**Customer Churn Prediction of a Telecom Company Using Machine Learning**

Hi Everyone, I am Sankalp Mahapatra and in this blog I will be showing you how to predict the Customer Churn of a Telecom Company by using different Machine Learning models.



**Introduction**

In this Project I have done through analysis of factors affecting directly or indirectly the Customer Churn. In this Project first I have imported the Customer churn dataset from github and then I have performed the Exploratory Data Analysis(EDA) on the dataset and then I have used Different Imputing techniques, Encoding techniques and deleted the unwanted columns where ever needed. Then I have Performed the Data visualization by using different plots. I have noted down all the observations while building the model. And at the end I have used different machine learning models to predict the Customer churn and After that I have used the Hyperparameter tuning technique on the best Machine Learning model and then I have saved the best machine learning model.

**1.Problem Definition**

The problem statement is briefly explained below.

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.

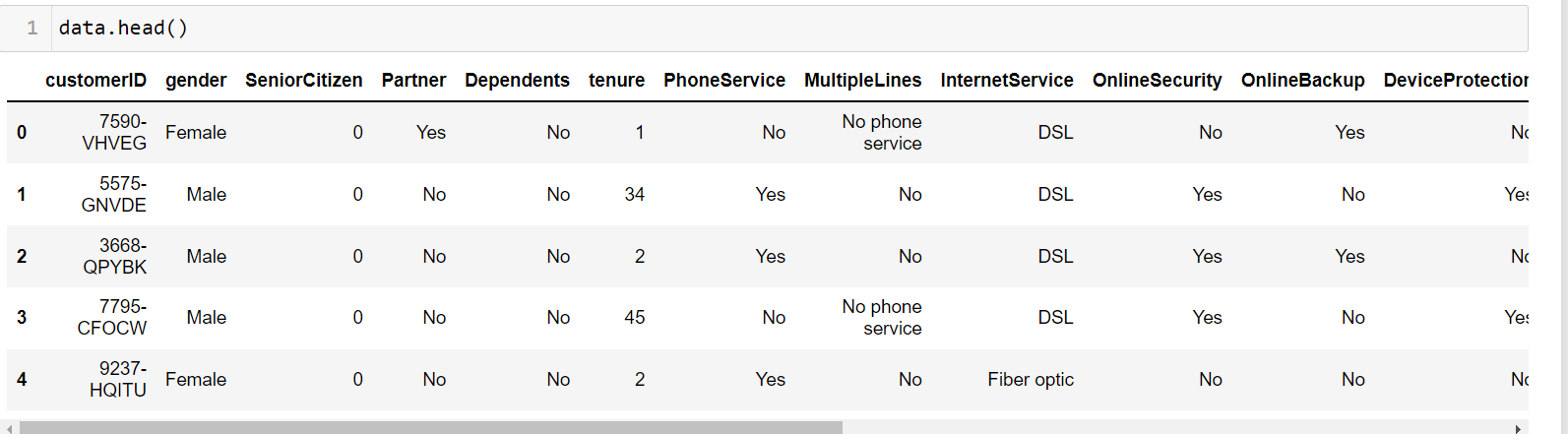
Existing customers will often have a higher volume of service consumption and can generate additional customer referrals.

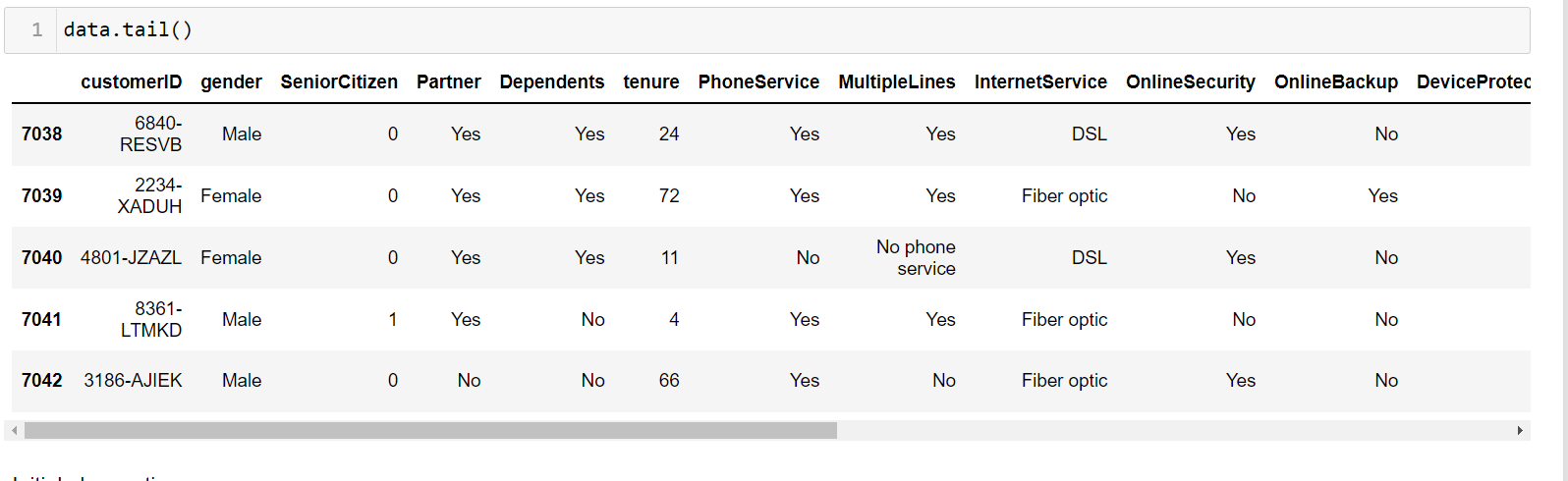
Customer retention can be achieved with good customer service and products. But the most effective way for a company to prevent attrition of customers is to truly know them. The vast volumes of data collected about customers can be used to build churn prediction models. Knowing who is most likely to 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.

**2. Data Analysis**

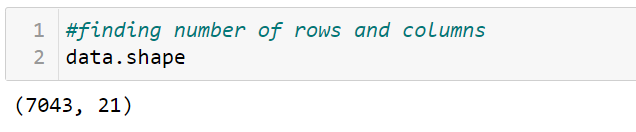
Now let’s analyse our dataset and see what initial information we can extract from the dataset. I have used head() and tail() methods to print the first 5 and last 5 records of the dataset.

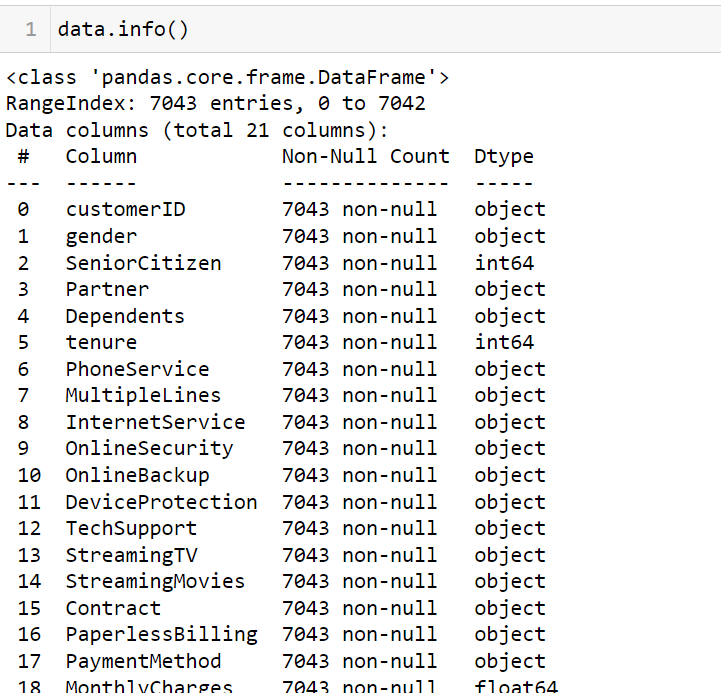




* The dataset has 7043 records 21 columns
* Both numerical and categorical data are present in the dataset.
* The column names are 'customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents','tenure', 'PhoneService', 'MultipleLines', 'InternetService','OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport','StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling','PaymentMethod', 'MonthlyCharges', 'TotalCharges', and 'Churn'.
* We can observe many columns having object(string) value in the dataset.
* Based on all the features we have to predict the costumer churn of a telecom company.
* The 'churn' column has only values of yes and no, so we can clearly say that this is a classification problem as the target variable or the label has categorical data.

Now we will analyse more deeply about the dataset. We will find out about the differnt datatypes present in the dataset. And We have to make sure every column is in numerical format as we can only build machine learning models with numerical data.





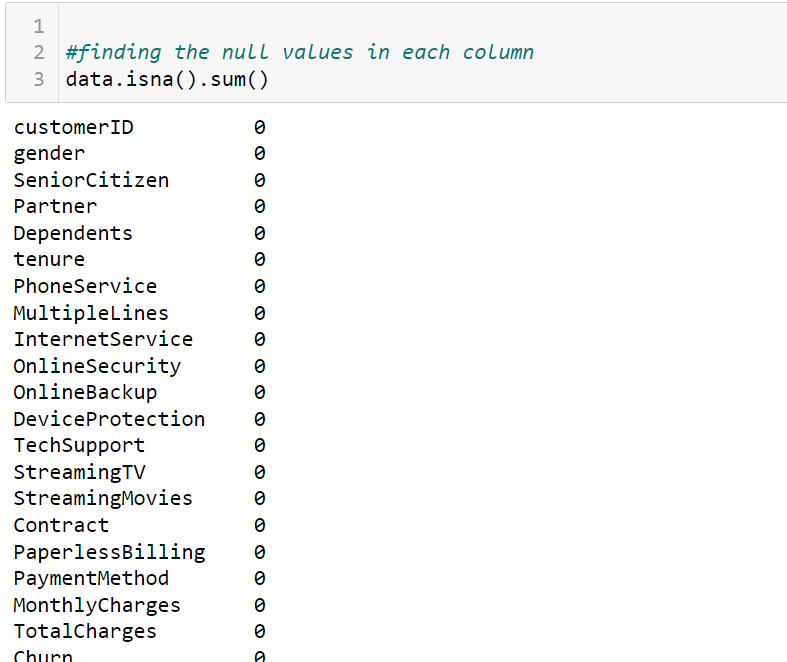
Most of the columns in the dataset are having object(string) datatype values. We need to encode the object values into numbers for better analysis.

**3.EDA Concluding Remarks**

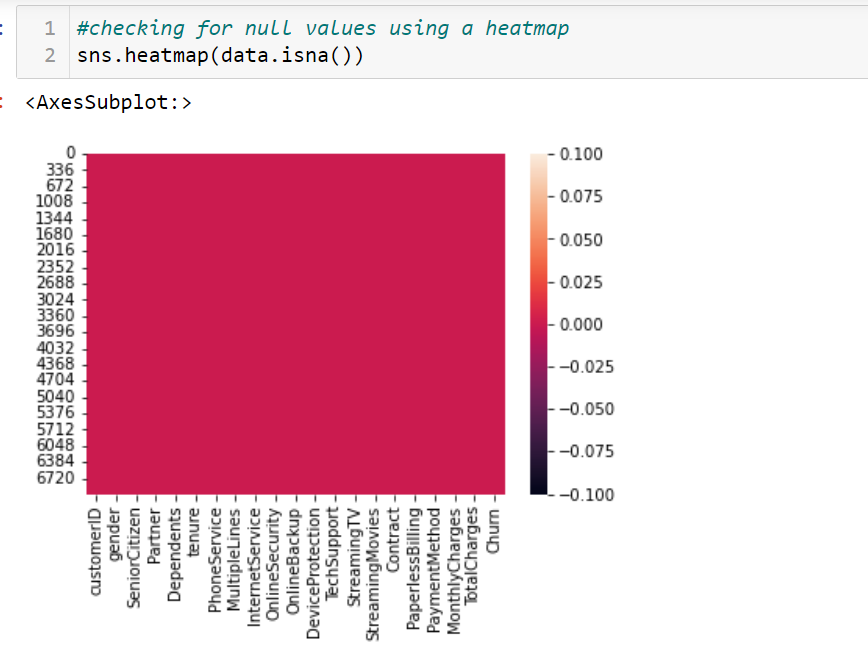
Now we will perform the Exploratory Data Analysis(EDA) on the dataset by involving the following steps.

* Checking null values
* Statistical Summary
* Plotting Count plots
* Checking the distribution of all the features
* Removing outliers
* Checking relationship between the features (multicolinearity)
* Checking relationship between the features and labels

**Checking null values**

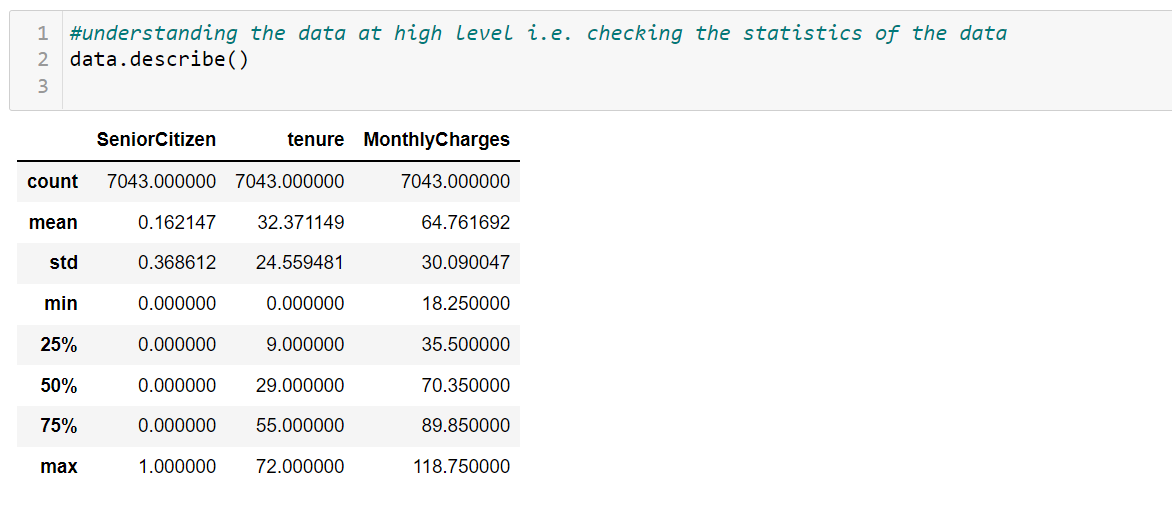
****

No null values are present in the dataset, let’s recheck our findings by plotting a heatmap plot.



**Statistical Summary**

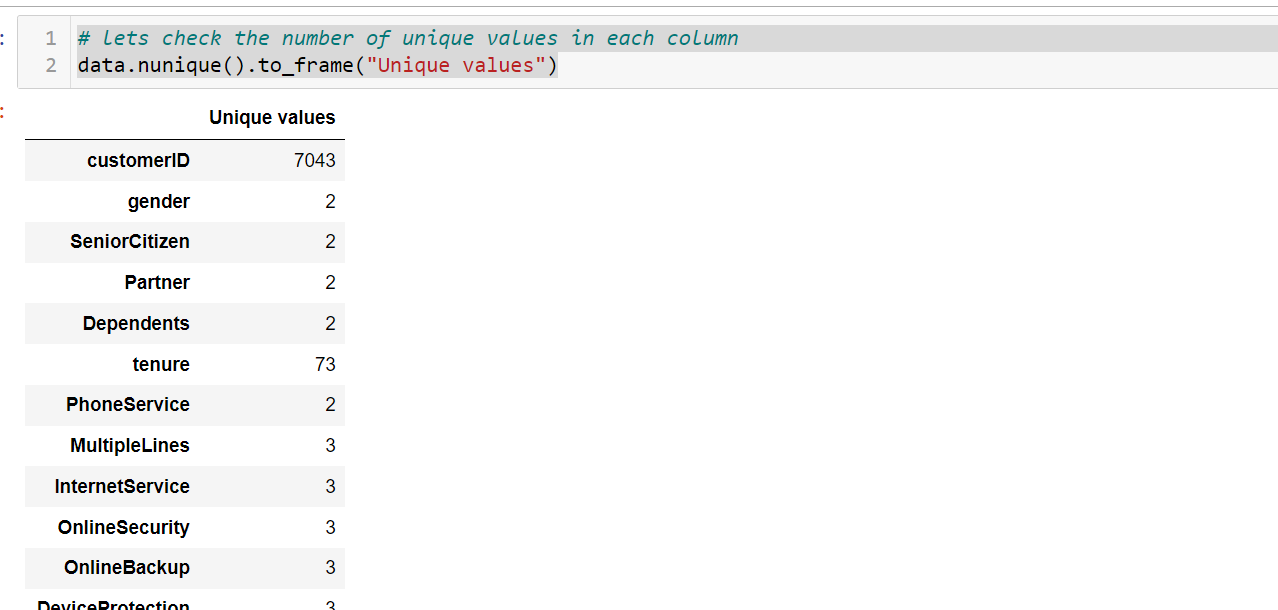
We can have a basic overview about the dataset after finding out the statistical summary of its data. Here we can know about the mean, median, standard deviation min and max values of each column etc.



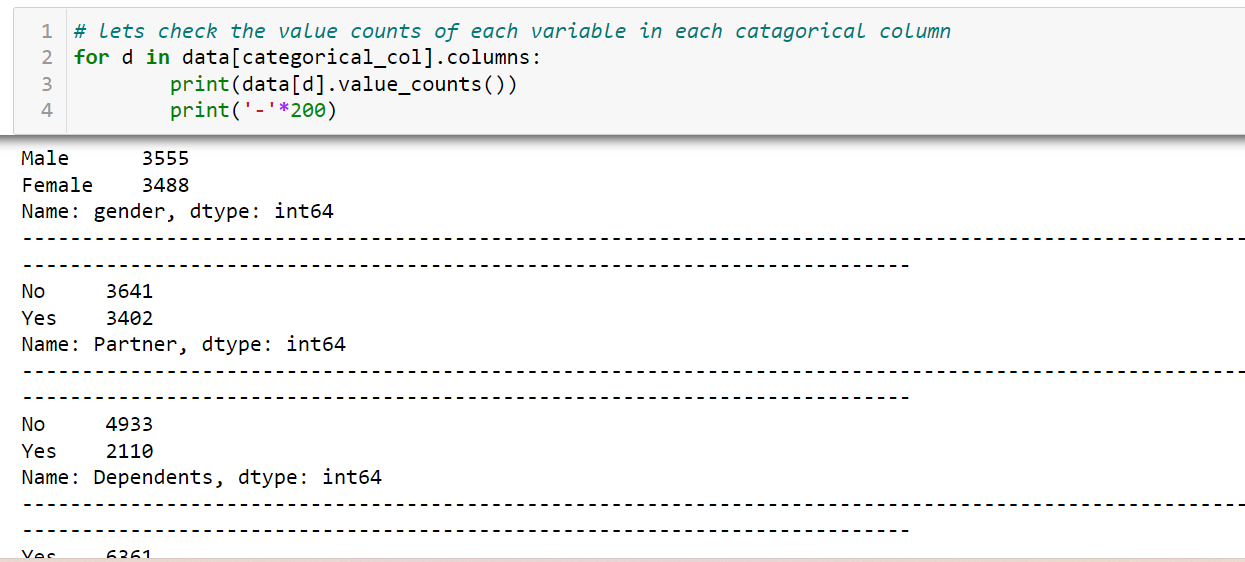
* Senior citizen is a categorical column. So its min value is zero and max value is one.
* The difference between the mean and standard deviation in tenure does not appear to be good. It is may be because of the variation in Tenure column. Also the min value of tenure is zero. Apart from that the difference between all the quantiles in Tenure appears to be good.
* The difference between mean and standard deviation in MonthlyCharges does not appears to be good enough standard deviation is nearly 50% of the mean. And also the difference between all the quantiles is not good enough.
* Also In some of the columns the mean value is greater than the median value.
* rest all the columns have object data so only three columns are printed after using describe() method.

**Plotting Count plots**

Then I have found out the number of unique values in each column to find out more about the categorical data present in the dataset.

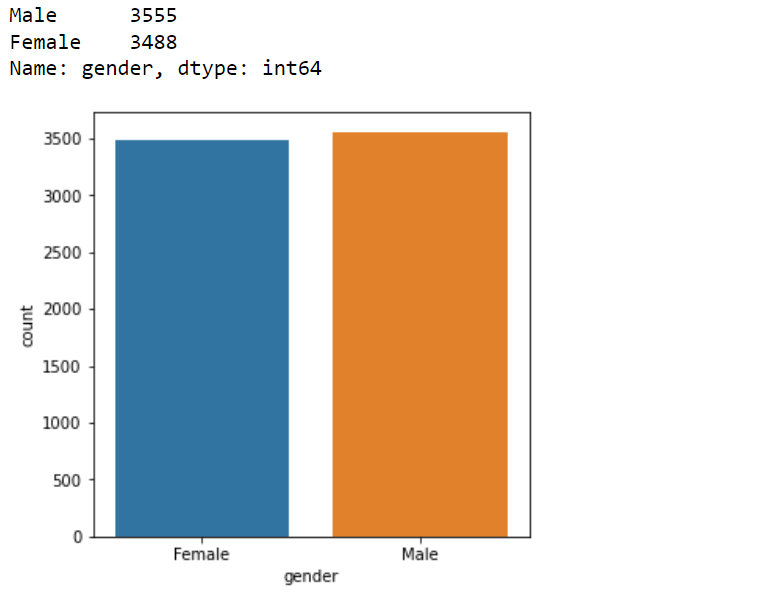


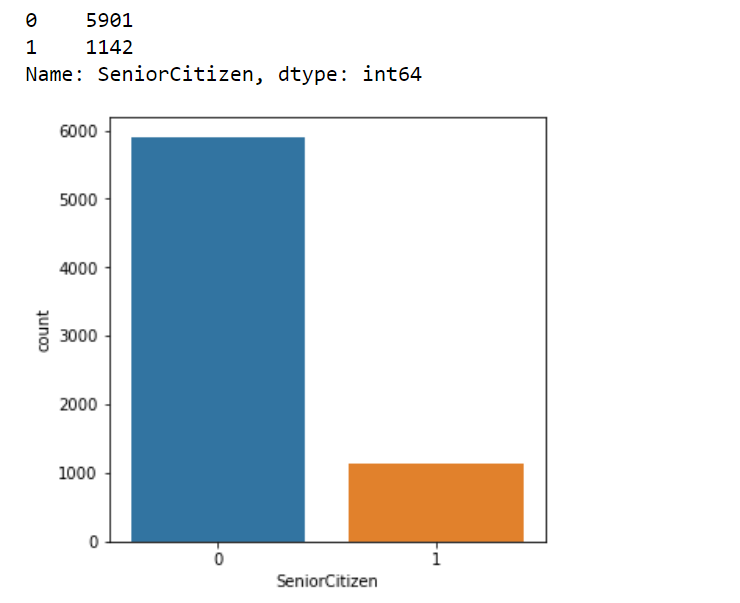
After finding out the unique values in each column, I have searched for the value counts of each unique value for each categorical column.

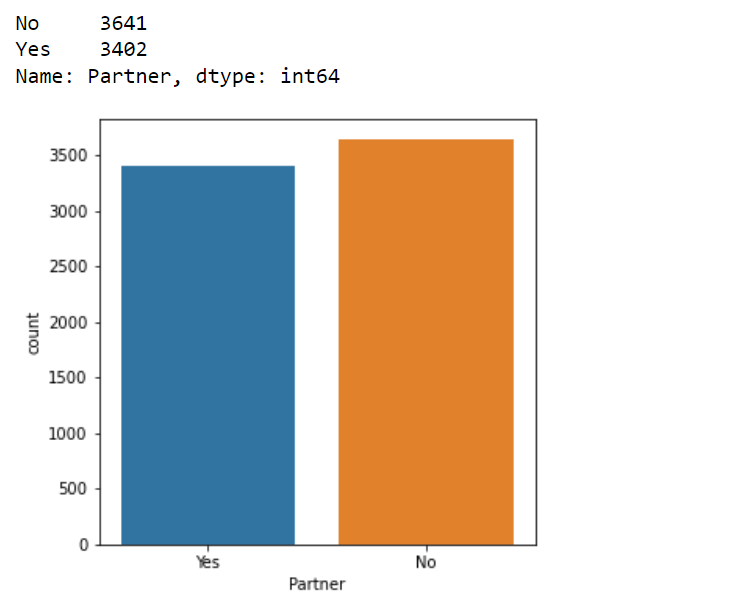


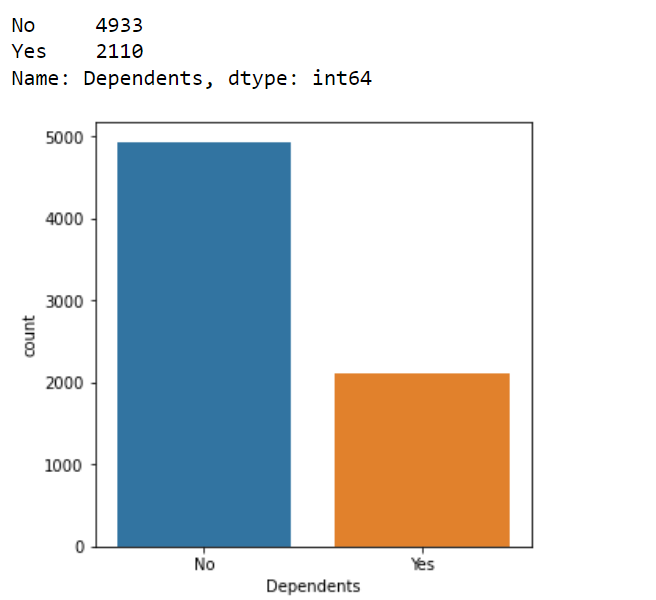
After that I have plotted the value counts using the count plot.

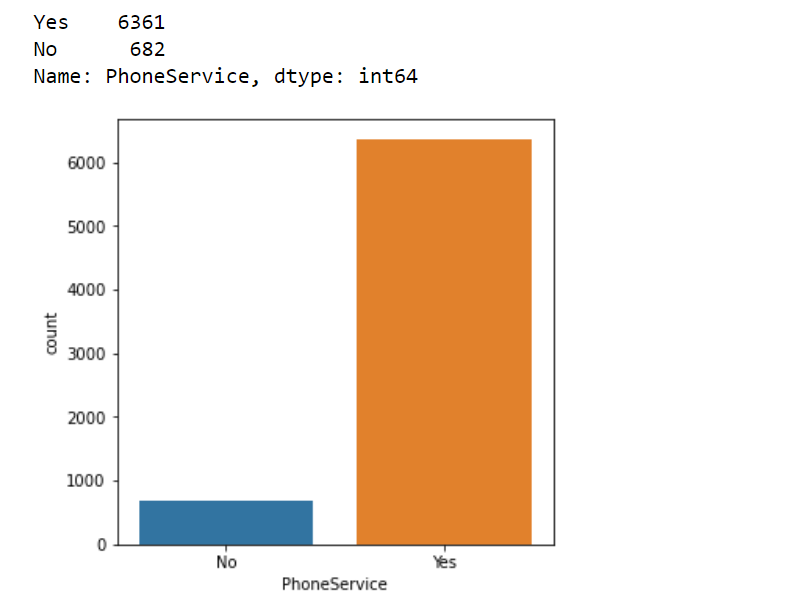


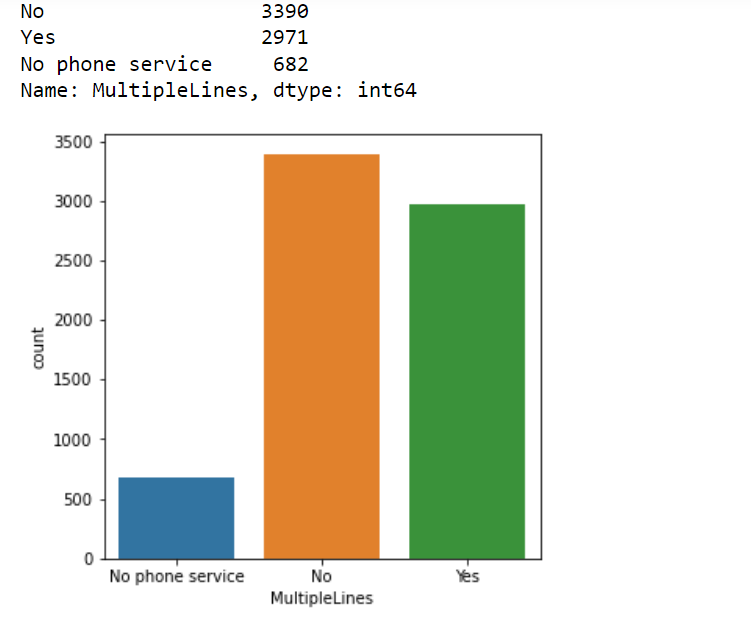


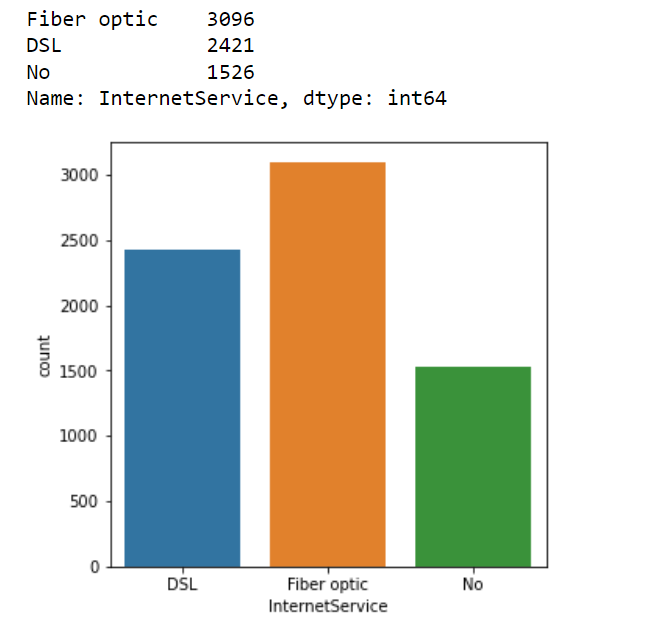


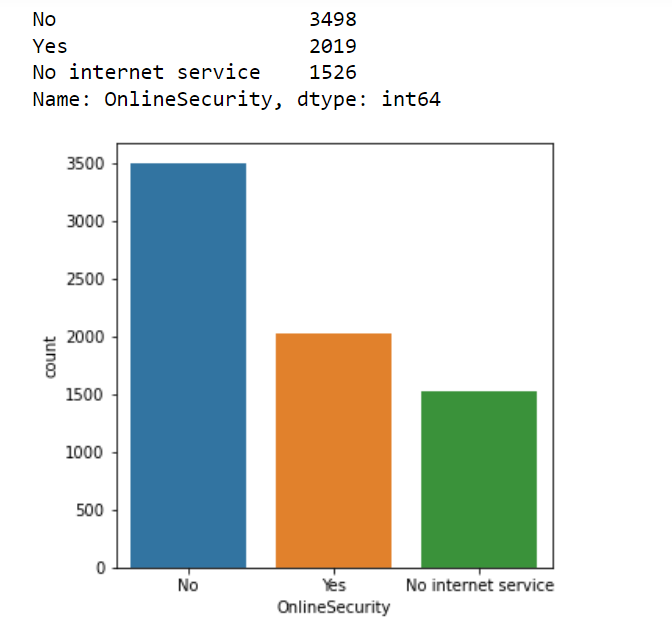


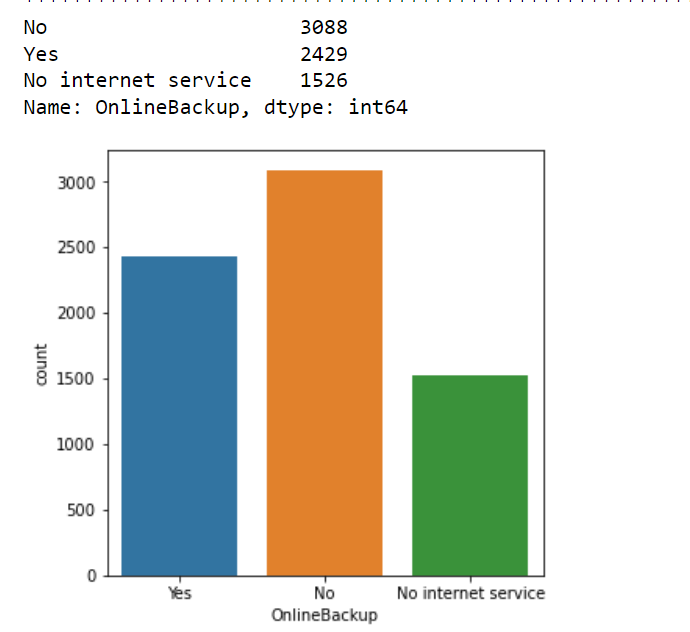


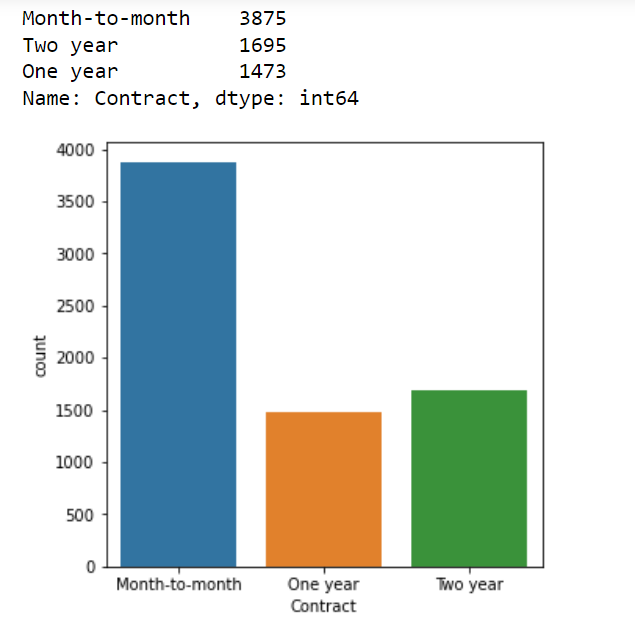


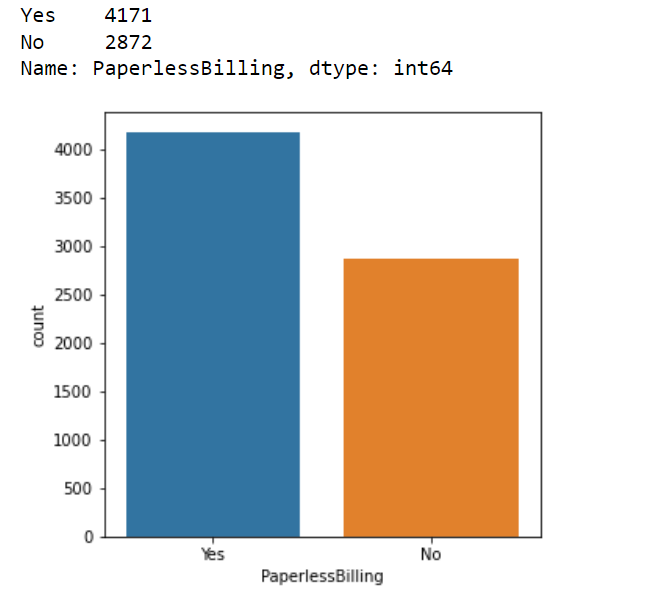








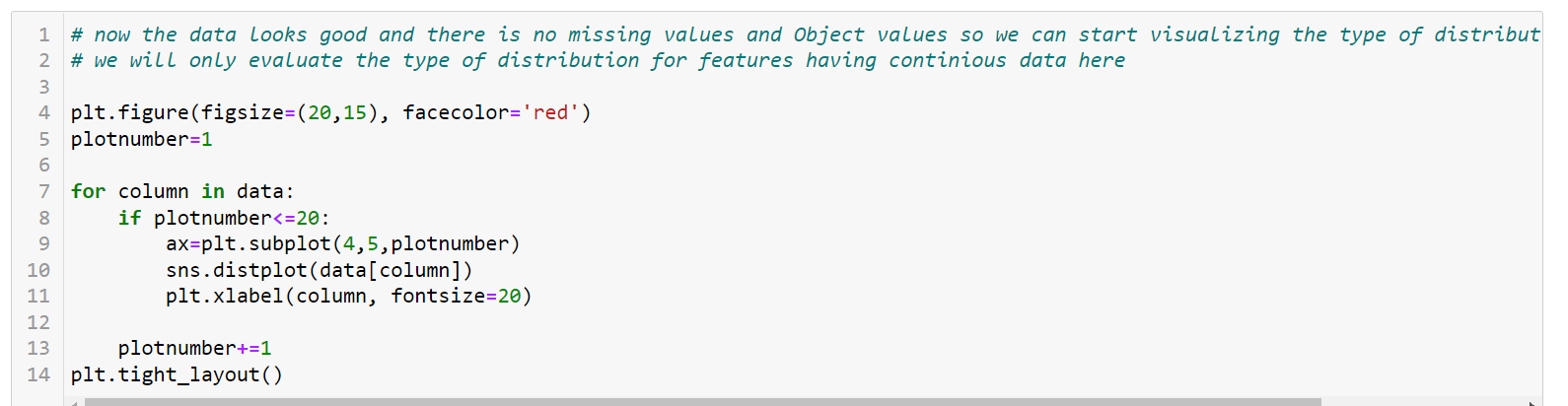


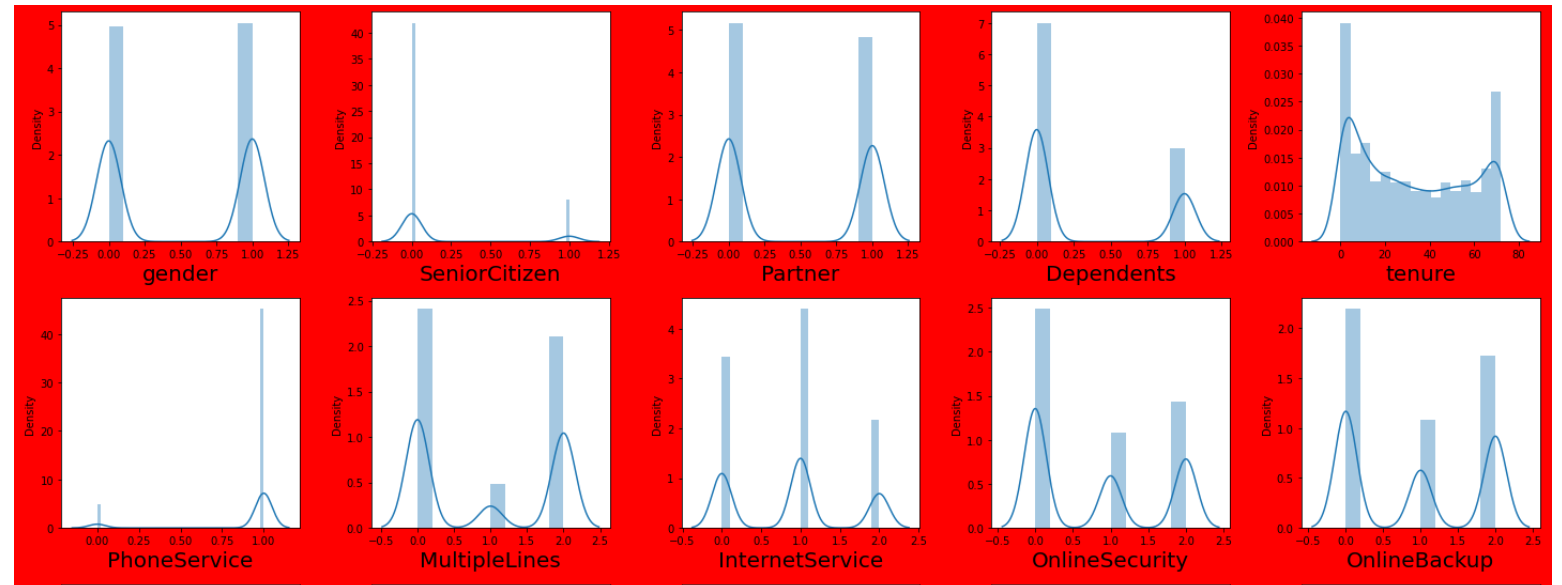


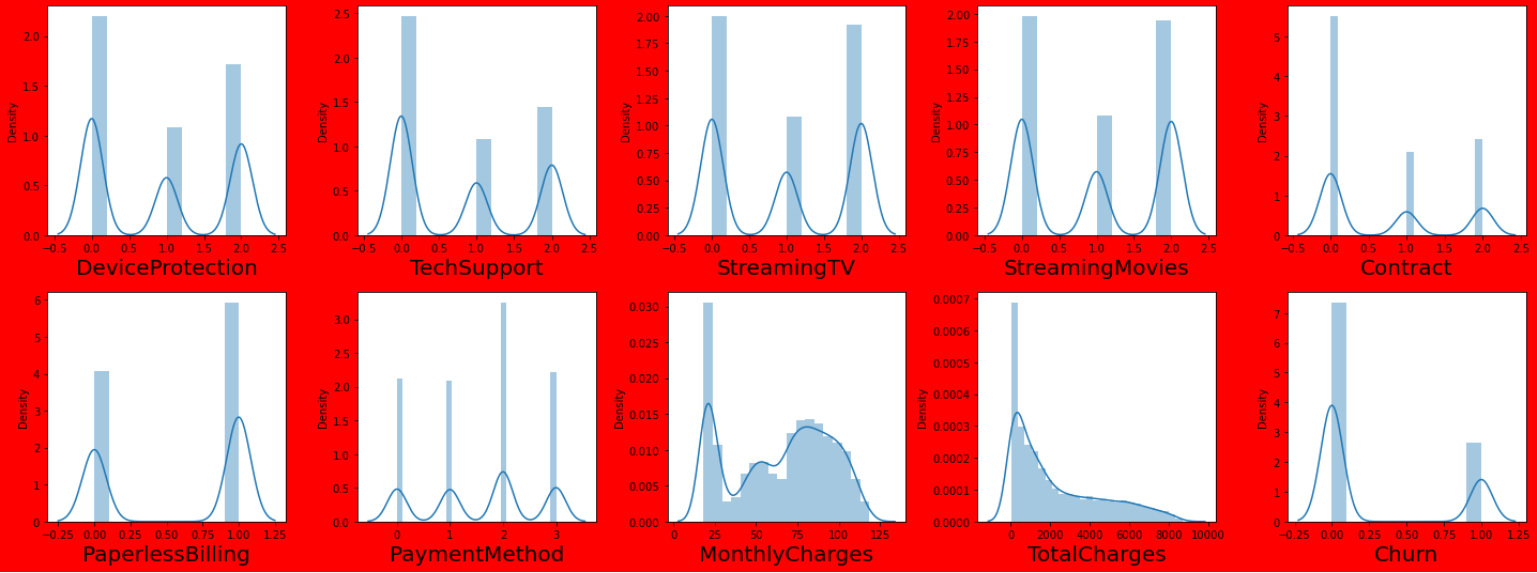
* The Telecom company has more males customers than female.
* Most of the telecom companys customers were not senior citezens.
* Most of the customers of telecom company does not have dependents.
* Most of the customers have opted for phoneservice.
* Most people has not taken multiple lines.
* Most people has taken fibreoptics as internet service.
* Most peopl have not opted for device protection, online backup and internet security.
* Most of the people have opted for monthtomonth contact and paperless billings.

**Checking the distribution of all the features**

So let’s find out the type of distribution of each feature.



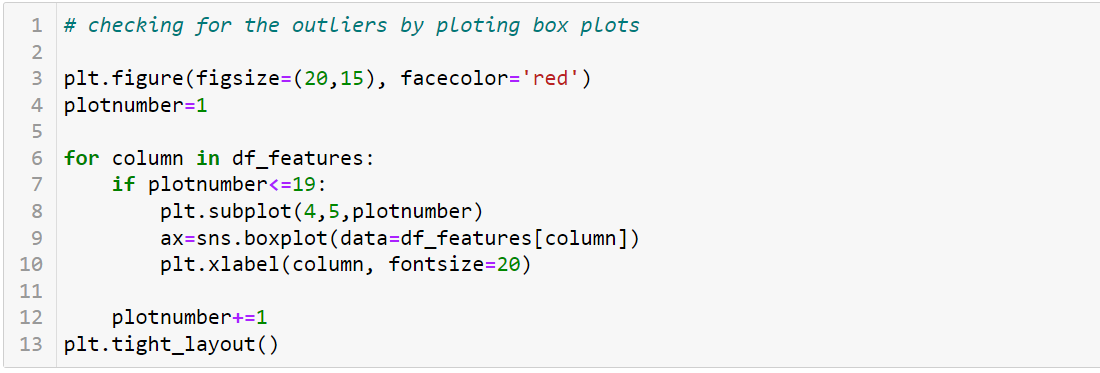


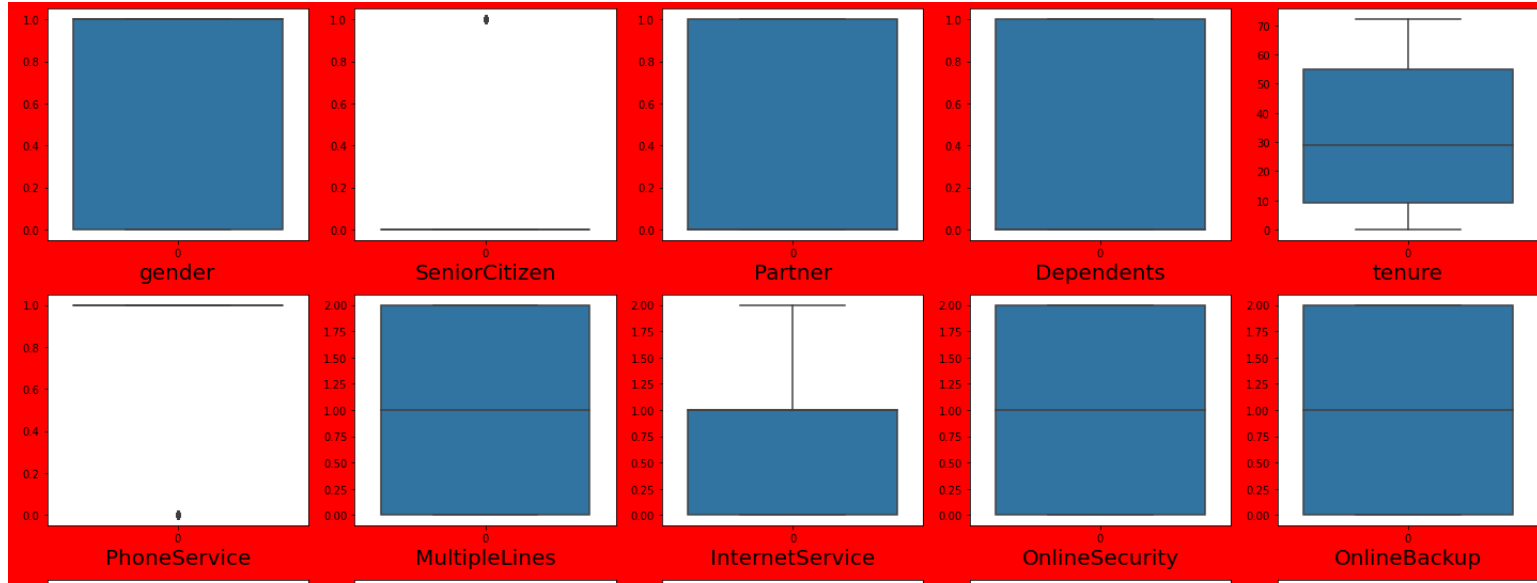


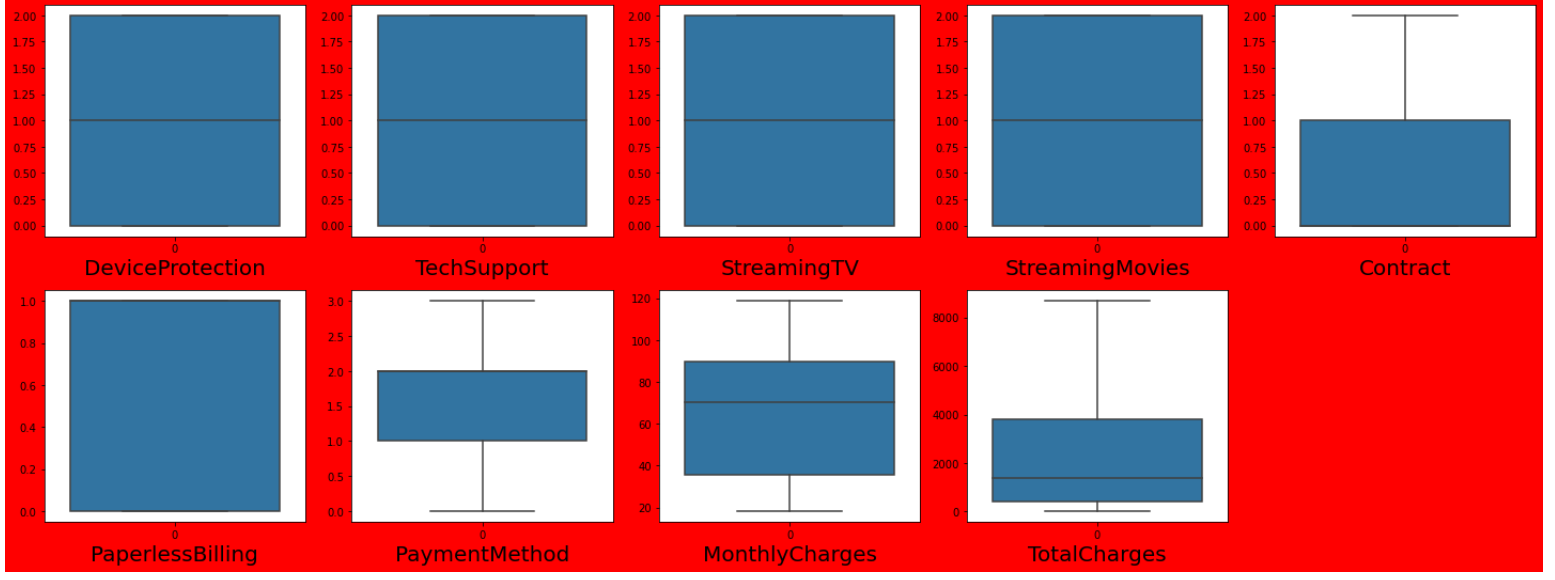
* Here Tenure, monthly charges and total charges are columns having continiuos data, so we will only consider their respective plots for analysis.
* In tenure and Monthly charges the curve does not appears to be having normal distribution (bell shaped Curve).
* In total charges the curve appears to be normal but it is observed to be having some skewness in the right side.
* now let’s check whether they have outliers in the distribution by plotting box plots.

**Removing outliers**

Let’s check the outliers by plotting Box Plots.



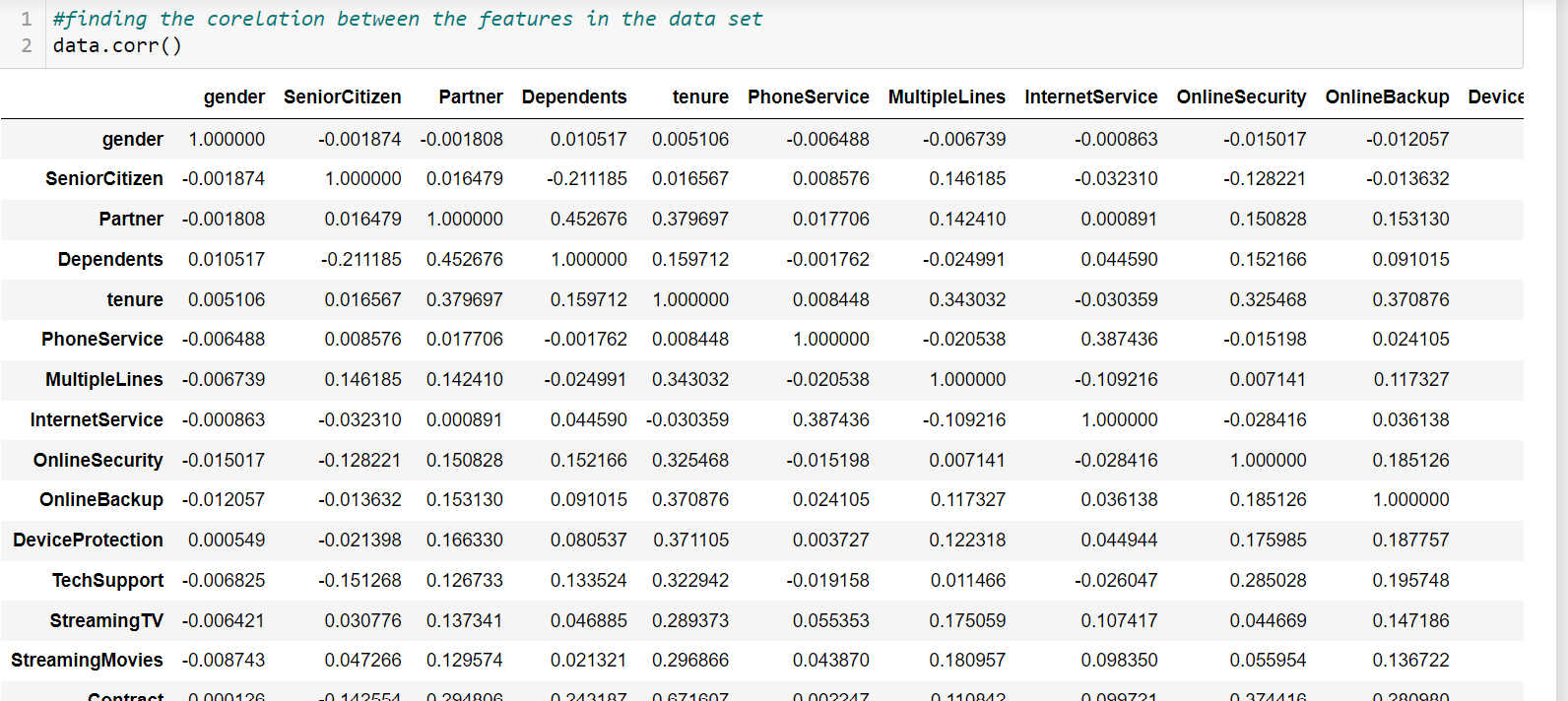


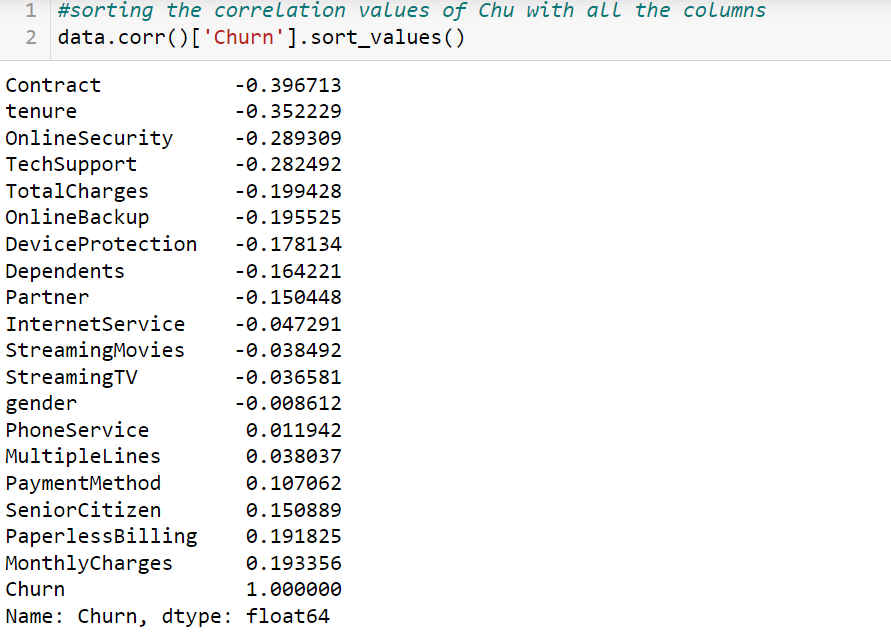


* Senior citizen and Phone service columns gave outliers but they are categorical columns so we will leave the outliers as it is.

**Checking relationship between the features (multicolinearity)**

Now let’s check if the features have co linearity among them. We need to remove multicolinearity for building a good machine learning model.



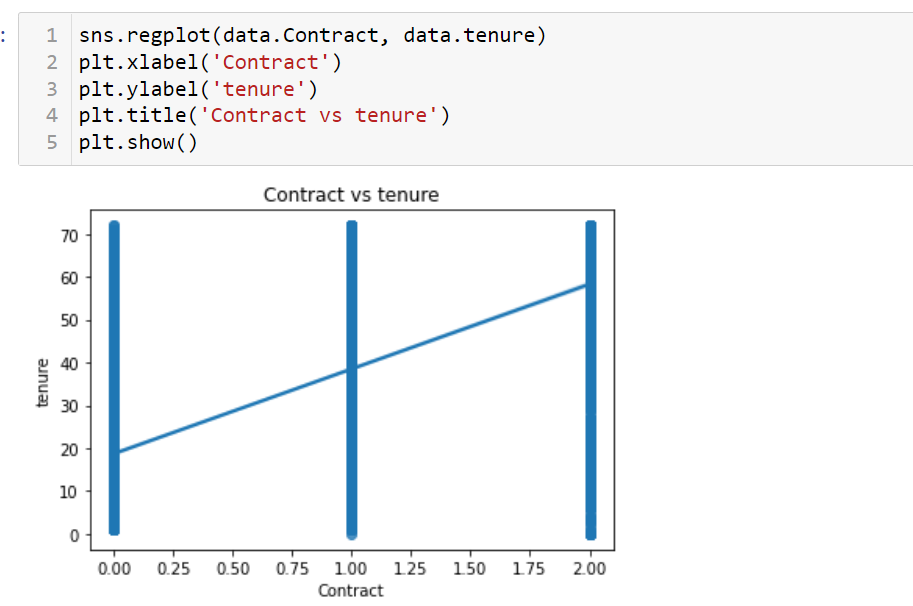


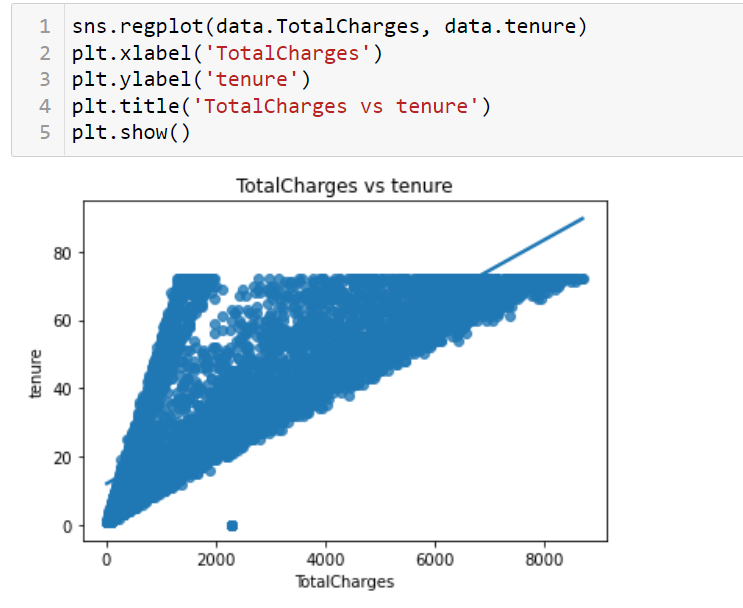
* We can observe that all the no feature is strongly correlated with the label(Churn). To get a more clear view lets plot the heatmap.

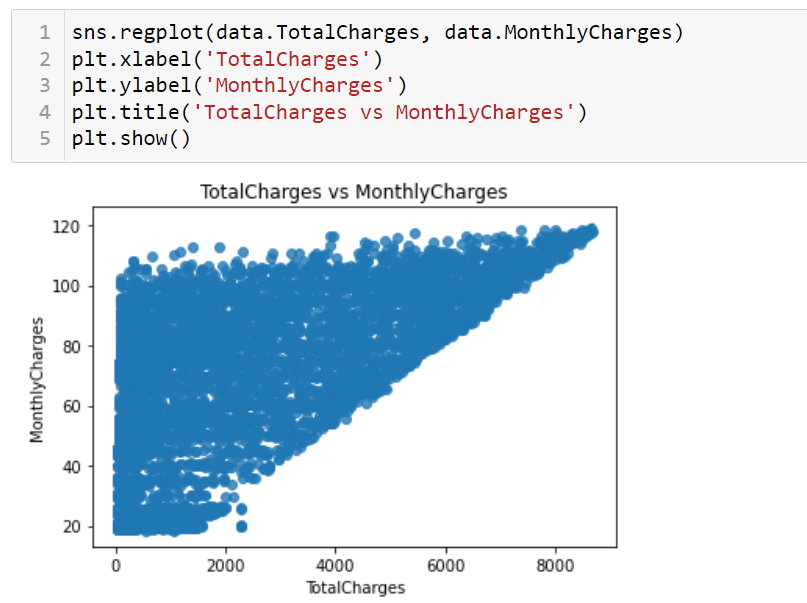


* Contract is highly correlated with tenure.
* Total\_charges is highly correlated with tenure.
* Monthly charges and total charges are highly correlated with totalcharges.

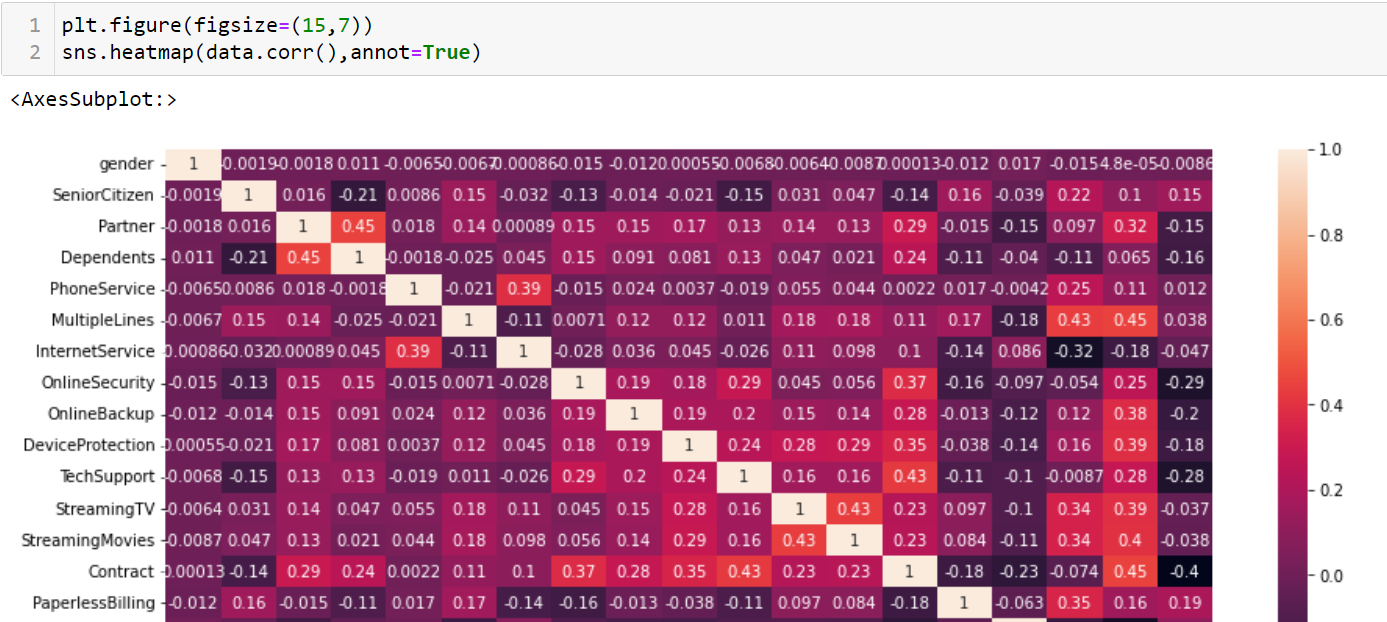
Now let’s plot scatter plots among the above features and analyze the trend.







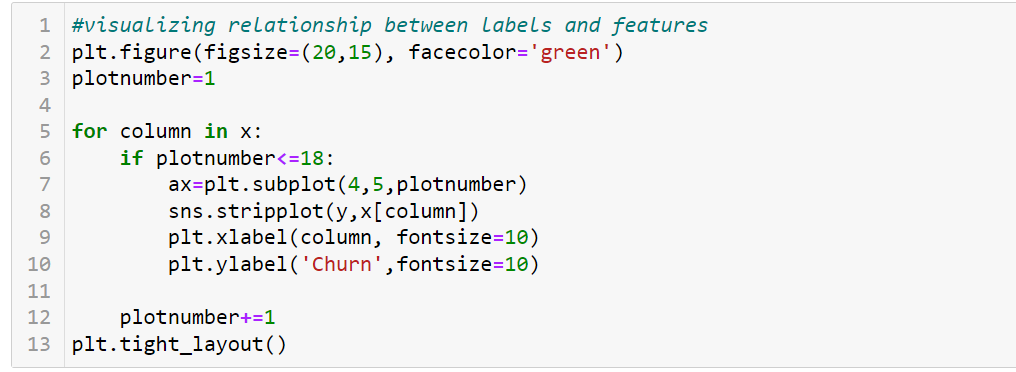
I have deleted the tenure column from the dataset as it is showing multicolinearity with other features. Now let’s check if multicolinearity issue is resolved after deleting the tenure column.

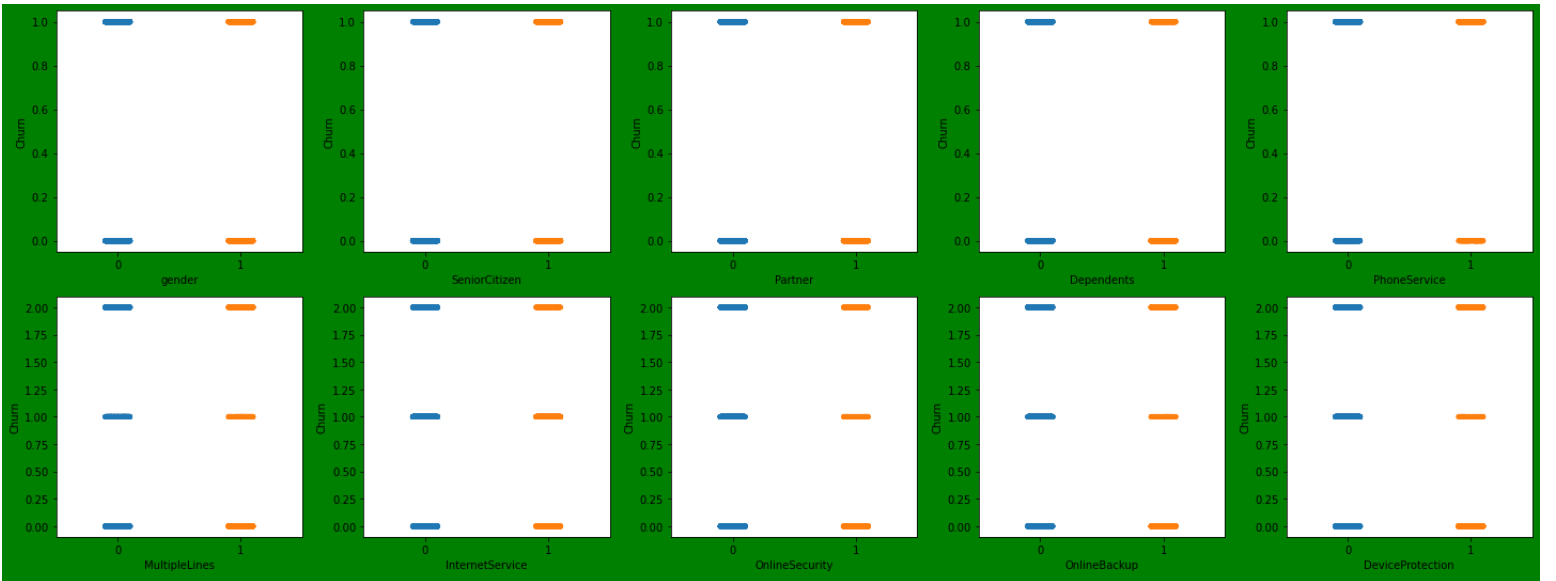


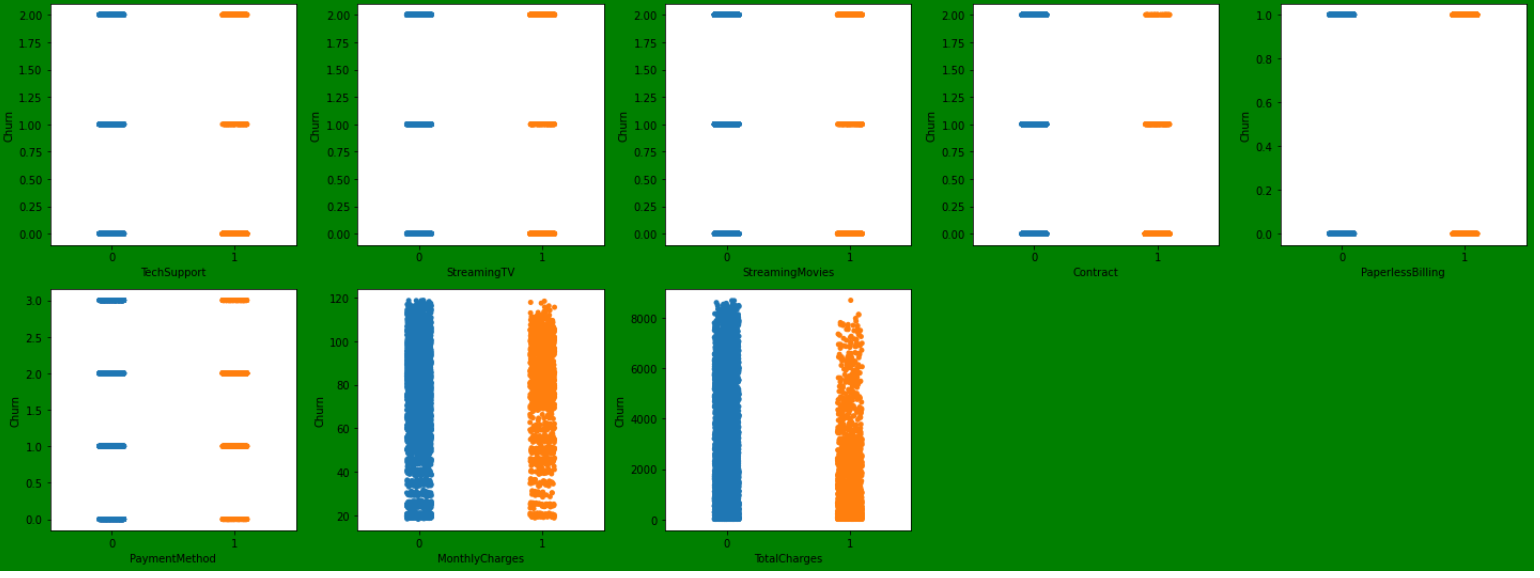
* We can clearly see that the multicolinearity issue has been resolved after deleting the tenure column.

**Checking relationship between the features and labels**

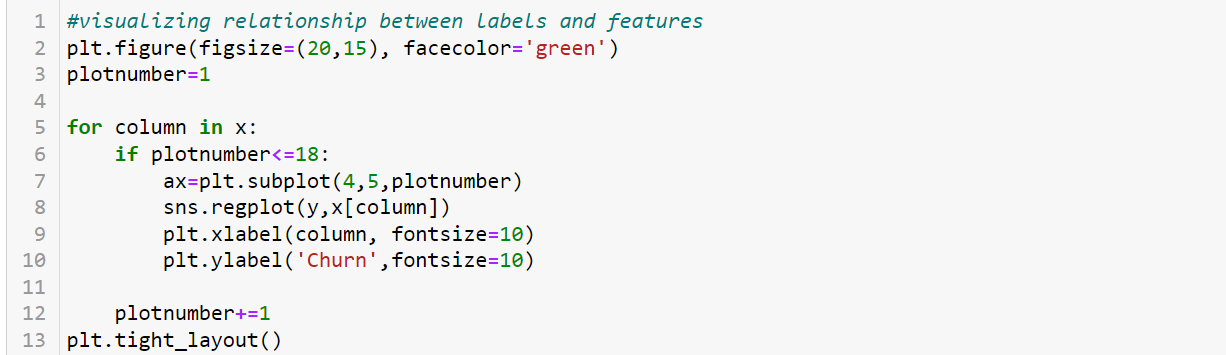
Let’s check the relationship between the features and the labels by using stripplots. This will tell whether the features have positive, negative or no impact on the label. This will help us to further in building different machine learning models.

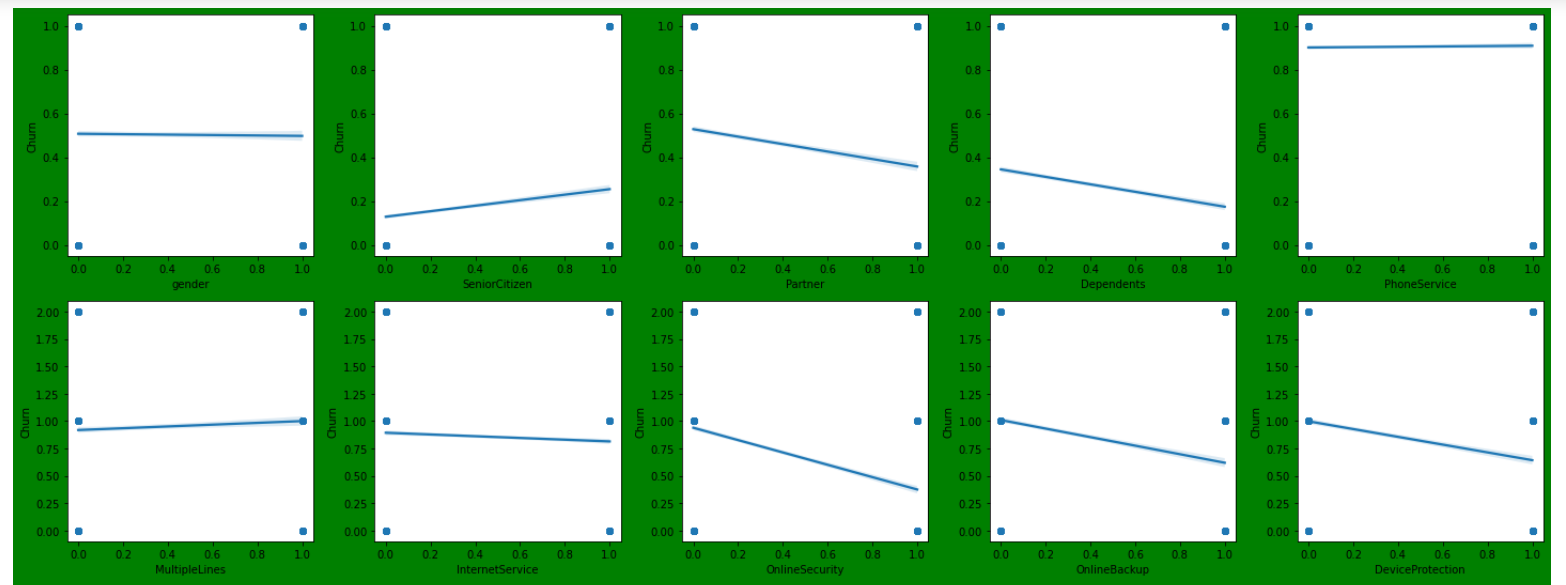


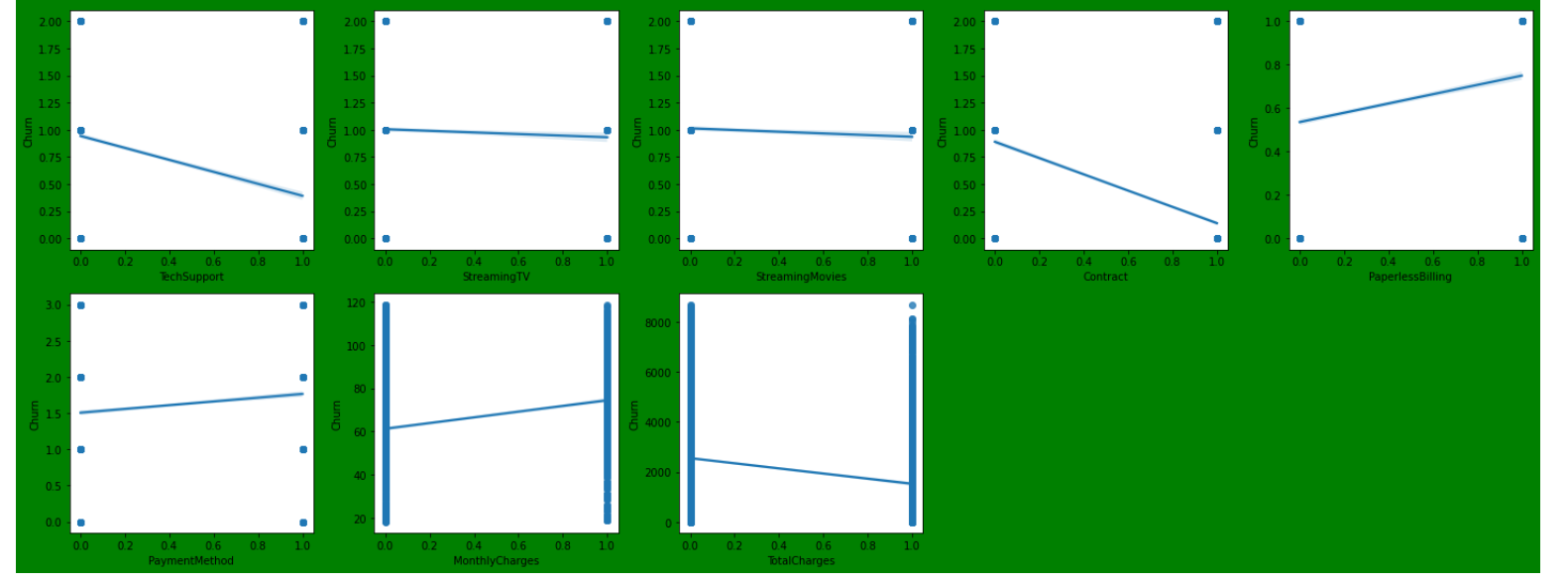




* The trend is not clear here, so let’s plot the regplot for a better view of the trend in relationship between the features and the label.







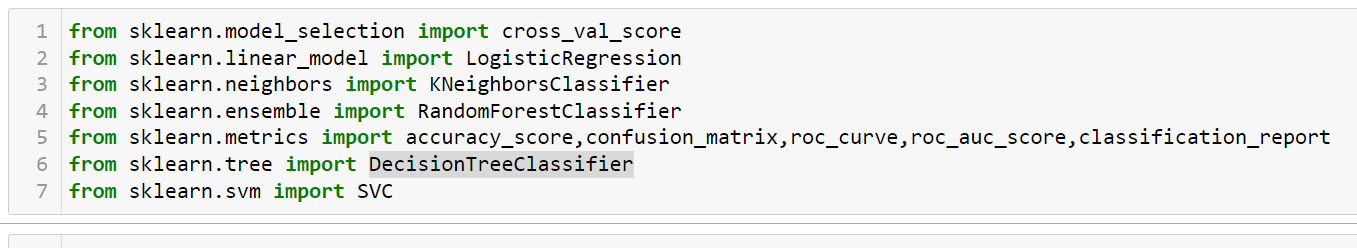
* We can observe the trend clearly from the regplots.
* Most of the features does not show a completely positive or negative relationship with the label Churn.
* Contract and techsupport shows somewhat negative relationship with the label.
* PaperlessBilling and monthly charges show somewhat positive relationship with the label.

**4.Pre-processing pipeline**

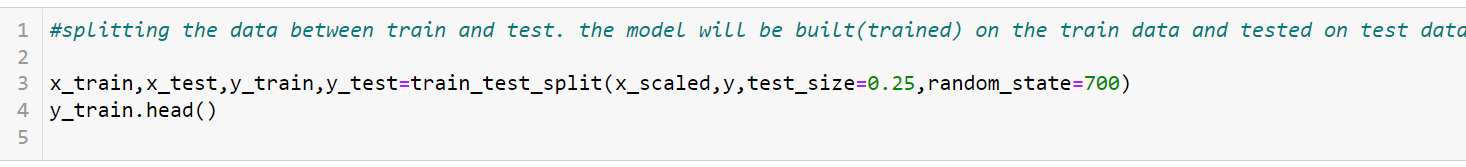
* We have performed many steps during pre-processing of the data like Data cleaning, Data reduction and encoding techniques. Let's discuss each of the steps performed during the pre-processing of the dataset in detail.
* We have drooped the columns which we thought would not impact the prediction of label (churn) like Customer\_id. We have also deleted the columns which show multicolinearity like tenure.
* Here we have observed that no NANs (null values) are present in the dataset provided so we have not used any imputing techniques for treating the null values.
* During removing the outliers we observed that the outliers are only present in the columns having categorical data, So we decided not to remove them. Hence Senior citizen and Phone service column outliers were not removed.
* After that we have checked whether any column has Object(string) datatypes or not, we have split the whole dataset into numerical and categorical columns and then we have encoded the categorical columns by using suitable encoding techniques. We have used LabelEncoder() encoding technique to encode the object values in the current study.
* After that I have used the Variance Inflation Factor(VIF) for all the features to find out whether multicolinearity still exists. All the features have vif scores less than 10,So we don’t need to worry about multicolinearity now.
* After that I have scaled the whole dataset to a single unit with the help of standardscalar() to help our model to learn better.
* The dataset has only 23 columns and from that we have already deleted some columns, so I have not used PCA to reduce the number of columns here.

**5.Building Machine Learning Models**

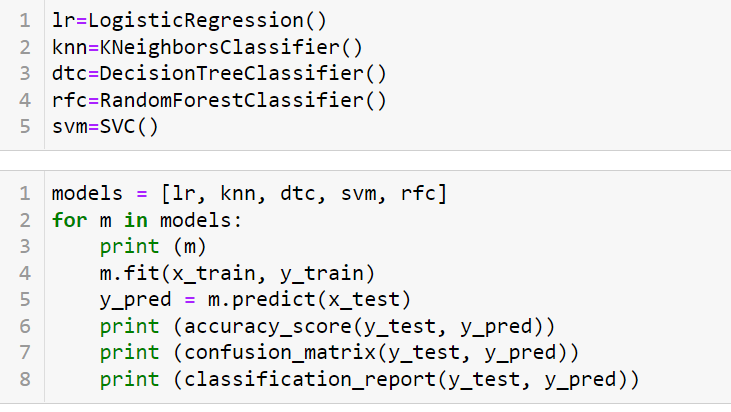
Now as the label has categorical data so this is a Classification problem. We will now be building some classification algorithms for this problem and we will find out the best machine learning model among them. We have imported some classification machine learning algorithms like LogisticRegression, KNeighborsClassifier, RandomForestClassifier, DecisionTreeClassifier and SVC. Then we have imported each one of these model in Jupyter Notebook.



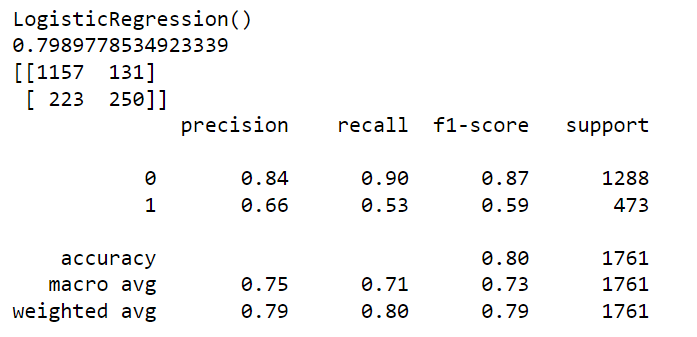
After that I have used train\_test\_split where the model would be trained on the training data and tested on the testing data. I have taken the test\_size as 0.25 which means the model would trained on 75% of data present in the dataset and would be tested on the remaining 25% data. We have taken the random\_state as 700 and used the scaled data we got from StandardScalar() for training the model.



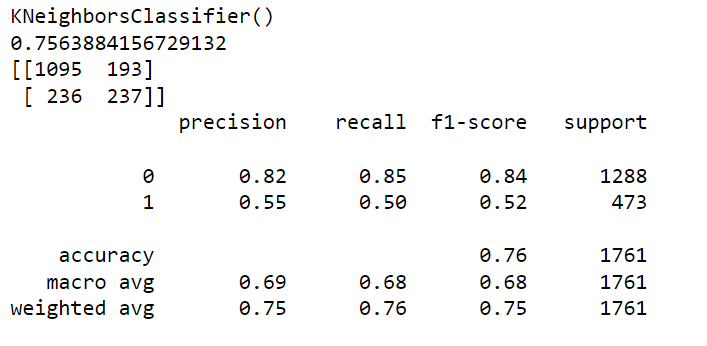
We are choosing the best model based on their respective accuracy scores and Cross validation scores. Along with that we are also printing their respective Confusion matrix and classification reports.



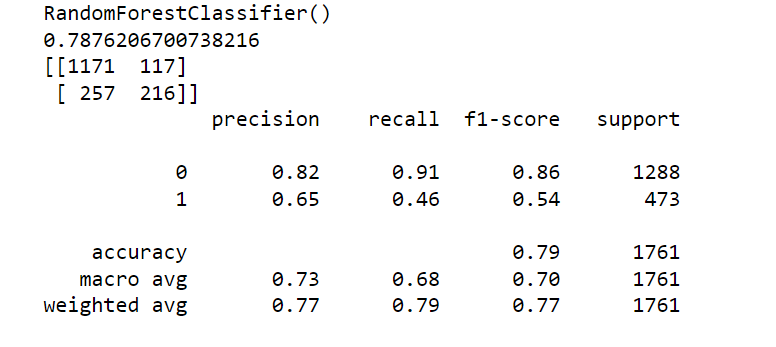
**LogisticRegression()**



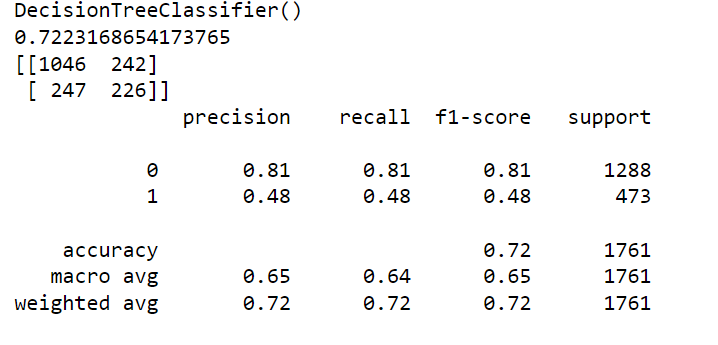
**KNeighborsClassifier()**



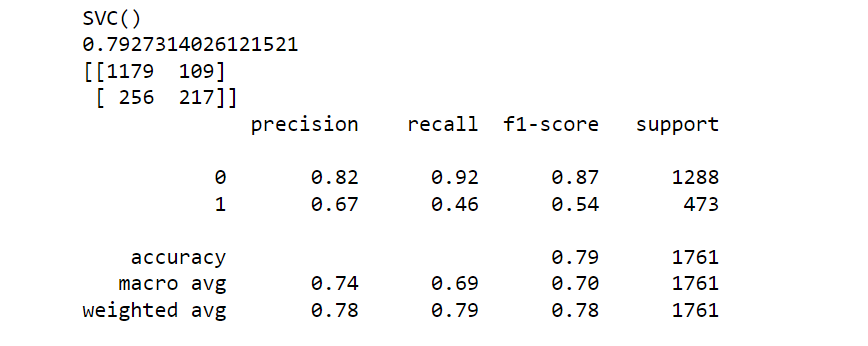
**RandomForestClassifier()**



**DecisionTreeClassifier()**



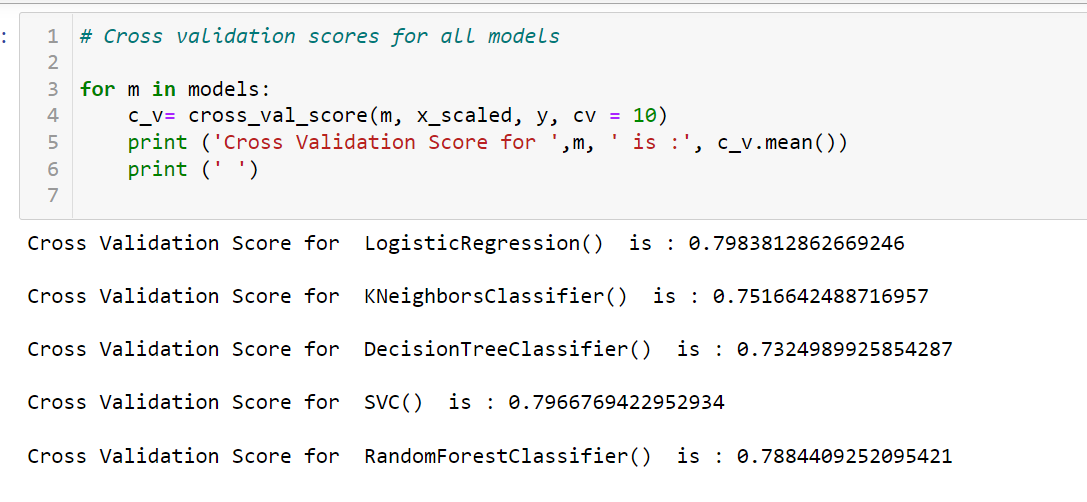
**SVC()**



We have observed that the LogisticRegression() model gives the best accuracy score with 80% accuracy.

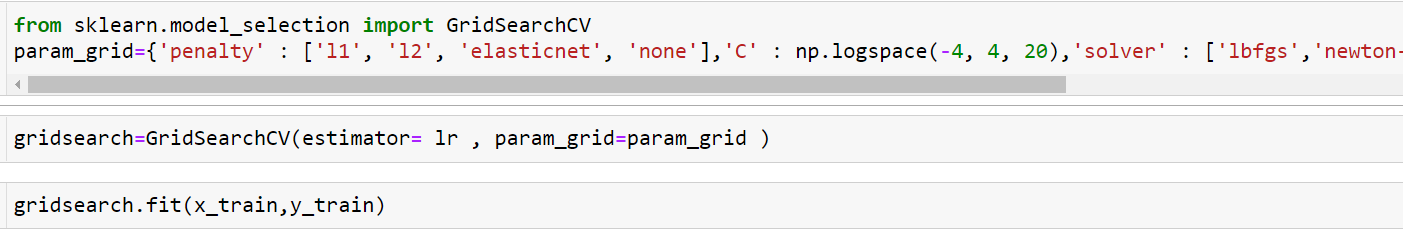
After that we have used the cross-validation technique because the model may have over fitting issue thus in order to be 100 percent sure that the model is not over fitting we have used cross validation method. I have used the K-fold cross-Validation technique here.

In this Study I have taken the cv value as 10 for the cross validation.

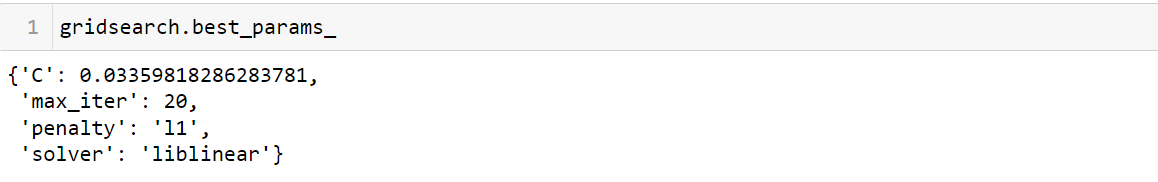


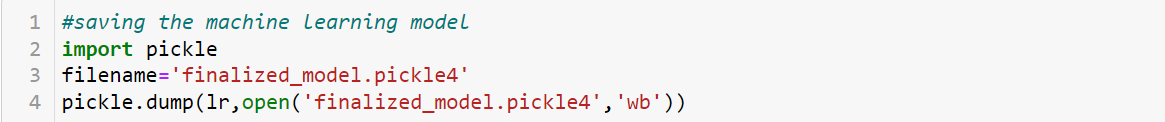
After finding the cross validation values for each model we observed that LogisticRegression() model has the best cross-validation value.

We have concluded that the LogisticRegression() is the best model based on the accuracy and Cross validation scores of all the models. So now let’s use LogisticRegression() for further evaluation and let’s check whether we can improve the accuracy of the LogisticRegression()model by using Hyperparameter tuning using GridSearchCV.



The best parameters for LogisticRegression() was found to be



After that I have saved the model using pickle for future predictions.

**6) Concluding Remarks**

* The main goal of our project is to predict the Customer churn of a telecom company.
* In this study we have used different machine learning algorithms, visualization, encoding and data cleaning techniques.
* We have done the complete analysis of the data using Exploratory Data Analysis(EDA), Checking the distribution of all the features, Plotting count plots for different features ,checking correlation among the features and correlation between the features and label , checking presence of outliers, checking for missing values.
* We have also used StandardScalar() for Standardization of the data.
* During the model building phase I used classification algorithm and different evaluation matrix to prepare the models and I found out Logistic regression model as the best model.
* So now we can make predictions for Customer churn for the telecom company using our classification model.

Author-Sankalp Mahapatra

DataTrained, Batch-1842