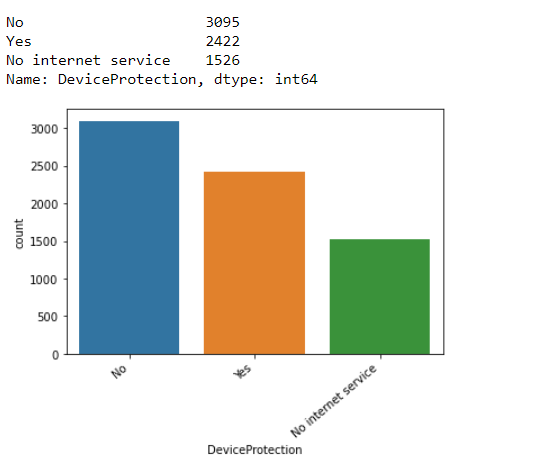
3498 customers do not have online security and 2019 customers have internet security. 1526 are not having internet services only.

Online Backup:



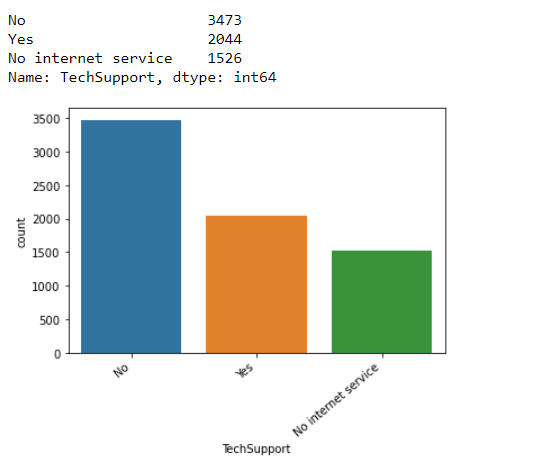
3088 customers are not having any online back-up and 2429 have online backup. 1526 are not having internet services only.

Device Protection:



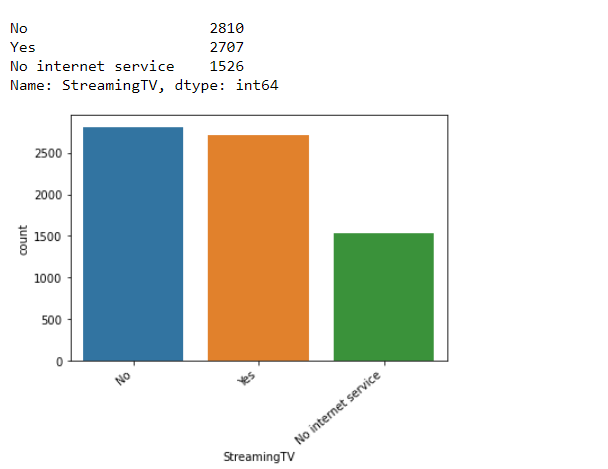
Customers who are using device protection are 2422 and those who are not using 3095,1526 are not having internet services only.

Tech Support:



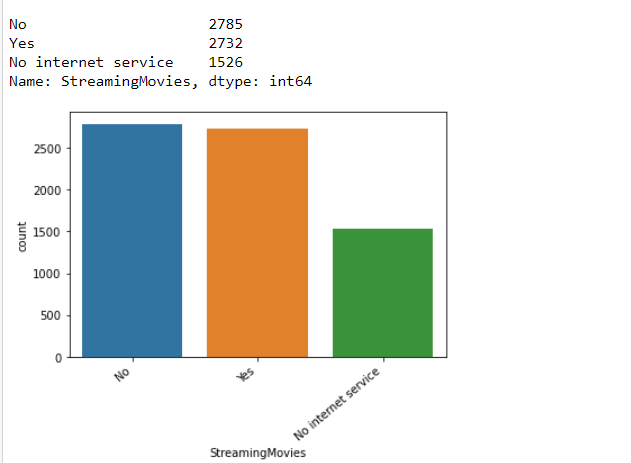
Customers who have technical support are 2044 and who do not are 3473. 1526 are not having internet services only.

Streaming TV:



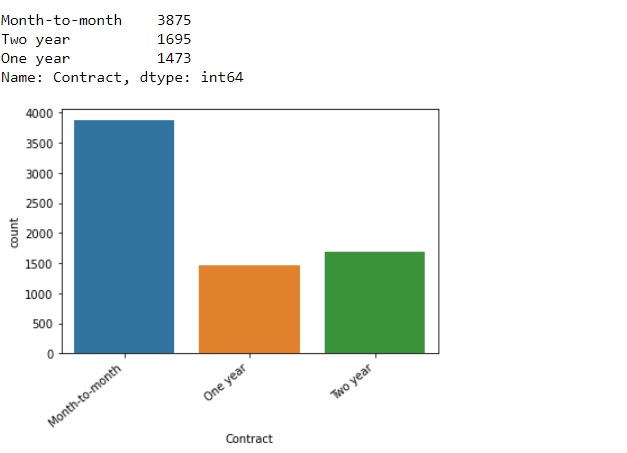
Customers who are streaming TV are 2707 and 2810 are not streaming TV. 1526 are not having internet services only.

Streaming Movies:



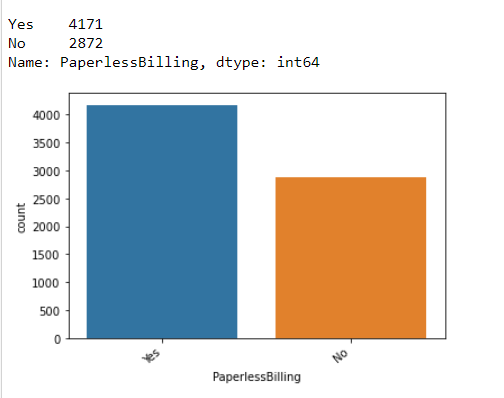
Customers who are streaming Movies are 2732 and 2785 are not streaming TV. 1526 are not having internet services only.

Contract:



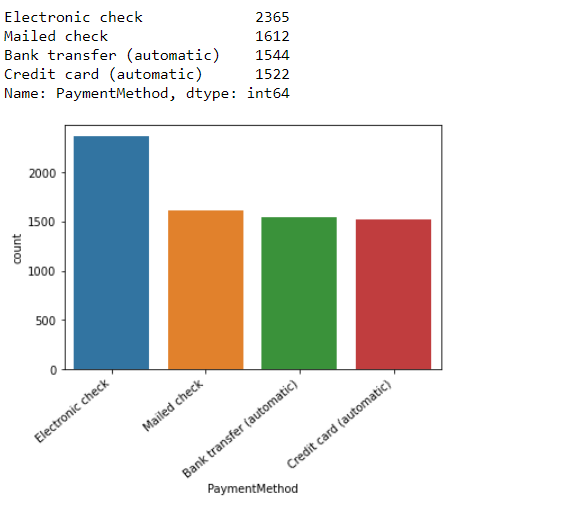
Customers with month-to-month contract are 3875 and two-year contract is 1695 lastly one-year contract are 1473.

Paperless Billing:



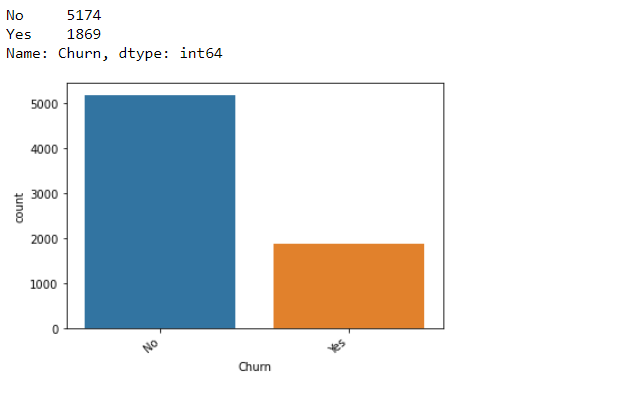
Customers with paperless billing are 4171 and not paperless billing are 2872.

Payment Method:



Customers with electronic check payment mode are 2365, mailed checks are 1612, 1544 customers are paying with bank transfer(automatic) and 1522 are doing payment with credit card (automatic).

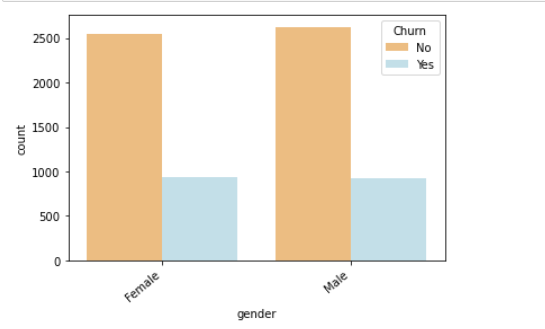
Churn:



Churned customers are 1869 and 5174 are not churned.

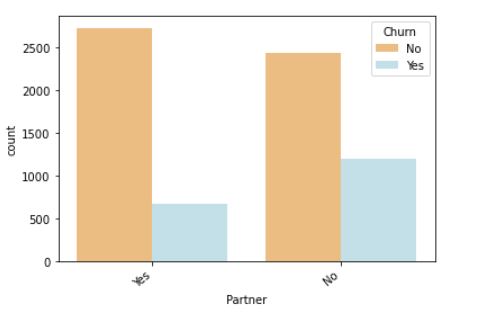
Then with the help of **countplot** I have compared all object type columns with our target variable churn.

Gender:



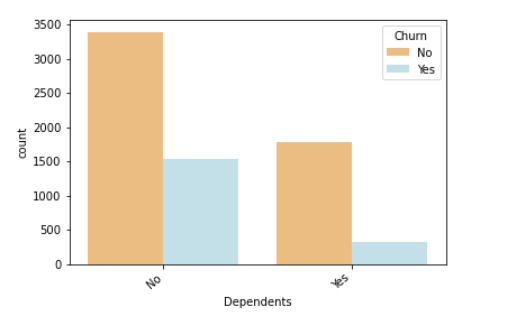
Churned male and females are almost the same.

Partner:



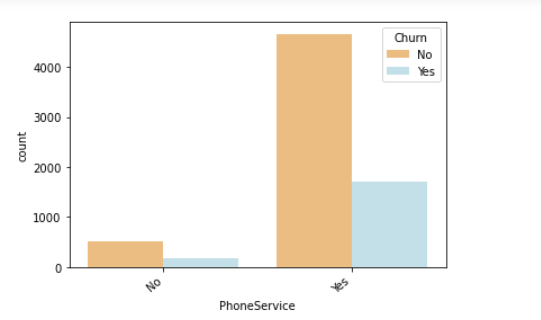
Customers who have partners are churned less than those who do not have partners.

Dependents:



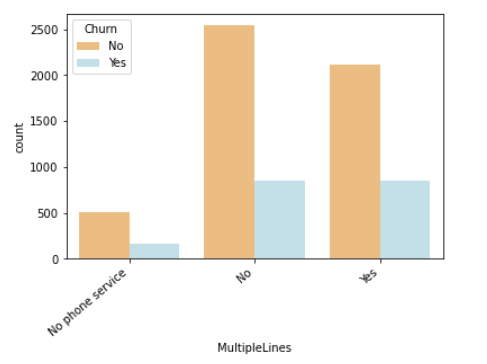
Customers with no dependents are churned more than with dependents.

Phone Service:



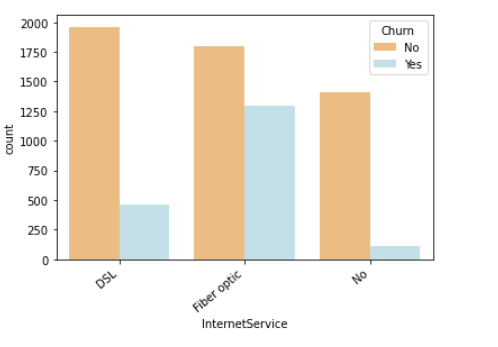
Customers who have phone service are churned more.

Multiple Lines:



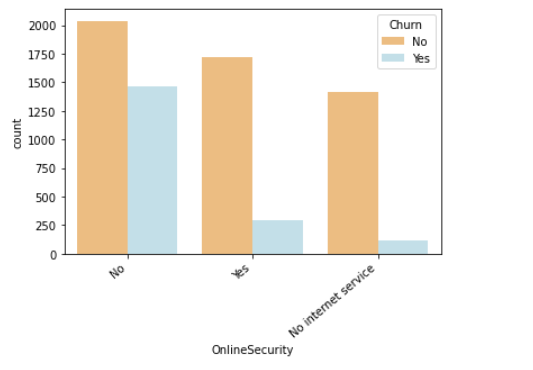
Customer does not have multiple lines and has multiple lines are churned almost by the same numbers.

Internet Service:



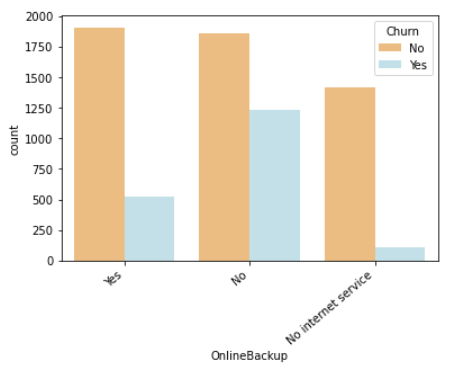
Customers with Fiber optic are churned more than DSL churned.

Online Security:

****

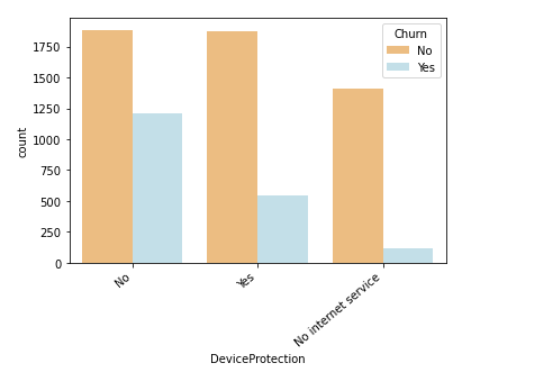
Customers who do not have online security are churned most.

Online Backup:



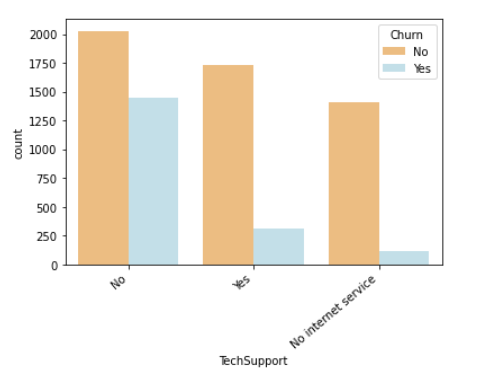
Customers not having any online back-up are churned most.

Device Protection:



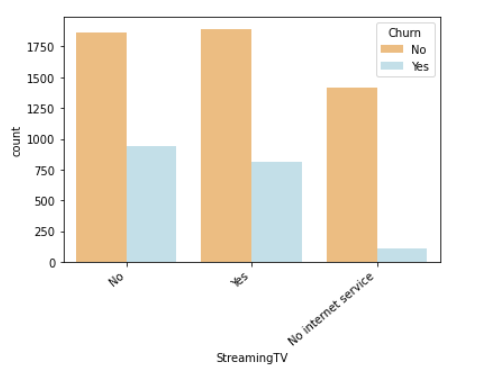
Customer who are not having device protection are churned the most.

TechSupport:



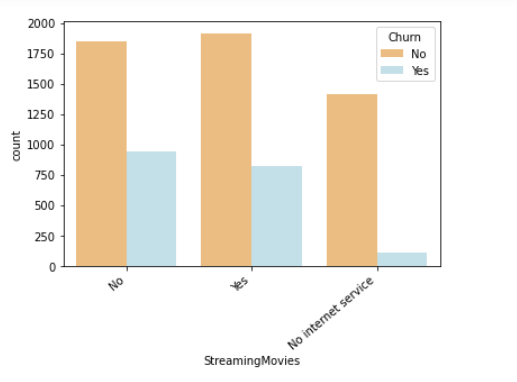
Customers who do not have technical support are churned more.

Streaming TV:



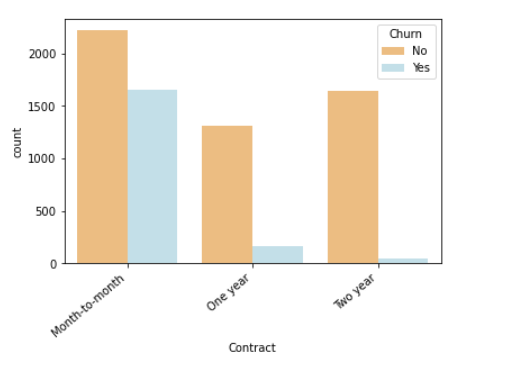
Customers who are not streaming TV are churned more.

Streaming Movies:



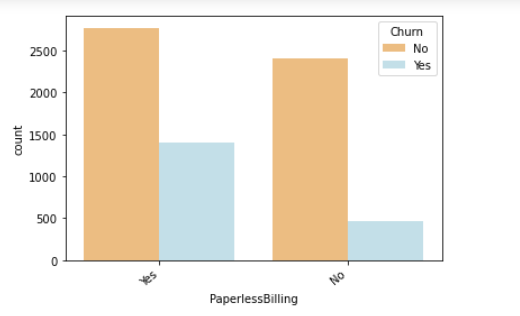
Customers who are not streaming Movies are churned more.

Contract:



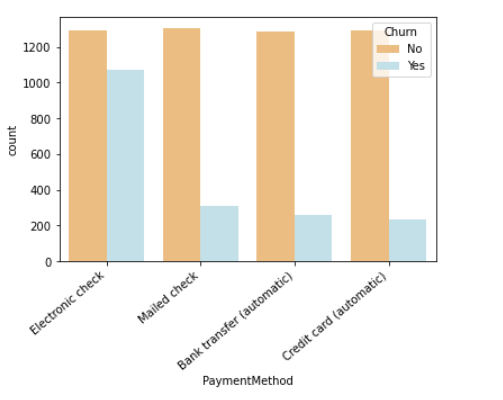
Customers with month-to-month contract are churned more.

Paperless Billing:



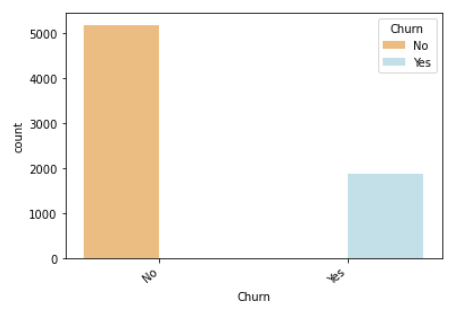
Customers with paperless billing are churned more.

Payment Method:



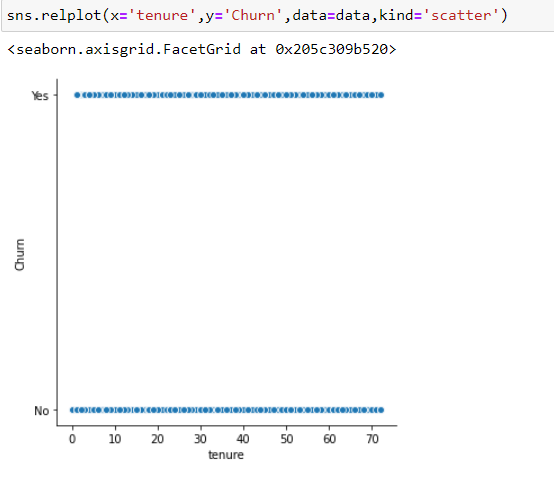
Customers with electronic check payment mode churned more.

Churn:



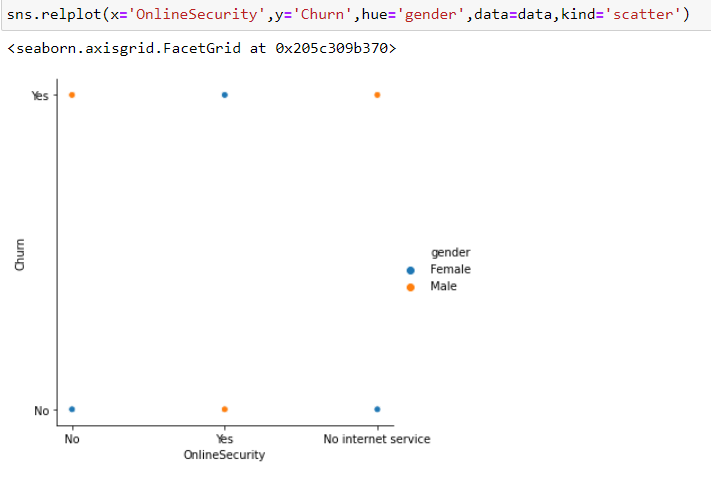
Not churned customers are more.

With the help on **relplot** let’s compare Tenure and churn column:



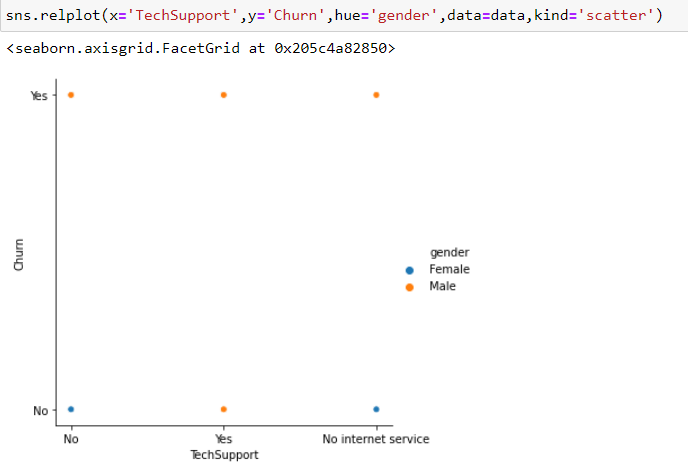
Customers who are churned and not churned are distributed in the entire range of tenure.

Then I have compared online security and churn with respect to gender.



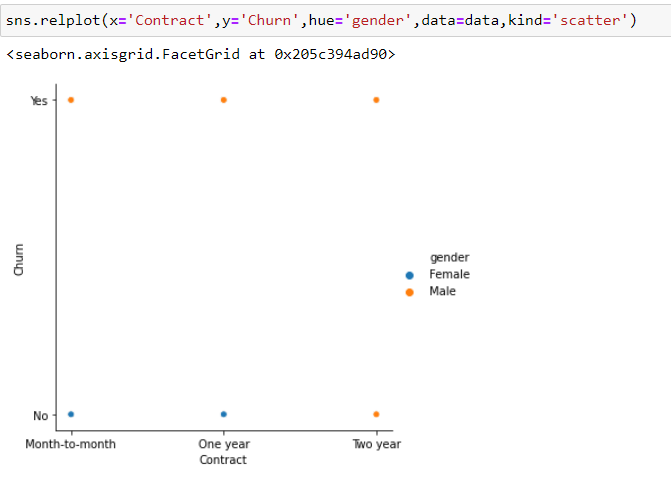
Churned male customers are with no online security or no internet service and female churned customers are with online security.

Then I have compared TechSupport and churn with respect to gender.



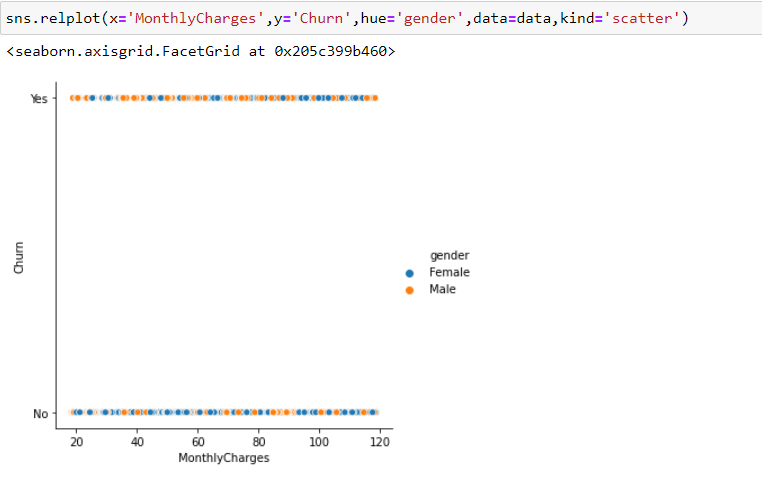
All churned male customers are having technical support.

Then I have compared contract and churn with respect to gender.



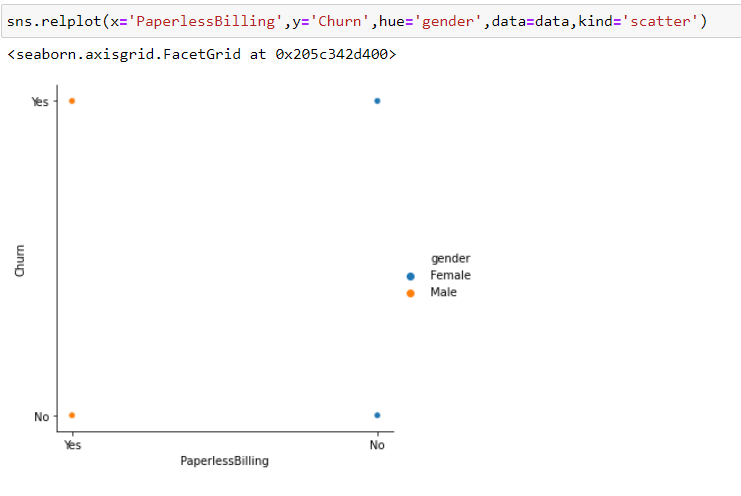
All churned males are having all three types of contract i.e., month-to-month, one year and two-year contract.

Then we will see Monthly charges and churn with respect to gender.



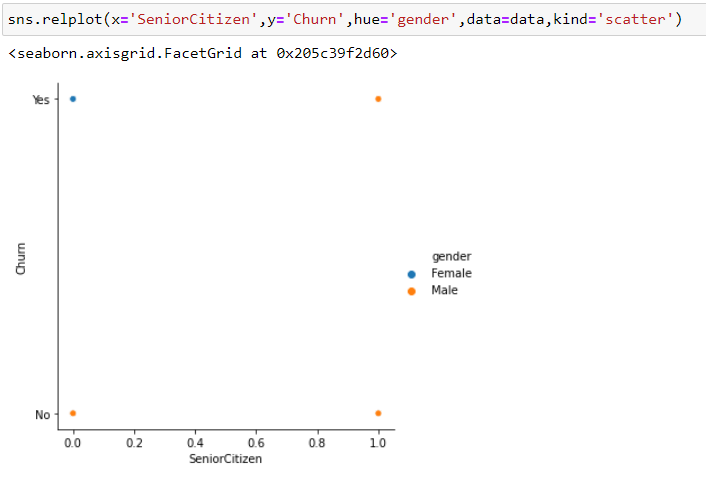
For the entire range of monthly charges, we have churned and not churned males and females both.

Then I have compared paperless billing and churn with respect to gender.

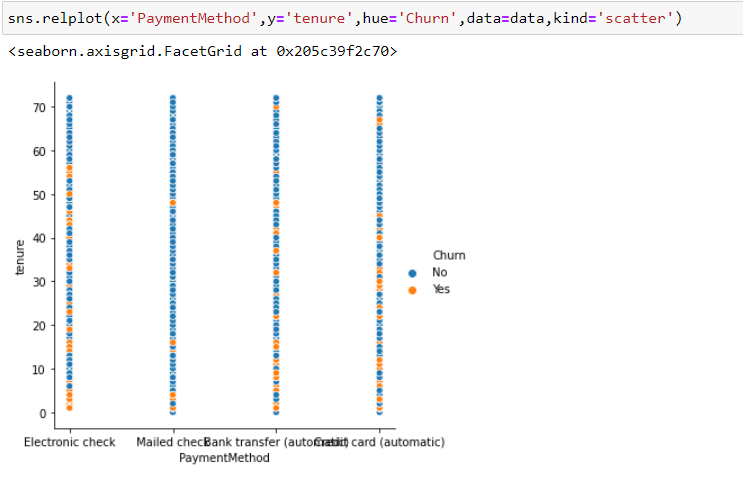


Both male and female customers who are churned have opted for paperless Billing.

Next is Senior Citizen and churn with respect to gender.

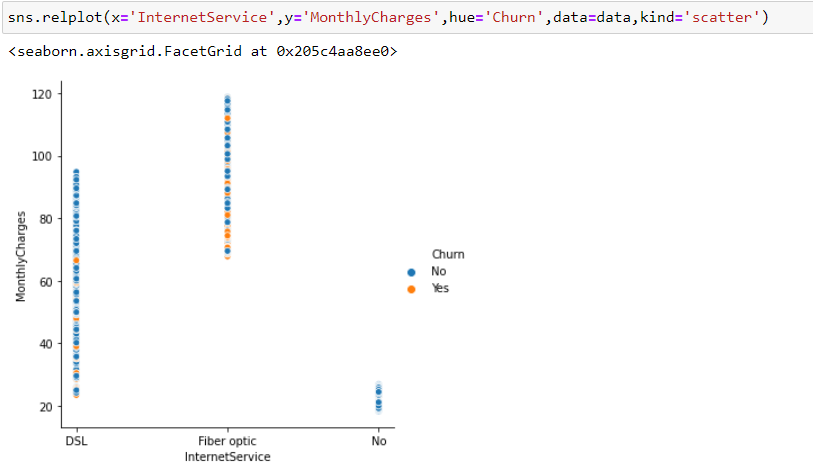


Churned male customers are senior citizen however churned female customer are not senior citizens. Then Payment Method and tenure with respect to churn.



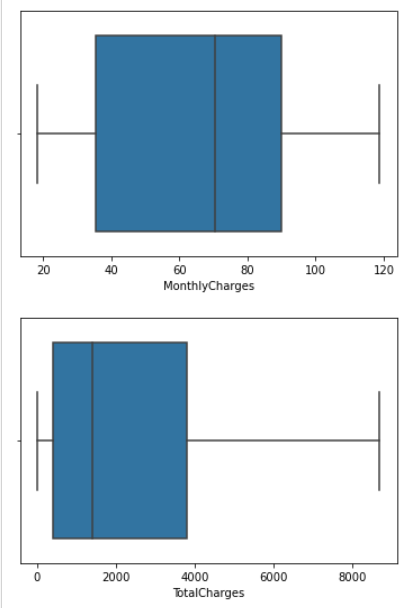
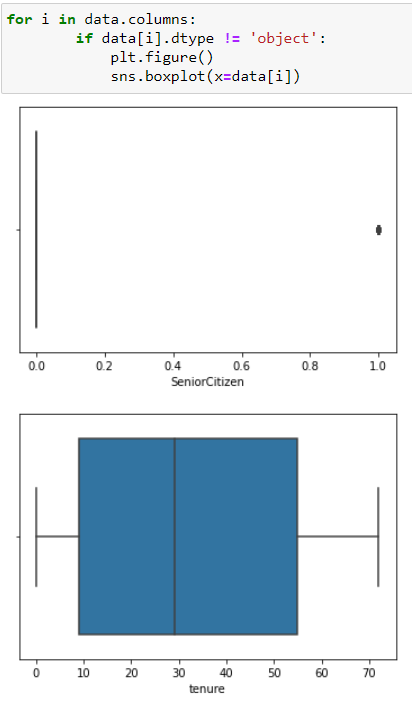
For the entire range of tenure and in all types of payment method, we have churned and not churned both the customers.

Next, I have compared Internet Service and monthly charges with respect to churn.

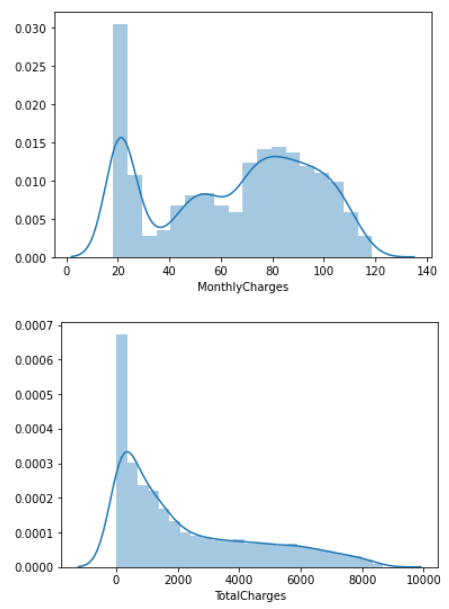
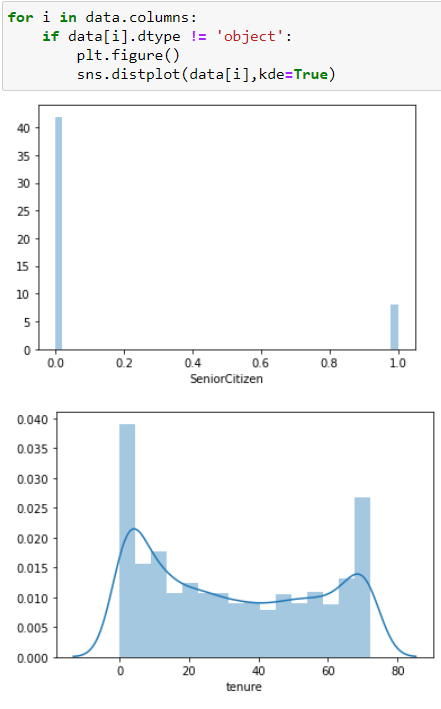


Monthly charges for DSL internet service are ranging from 20 to 100 so for fiber optic it’s ranging from 60 to 120. We have some customers for DSL and for fiber optic numbers are for churned customers.

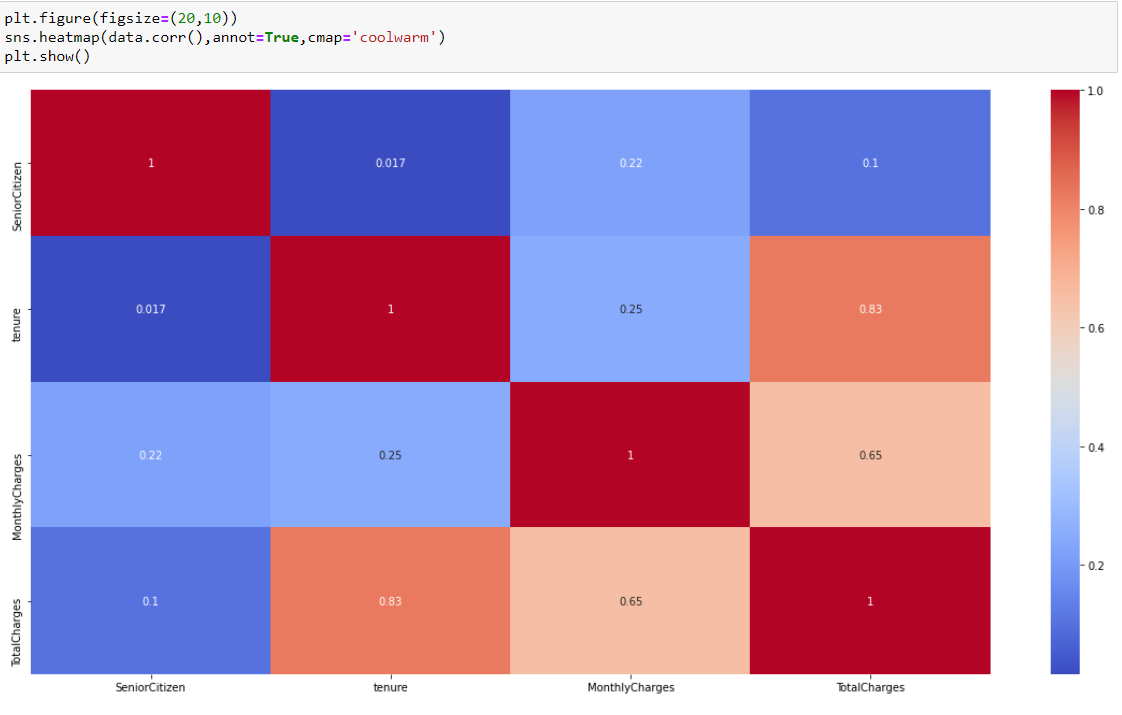
Now let’s view the numerical columns data with the help of Distplot and boxplot.



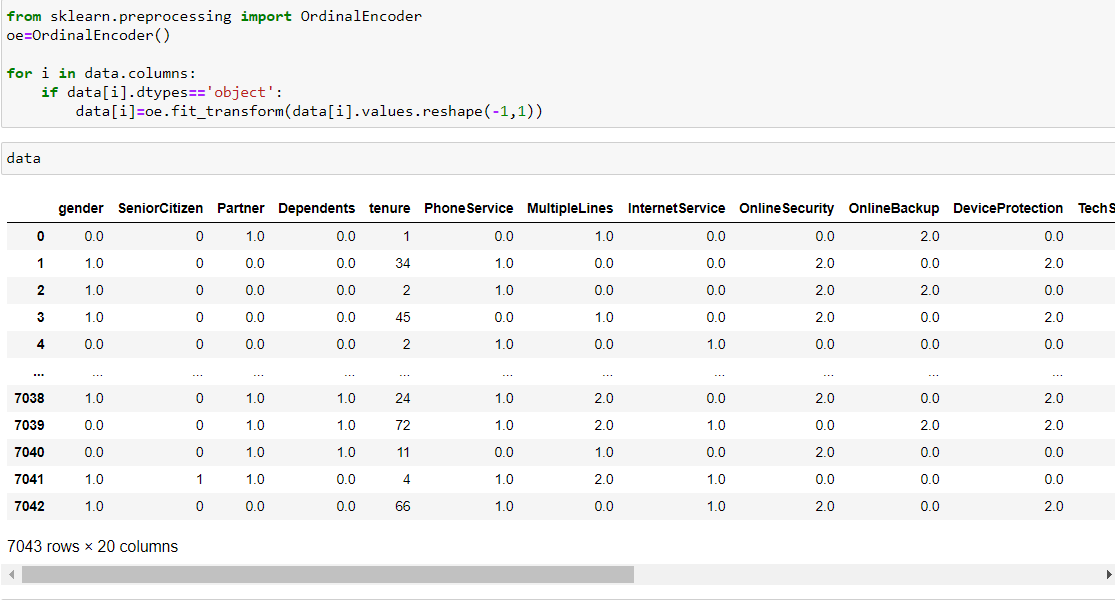
We cannot see any outliers in the numerical type of column. Senior citizen is actually an object type of column.

We can see uneven distribution in all numerical columns.

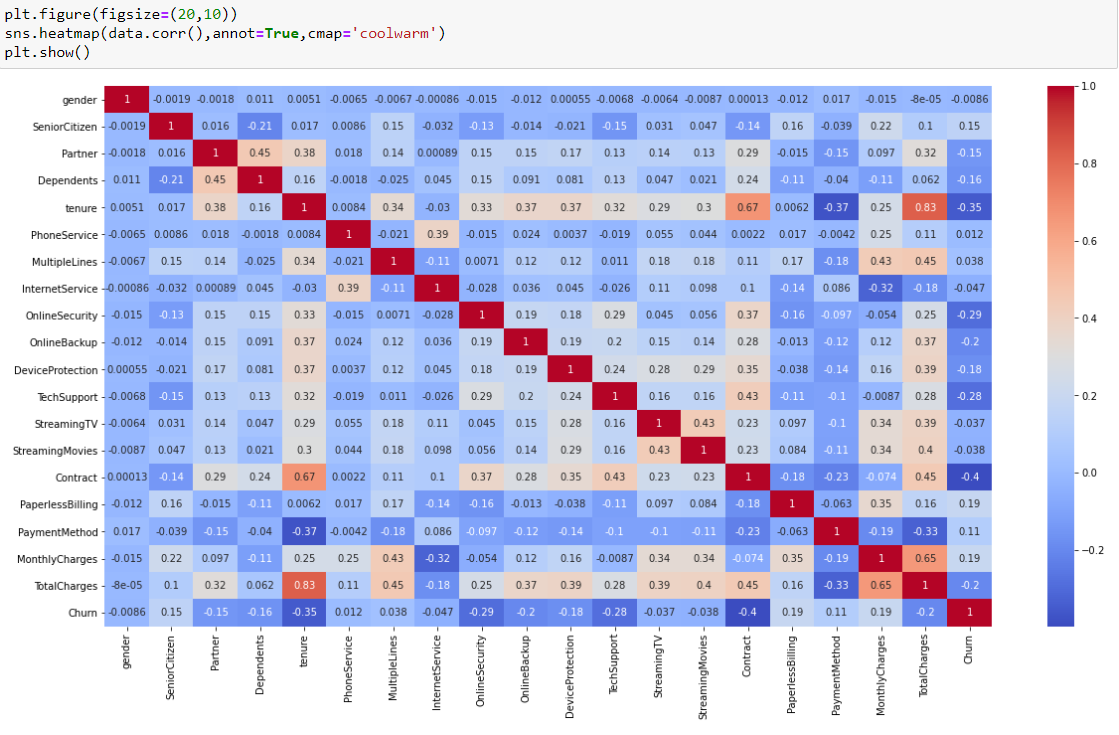
Then I have checked for correlation:

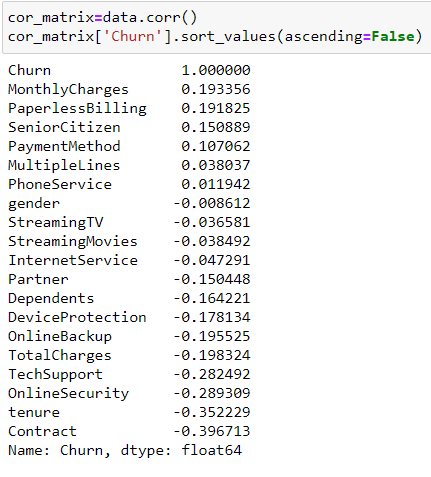


We need to check the correlation of all columns with or target column ‘Churn' however datatype of target column is Object hence we cannot see that column. Let's change the data type of target column and other object type columns too. I am using an ordinal encoding method for encoding.



We have changed all object type columns. We can see the changed details of the dataset above. Let’s check the correlation of the entire dataset now.

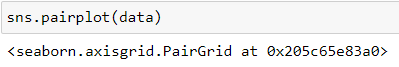


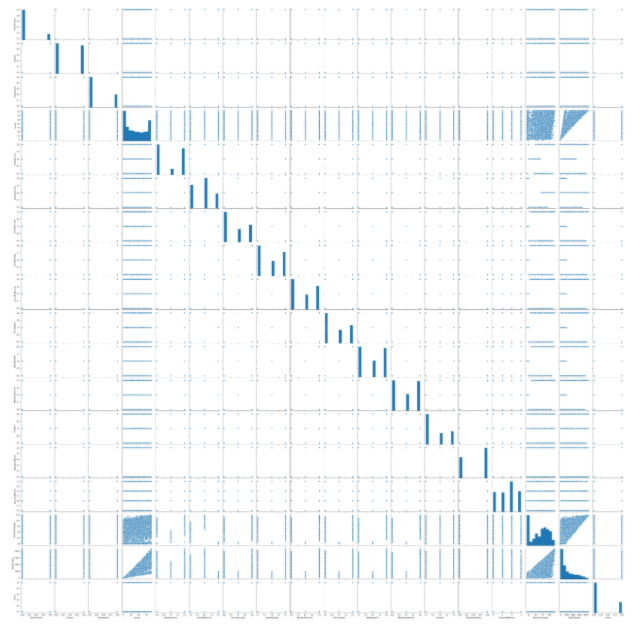


Above are the details of correlation of all columns with our target column Churn. We can see Churn is positively correlated with Monthly Charges, Paperless Billing and Senior Citizen. And it has a snegative correlation with Contract, tenure, Online Security, TechSupport, Online Backup. Churn has very less correlation with Phone Service and gender hence we can drop these columns.



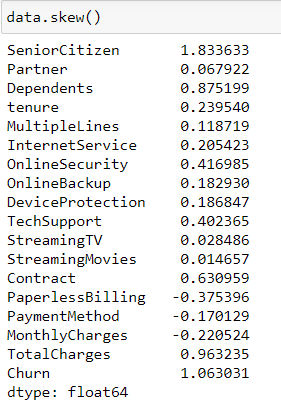
Let’s see the relation of columns with all columns with the help of pairplot.





We can see a linear relationship between total charges tenure and Also in between total charges and monthly charges.

Then I have checked skewness and outliers:

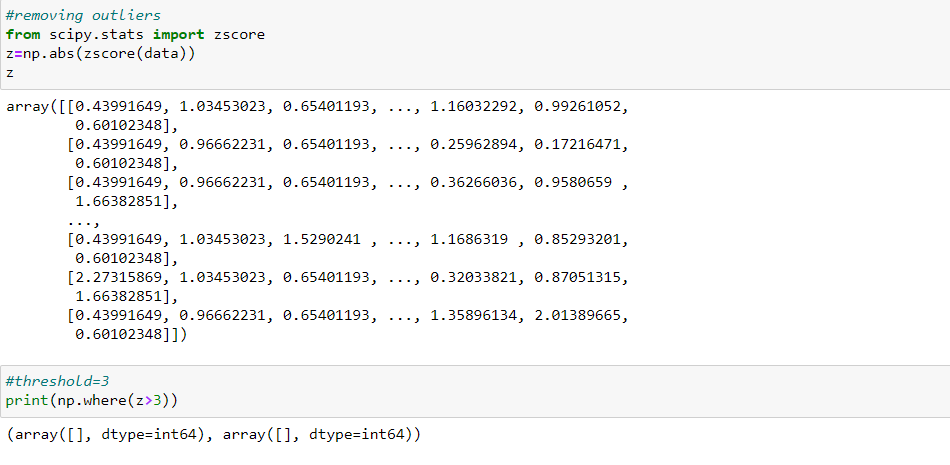


Skewness range is considered as +/-0.5. Below is a list of columns which does not fall in this range.

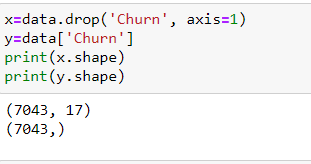
1. Senior Citizen
2. Partner
3. Dependents
4. Contract
5. Total Charges
6. Churn

For Categorical type of columns, skewness is not considered. And Churn is our target column.

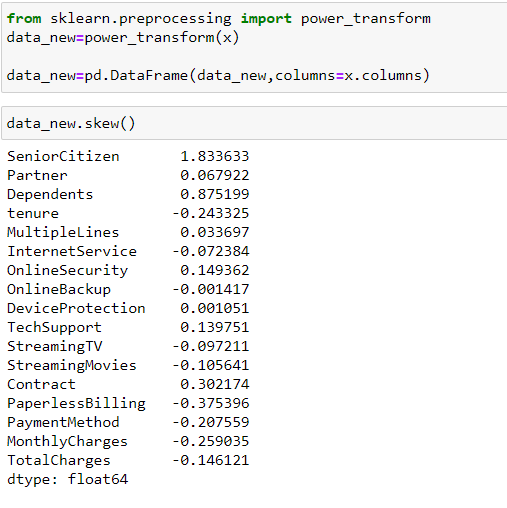
Then I have used zscore method for removing outliers.



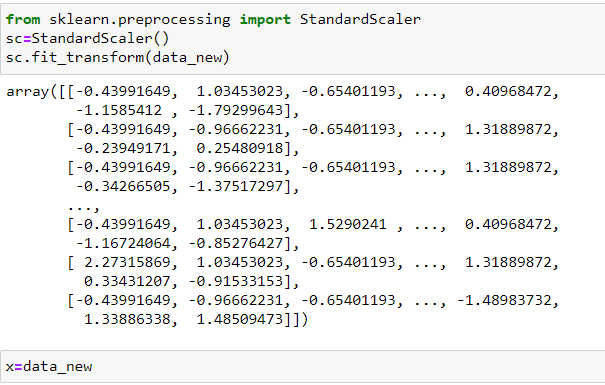
We cannot find any outliers hence we will move on to the scaling but before that we will split the dataset in dependent variables (y/target column) and independent variables(x).



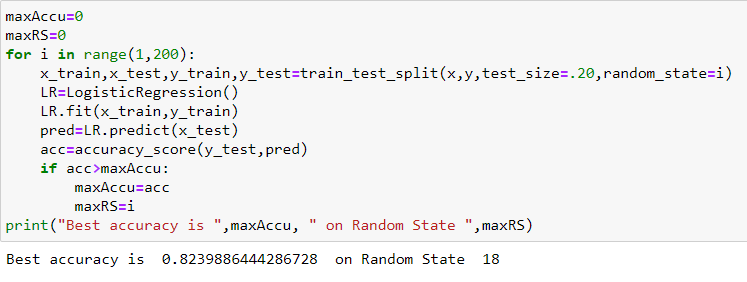
Then I have used the power transform method to reduce skewness.



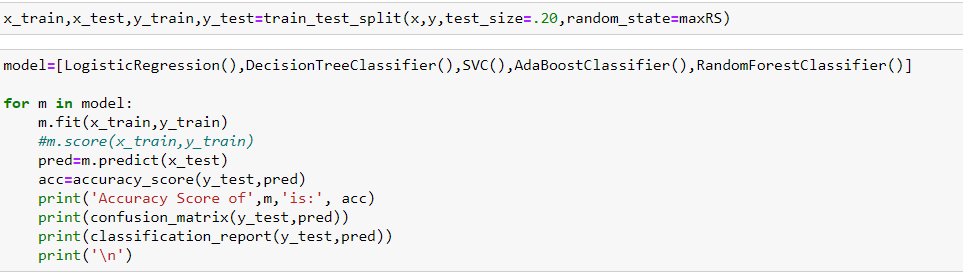
Skewness has now been reduced. Let’s do the scaling now.



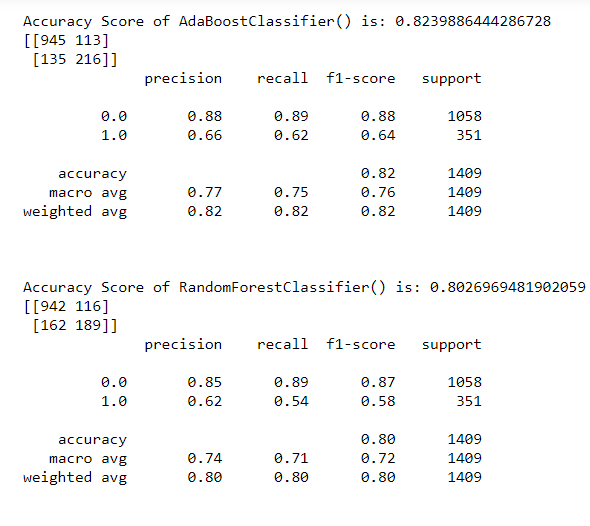
Before checking the best classification method, I am checking for the best random state.



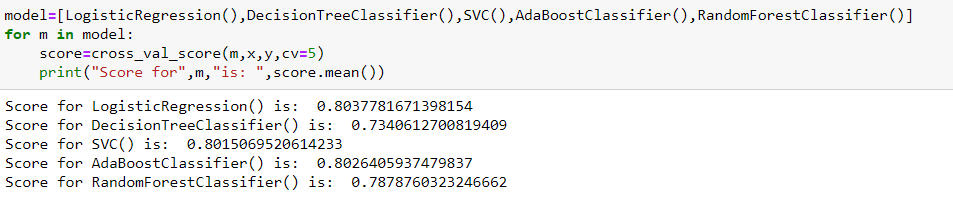
We found the best random state at 18. Let’s check which classification method gives us the best output.



Output of above code will give me Accuracy score, confusion matrix and classification report of all models.



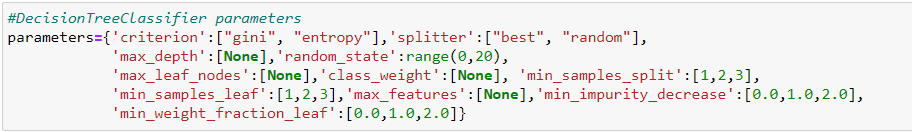
We can see the highest accuracy score is for Logistic regression and AdaboostClassifier however before proceeding ahead let’s check the cross-validation score.



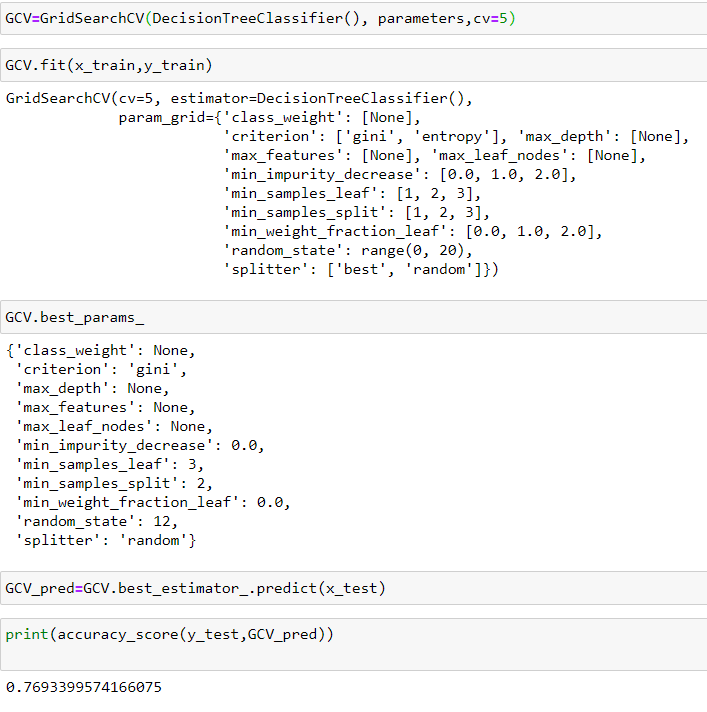
We can see the highest accuracy score is of logistic regression however With the DecisionTreeClassifier model there is very little difference in accuracy and cross validation score, Hence the best model is DecisionTreeClassifier.

Let’s try to improve performance of the model by doing hyper parameter tuning with GridSearchCV. Let’s import required libraries and create a dictionary for parameters.

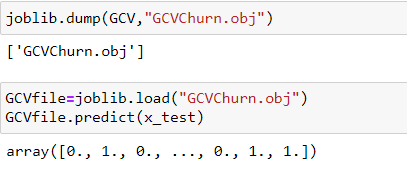




I have then called the model and tried fitting out the train set i.e., x=train (Input variables), y\_train (target variable). Let’s see the best fitted parameters found then predict output for input variables(x\_test) and then check for accuracy percentage.



Now the improved accuracy score is 76.93%, then I have saved the model then called and predict the target variable.



Summary

We started with the data exploration where we got a feeling for the dataset, checked about missing data and learned which features are important. During this process we used seaborn and matplotlib to do the visualizations. During the data preprocessing part, we computed missing values, converted features into numeric ones. Afterwards we started training 5 different machine learning models, picked one of them (DecisionTreeClassifier) and applied cross validation on it. Then we looked at how DecisionTreeClassifier performs, look a look at the importance it assigns to the different features and tuned it’s performace through optimizing it’s hyperparameter values. Lastly, we have saved the model then recall the model and then predict the data.