# **Customer Churn Analysis – Blog**

Submitted by

***Alfinson John***

Batch No **1840**

I have chosen Customer Churn as my favorite analysis project, in this article, I’ll be going through the process of building a machine learning models to predict the number of customers who don’t purchase products or services and what are the major issues behind these and by customer retention to get back the customers. It will take through each step-in detail and helps you to understand the whole machine learning model building process.

A picture containing icon

Description automatically generated

# **INTRODUCTION**

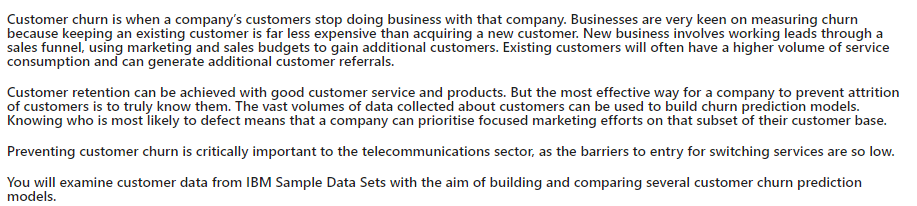
# Customer Churn is one of the most important and challenging problems for business such as credit card companies, cable services provide, SASS and telecommunications companies worldwide even though it is the most fun to look at, customer churn metrics can help business improve customer retention.

# Now a days customer churn problem in every industry because of heavy competitors but in telecom industries it become very competitive because of the way price reduced plan or coming with the less tariffs for these I had predicted on customer churn with the output accuracy of 87%, with the help of libraries which is used listed below.

# **LIBRARIES USED**

* Python
* Seaborn
* Numpy
* Matplot lib
* DateTime

# **Problem Statement**

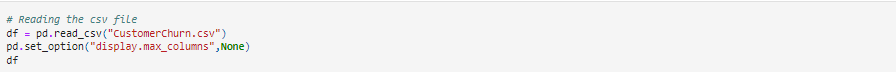


Preventing customer churn is critically important to the telecommunication sector, as the barriers to entry for switching are so low.

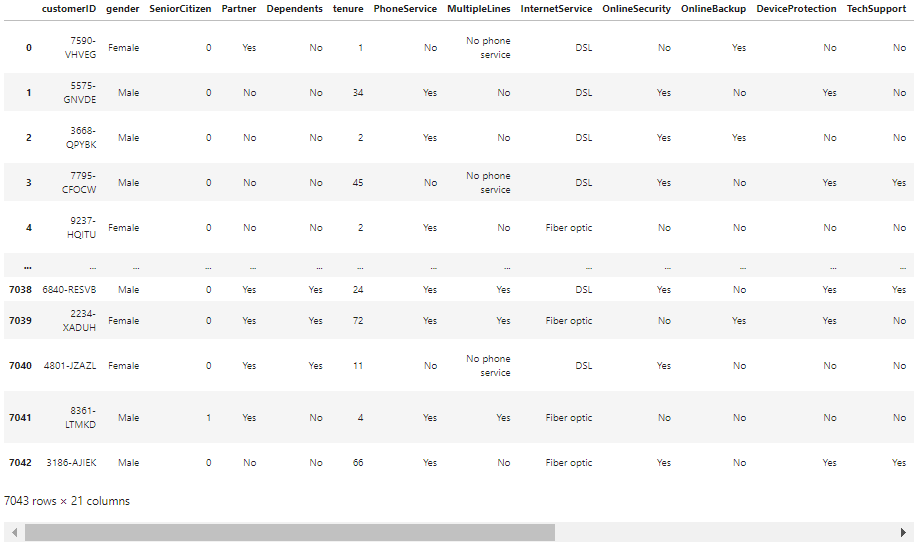
In this project, we are provided data from IBM datasets with the aim of building and comparing several customer churn prediction models.

Let’s do it step by step first analyzing the dataset and doing exploratory data analysis, data visualization, data cleaning, pre-processing, model building, model saving, and finally predictions.

## **Data Analysis**



Now we imported the dataset from IBM dataset which is converted into csv format and then the csv file had given the data frame(df) to read the file.



Above is my dataset. By looking into the features, I can say that I have both numerical and categorical with some unnecessary entries.

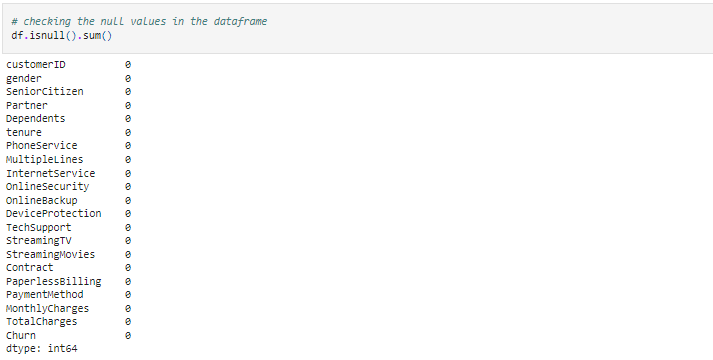
So, we must clean this data.

**Data Preparation and Cleaning**

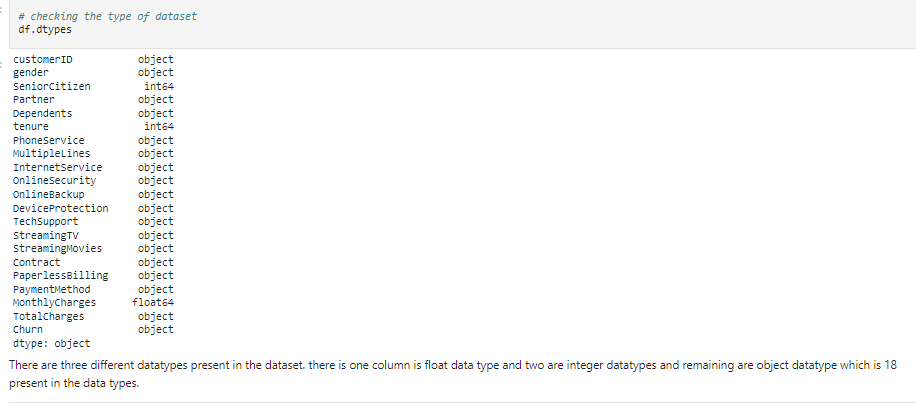
* Firstly, we must do some statistical analysis like checking shape, unique value counts, info etc…
* After reading the value counts if we find any unnecessary columns in the dataset, we can drop those columns.

**

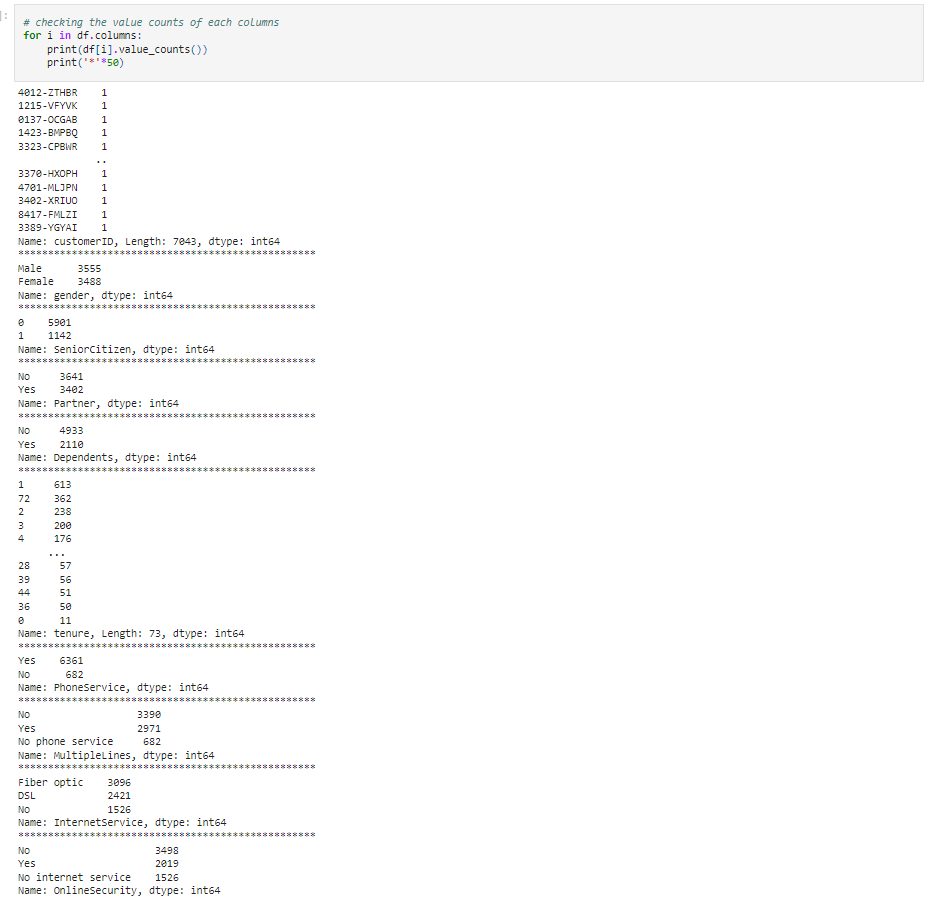
* Above we have both shape and columns contains.



* Above we have checked the null values in the data frame, and we got the result with there is no null values missing.

******

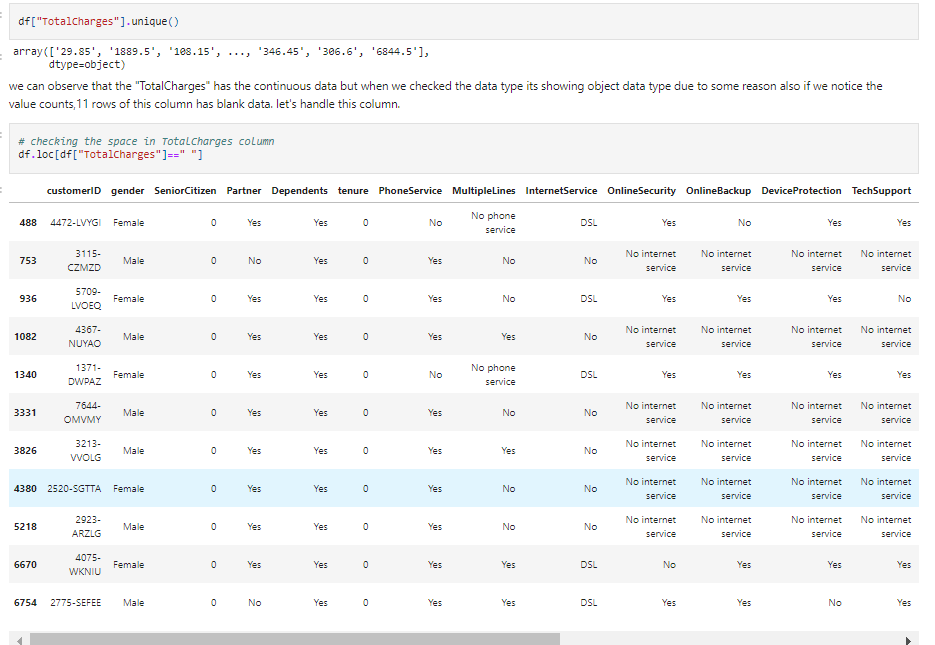
* Above there is three different data types exists in that which is float, integer and object datatype.

**

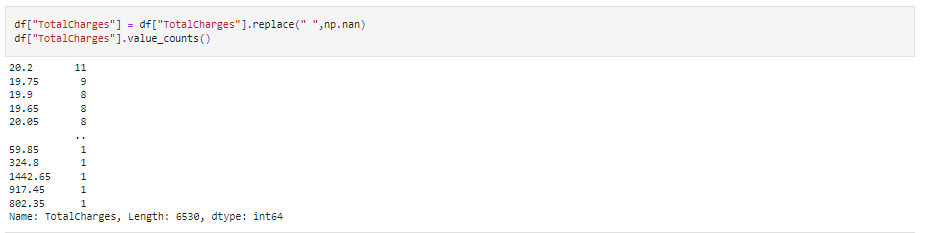
# 

* Above all are the value counts of each column present in the dataset and unique value of total charges present in the dataset. Chart, scatter chart

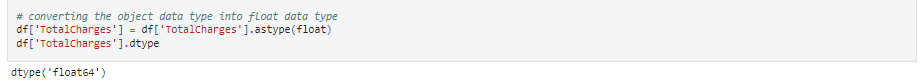
  Description automatically generated

**

* Above we check the space in the Total \_Charges.

**

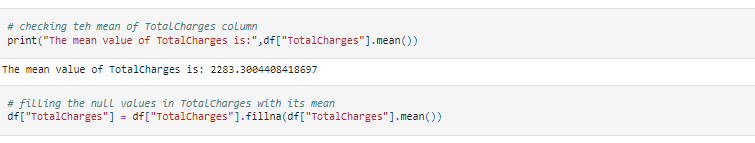
* Above we replace the Total Charges of space with nan values as Total Charges.

**

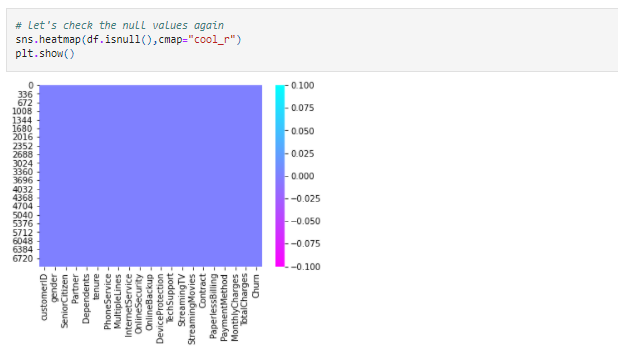
* Above we successfully converted the object data type to float data type for some reasons. Now Total charges data type is float data type.

# 

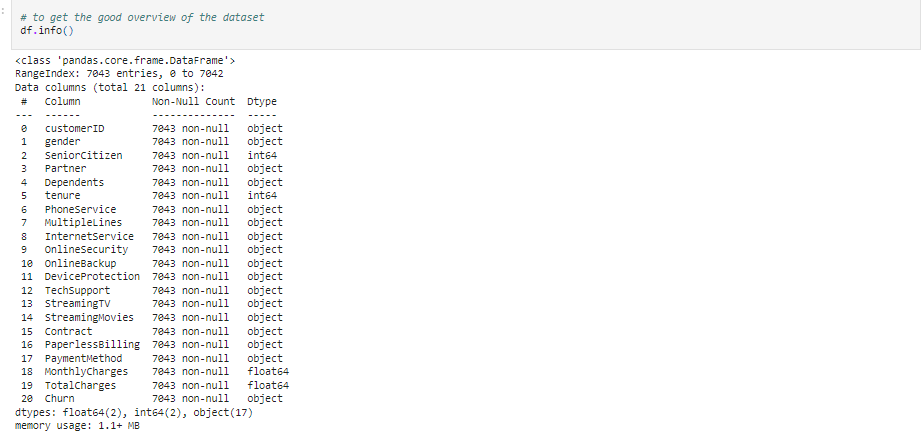
# Since we have filled the blank spaces in Total Charges column with nan values, it’s showing 11 null values in that column. replacing the nan values using mean method as the column Total Charges is continuous in nature.

**

* Here check the mean value of total charges and replaced the missing value with its mean values.

**

* There are no missing values in any column of the dataset.



* Above we will know the overview of the dataset.

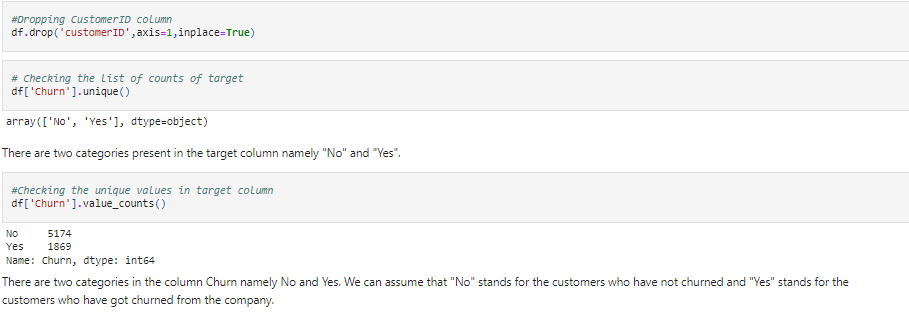
# Now we will separate numerical and categorical columns to know which column specifically belongs to



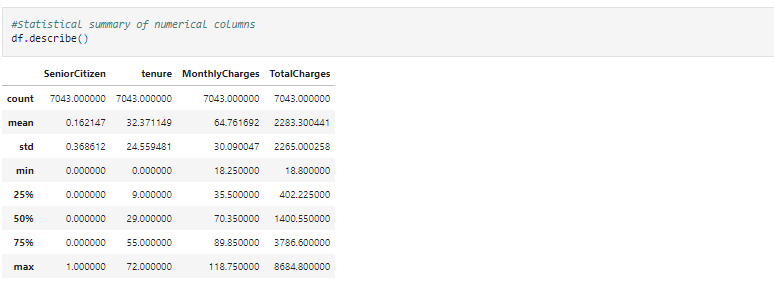
* Above it is Categorical and Numerical columns after separation.

**

**DROPPING COLUMNS**

**

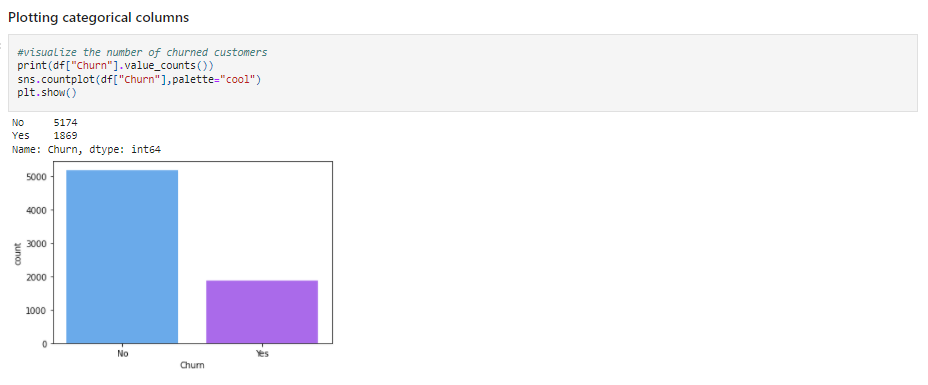
* Above we dropped the customer id and check the list of counts of target for Churn and we got to know by the yes or no manner where the “NO” stands for the customer who have not churn but whereas “YES” stand for customer who churned from the company.

**

* This gives the statistical information of the numerical columns. The summary of this dataset looks perfect since there is no negative/invalid values present.

From the above description we can observe the following things.

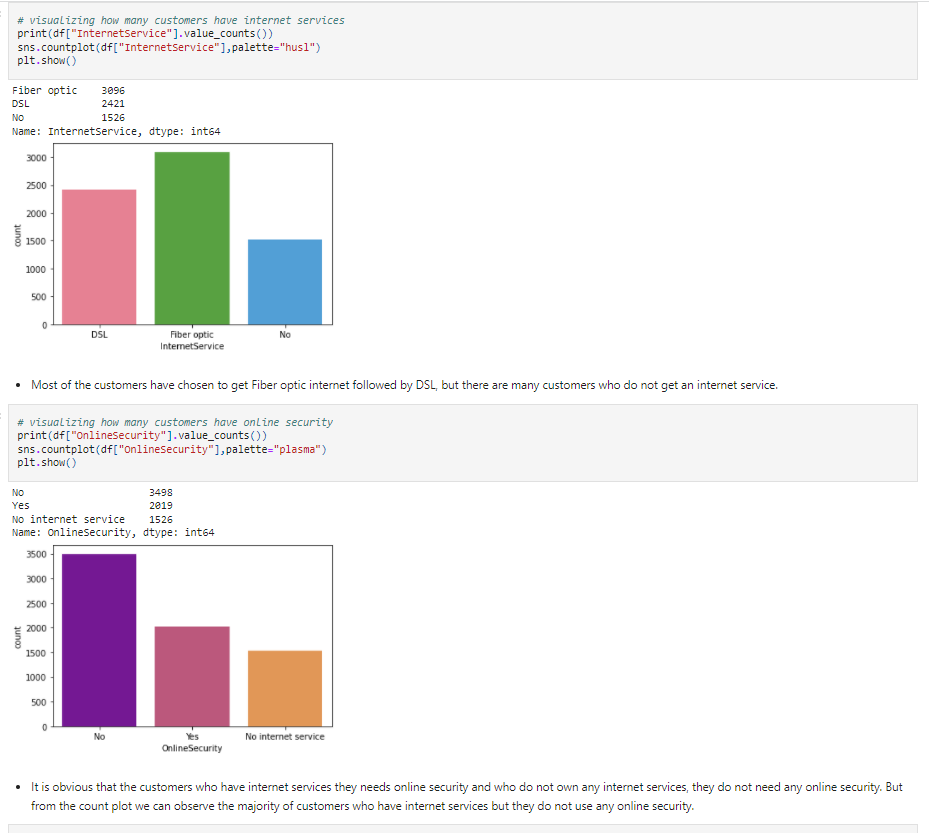
* The counts of all the 3 columns are same which means there are no missing values in the dataset.
* The mean value is greater than the median (50%) in tenure and Total charges columns which means the data is skewed to right in these columns.
* The data in the column Monthly Charges have mean value less than median that means the data is skewed to left.
* By summarizing the data, we can observe there is huge difference between 75% and max. hence there are outliers present in the data which we will remove them later using appropriate methods.
* We can also notice the Standard deviation, min,25% percentile values from this describe method.

**

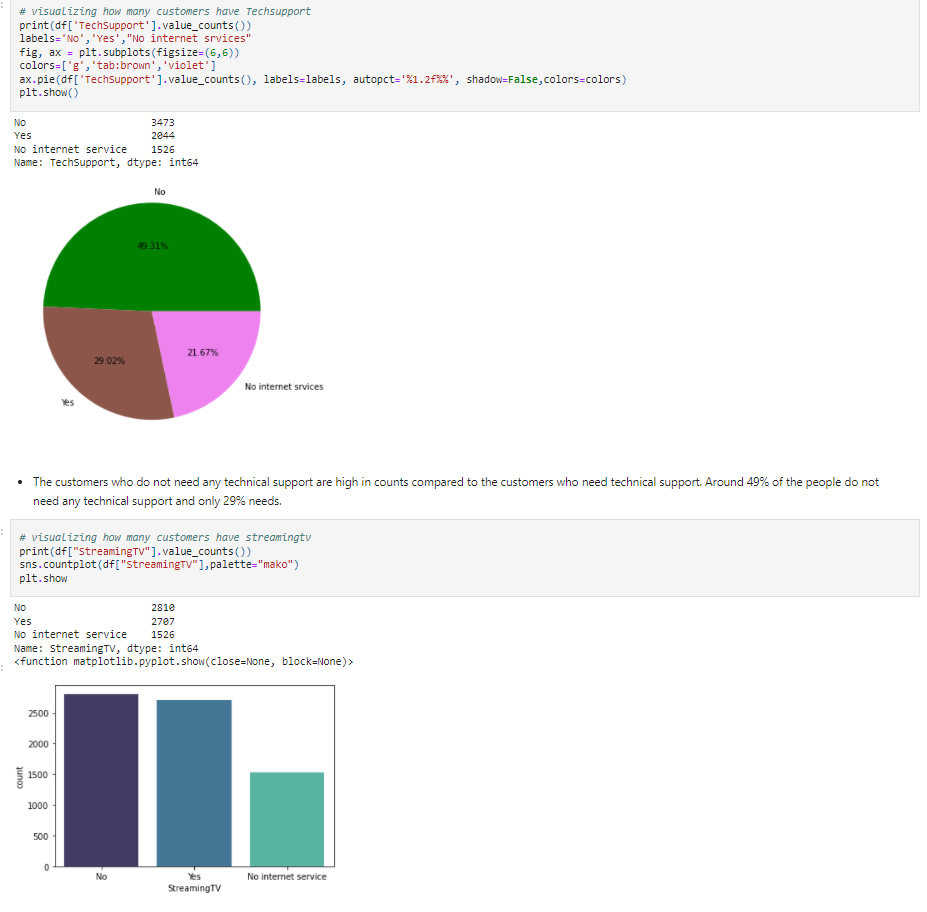
**

**

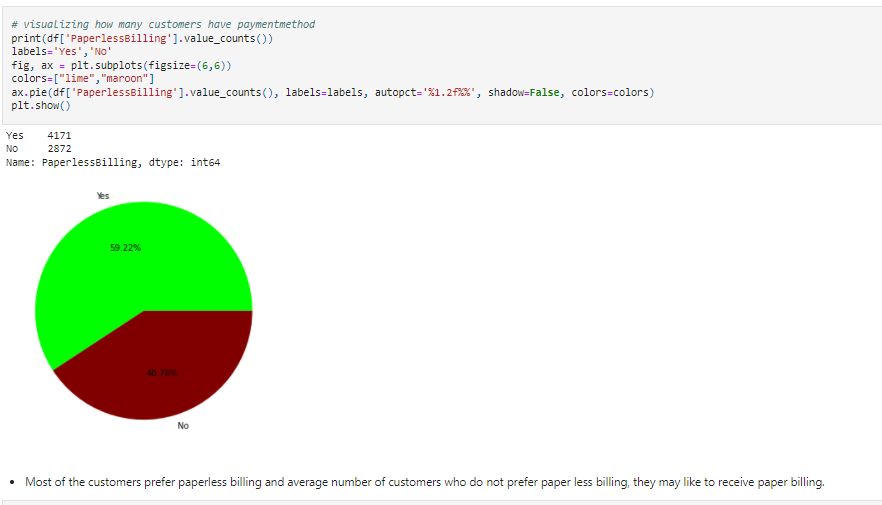
**

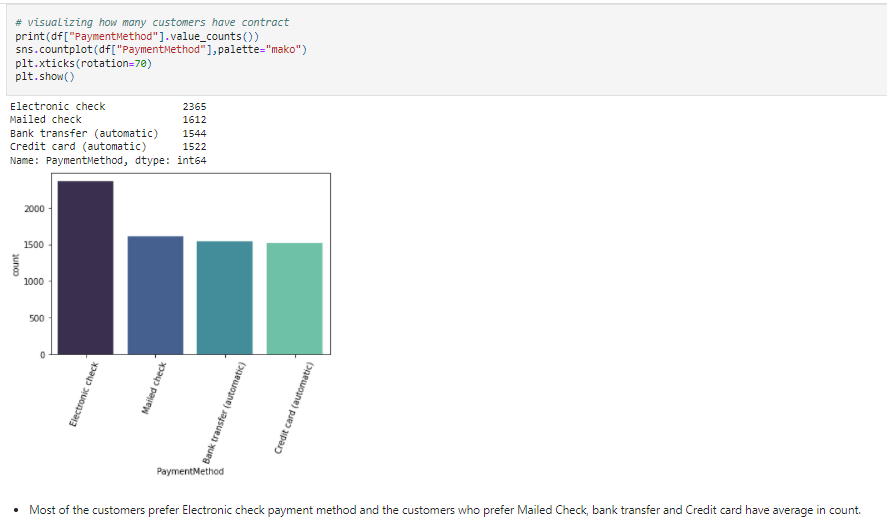
**

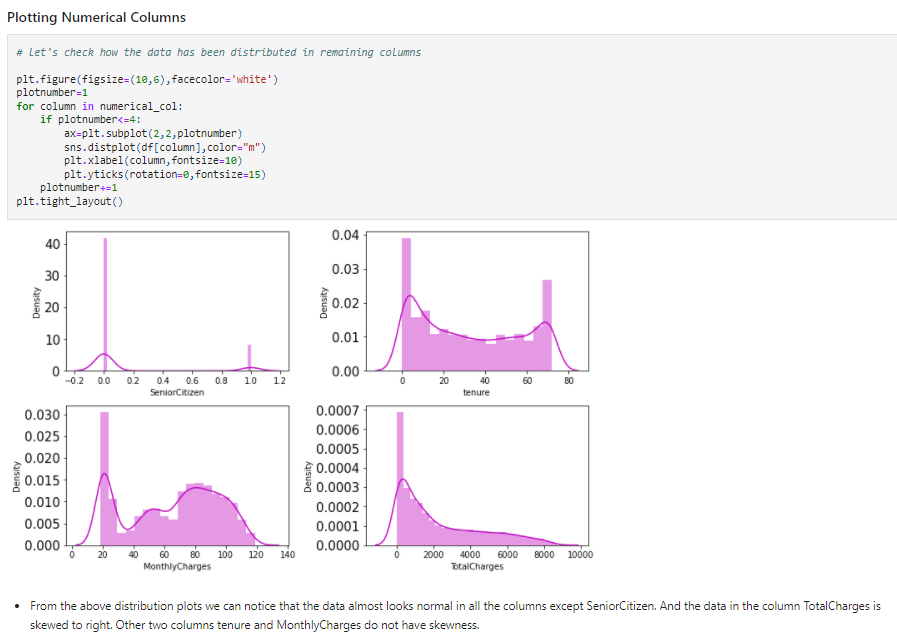
**

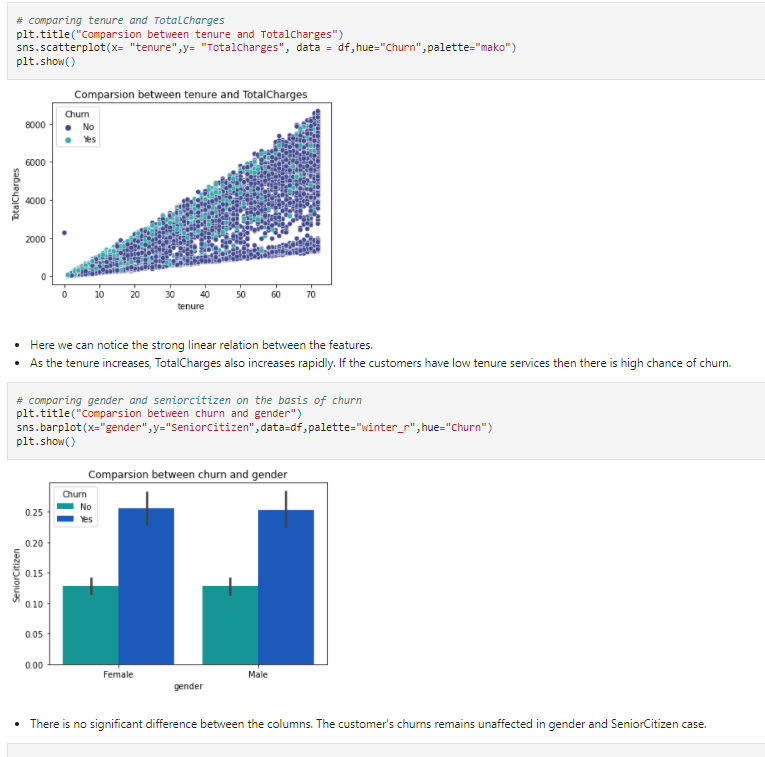
**

**

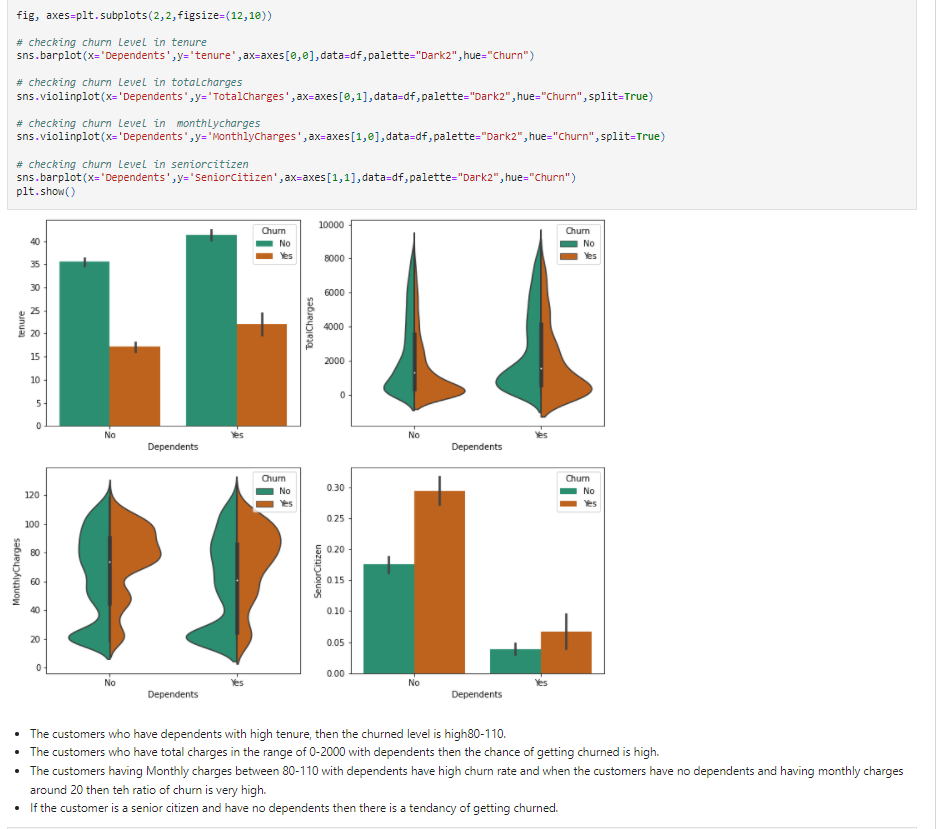
**

**

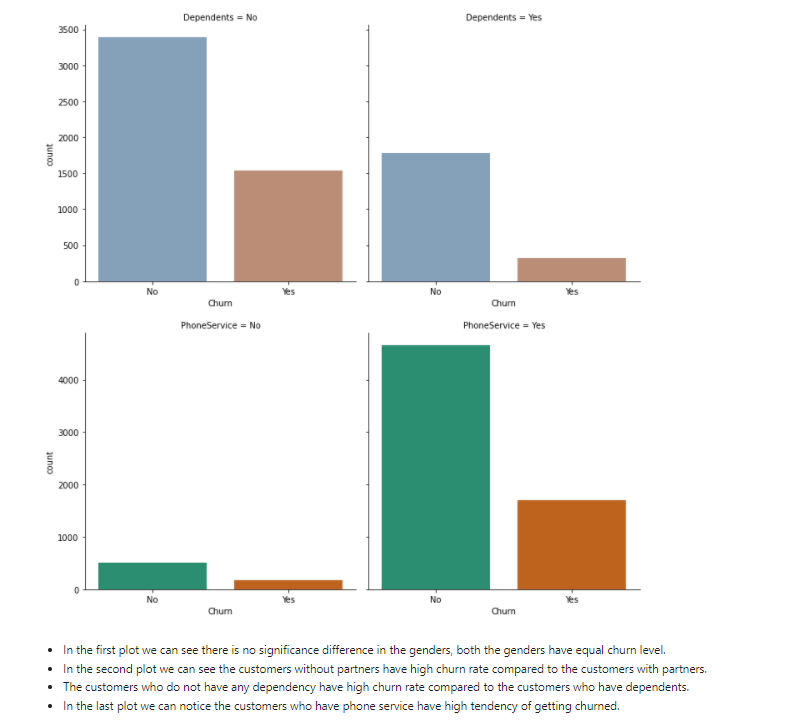
**

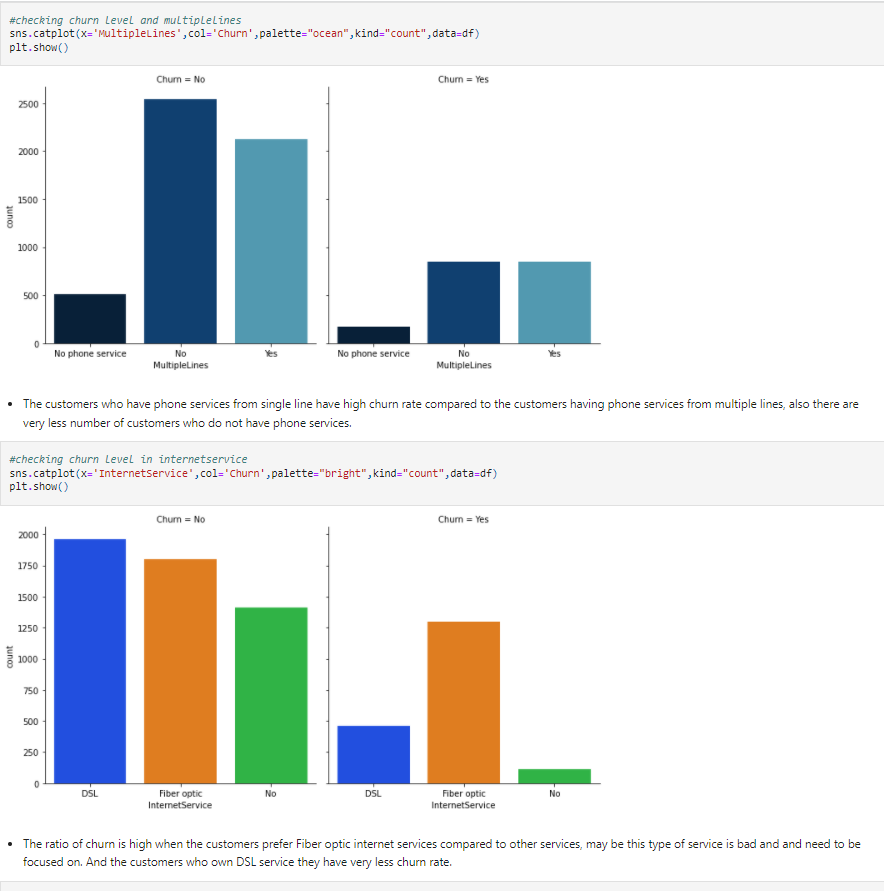
**

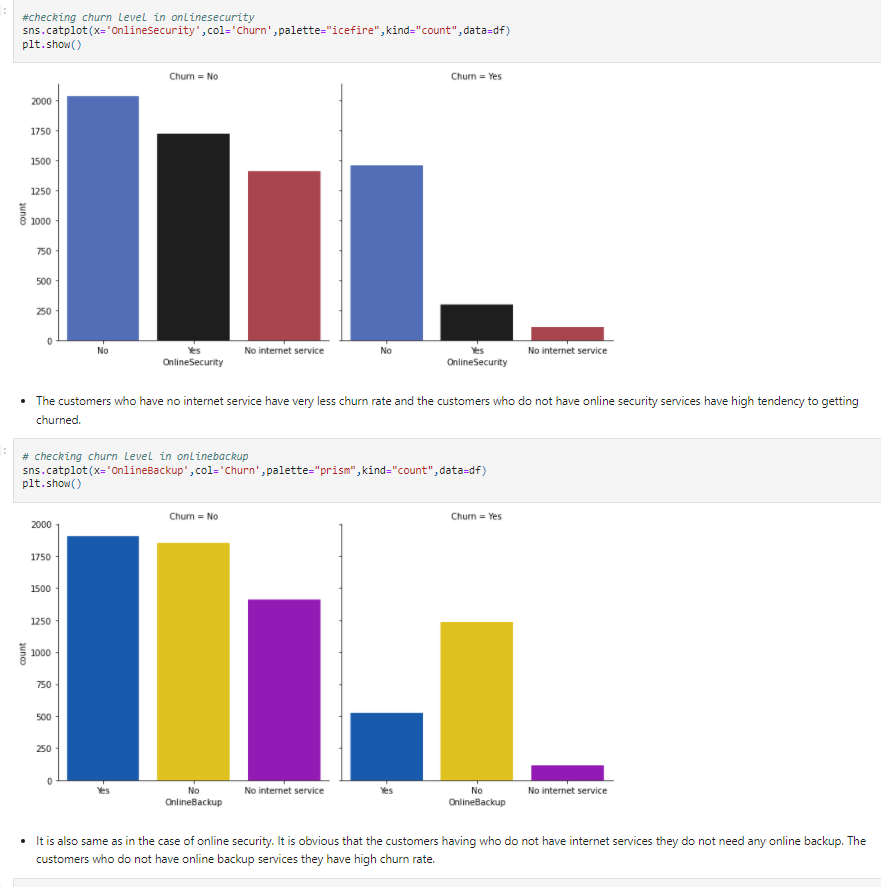
**

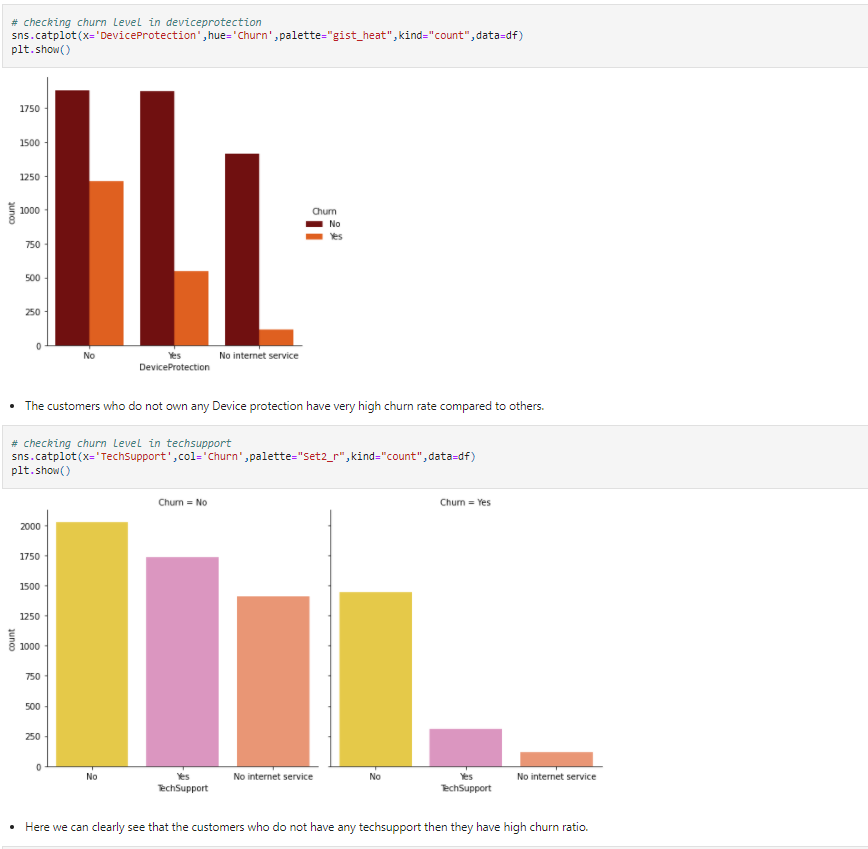
**

**

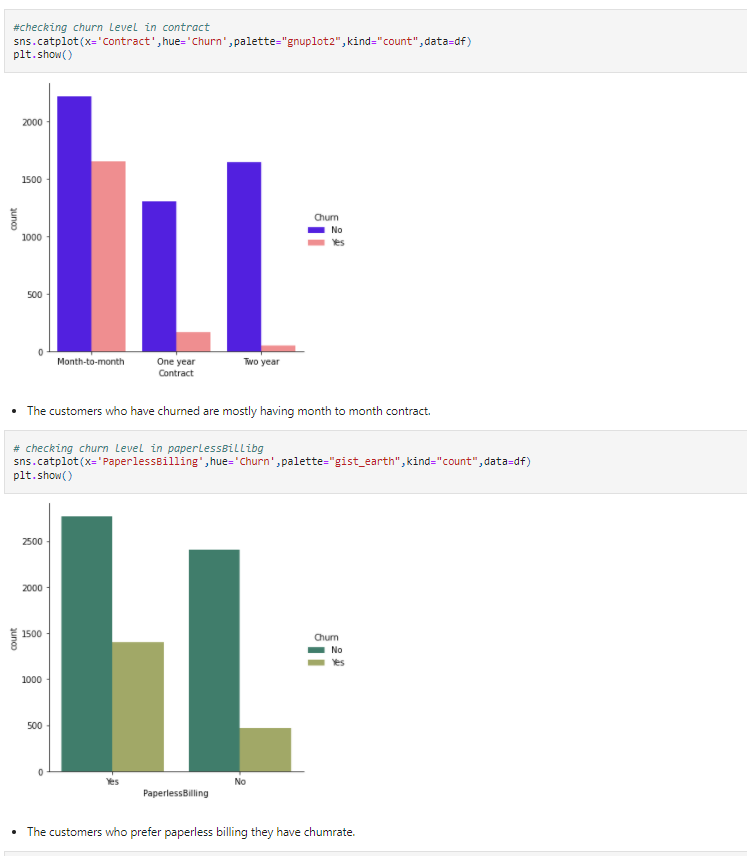
**

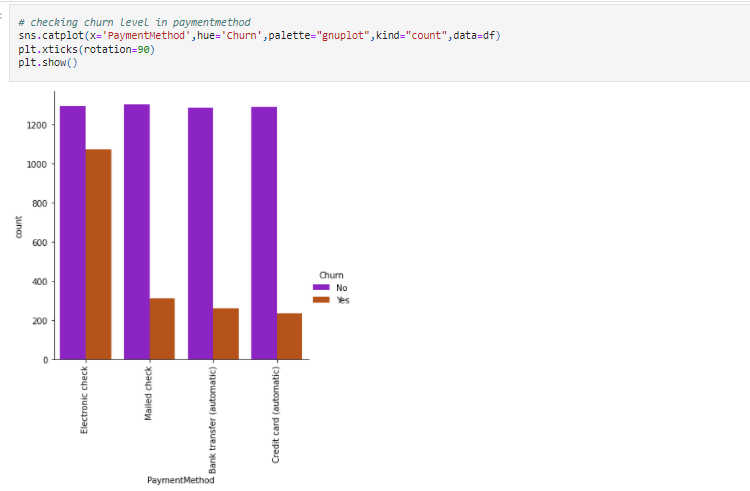
**

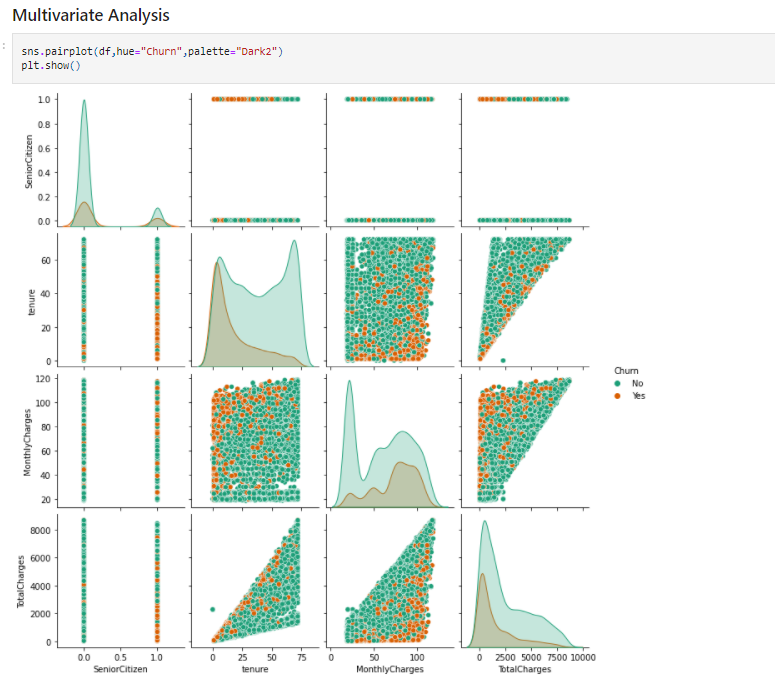
**

**

**

**

**

**

* The pair plot gives the pairwise relation between the features based on the target “Churn”. on the diagonal we can notice the distribution plots.
* The features tenure and Total Charges, Monthly Charges and Total Charges have strong linear relation with each other.
* There are no outliers in any of the columns.

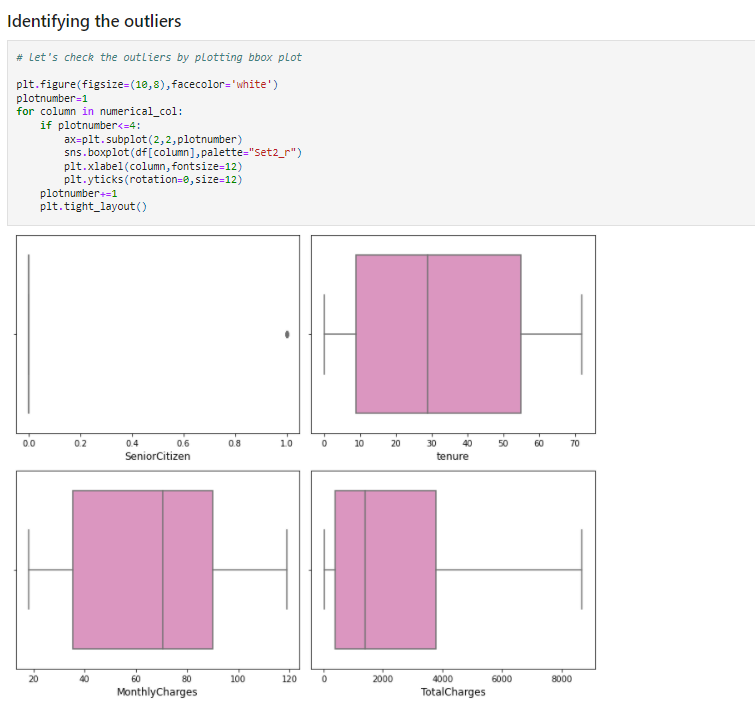
# **EDA Concluding Remark**

# We have observed the customer churn how they are using the services repeatedly in that who is majorly using the service with gender divided in that male and female are almost same but there is slide more in male compare to female .after this there is major customers who is non - senior citizen who is aware of it, this churn chain in these senior citizen who use this service is almost 16%,but the no senior citizen is high in utilized.

# In these the customers who have phone services have a large number whereas compared to non-phone services.

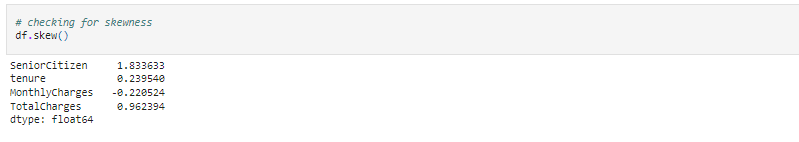
# There is also customer who don’t get the internet services but most of the customers have chosen the fiber optic internet.

* There are many more services which is done in above in form of plots to predict values.
* But we must identify the outliers where it is in senior citizen and solve the outliers using Zscore method as the data loss using this method is less compared to IQR method.

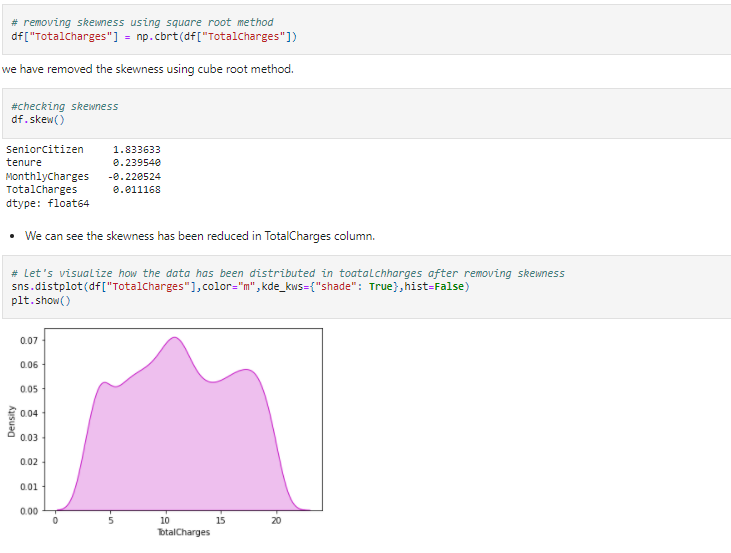
**

# As per said senior citizen has the outliers but it is in categorical data so no need to remove it. Apart from this none of the columns have outliers

# **Checking for Skewness**

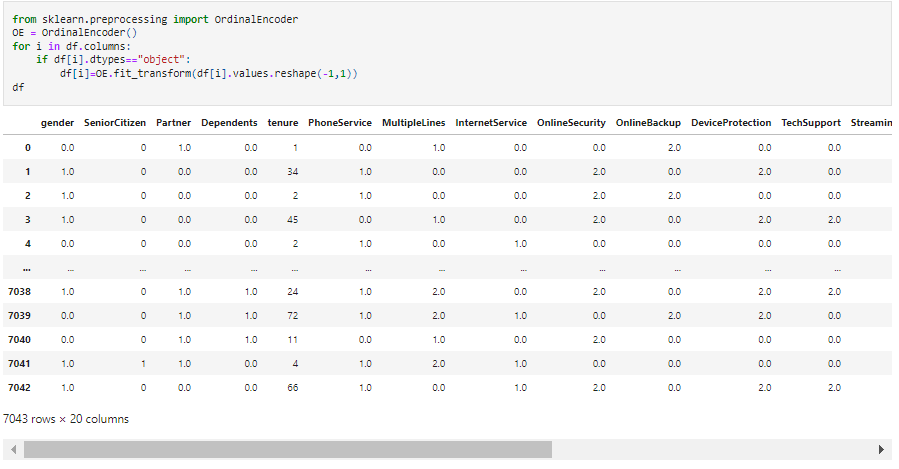
**

* But it is in the Total charges we must remove the outliers by using the cube root method.

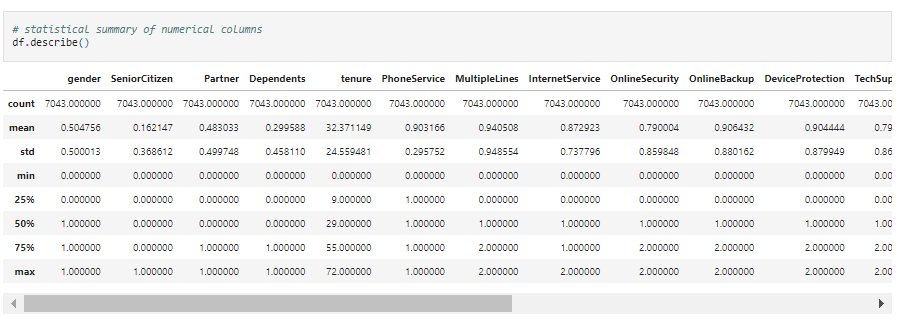
**

* It looks like normal, and the skewness is removed.

**Encoding the categorical columns using ordinal encoding**



* Now we have converted the categorical columns into numerical columns using the ordinal encoding method.

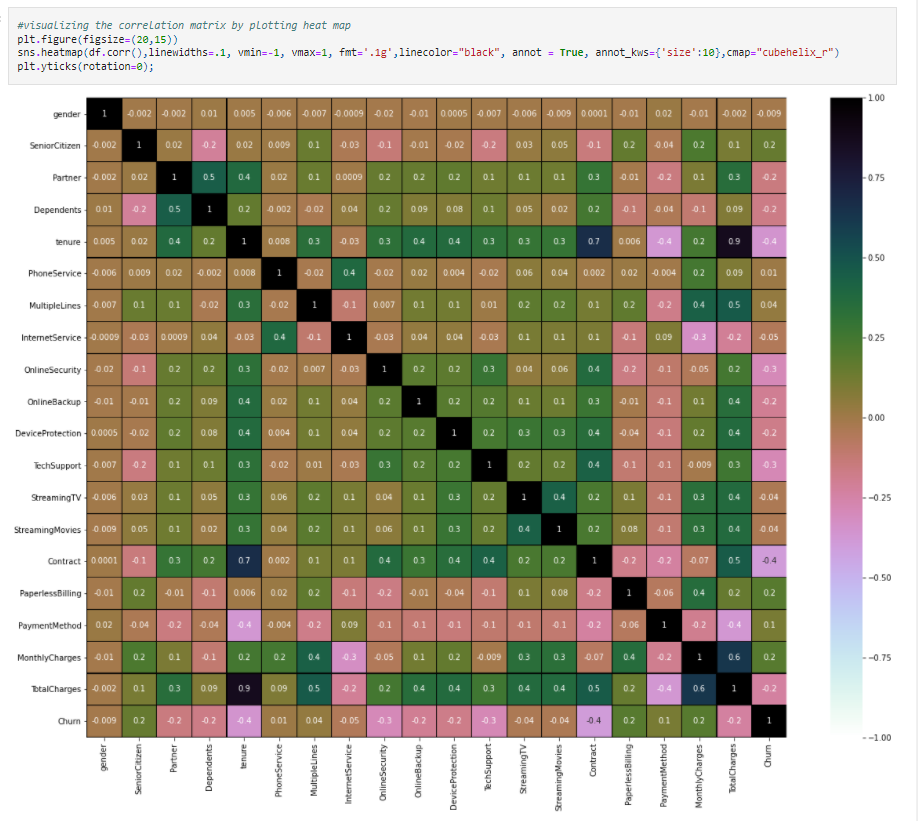
**

* After encoding the categorical column, we can see all the columns details here. The counts of all the columns are same that means no null values in the dataset.
* The describe method describes the count, mean, standard deviation, min, iqr and max values of all the columns.

# **Correlation between the target variable and independent variables using Heat map**



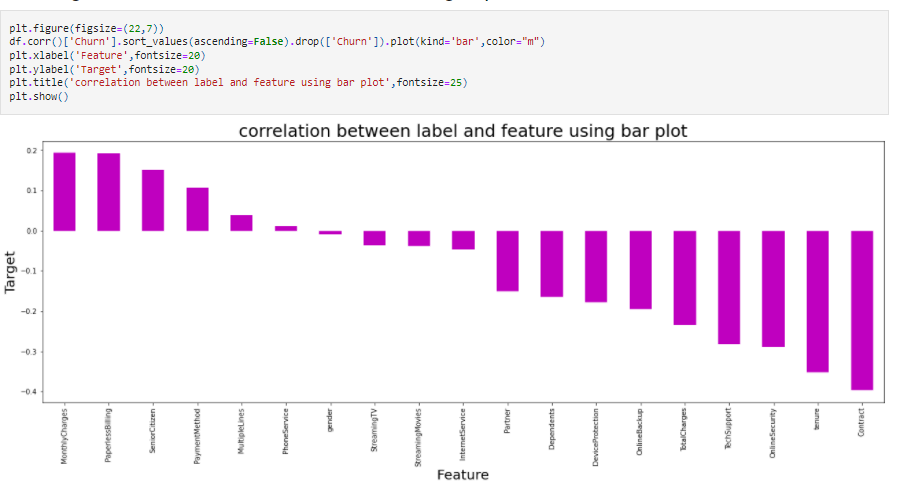
* This gives the correlation between the dependent and independent variables. we can visualize it in heat map.

**

This heat map shows the correlation matrix by visualizing the data, we can observe the relation between feature to feature and feature to label and this heat map have positive and negative correlation.

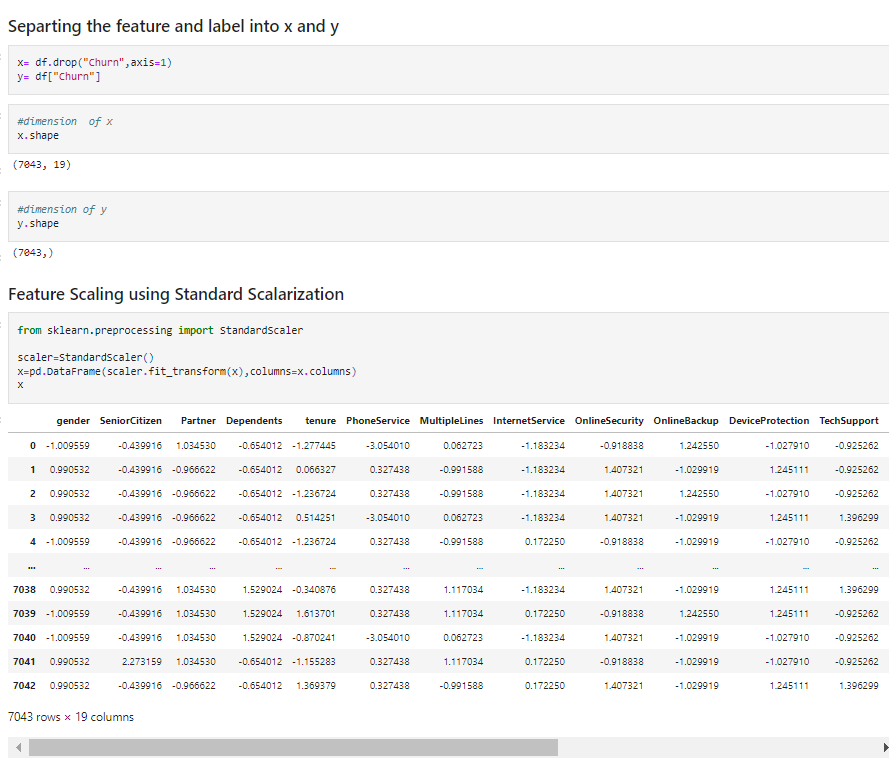
* There is not much positive correlation between the target and features.
* The columns Monthly charges, Paperless billing, Senior citizen and payment method have positive correlation with the label churn.
* The label is negatively correlated with contract, tenure, online security, Tec support, total charges, device protection, online backup, partner and dependents.
* Also, the column gender has very less correlation with the label, we can drop it if necessary.

# **Visualizing the correlation between label and features using barplot**



# From the above bar plot, we can notice the positive and negative correlation between the features and the target.

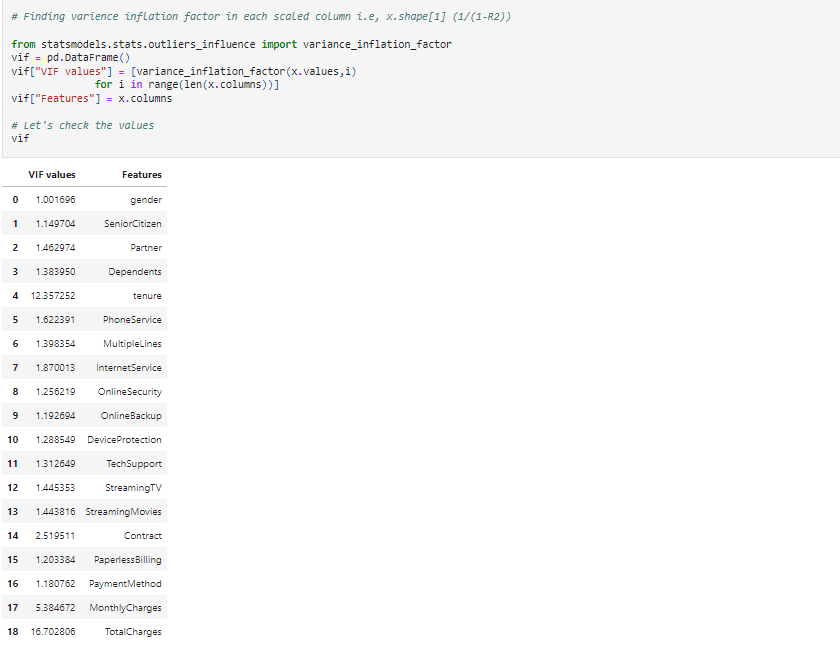
* Here the features gender and phone services have very less correlation with the column.

**

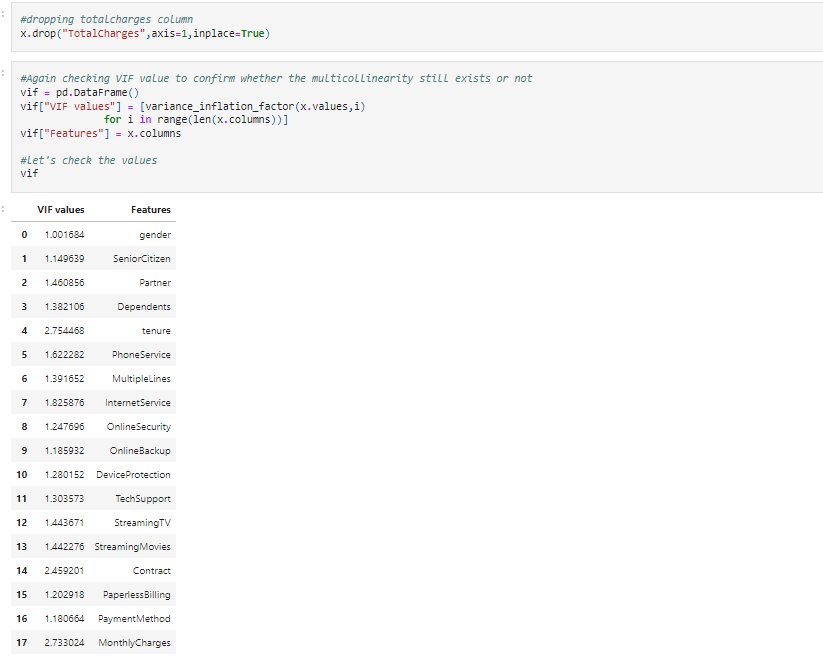
We have scaled the data using standard scalarization method to overcome with the issue of data biasness.

* In the heatmap we have found some features having high correlation between each other which means multicollinearity problem so let’s check the VIF value to solve multicollinearity problem.

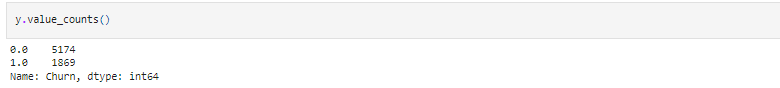
# **Checking Variance Inflation Factor (VIF)**



* By checking VIF values we can find the features which causing multicollinearity problem. Here we can find the feature total charges and tenure have VIF values greater than 10 which means they have high correlation with the other features.
* We will drop one of the columns first, if the same issue exist then will try to remove the column having high VIF (above10).

**

* Now all the columns have VIF less than 10 which means the data is free from multicollinearity problem.

**

* Here we can observe the data is not balance, since it is classification problem, we will balance the data using over sampling method.

# **OVER SAMPLING**

# 

# **Pre-processing**

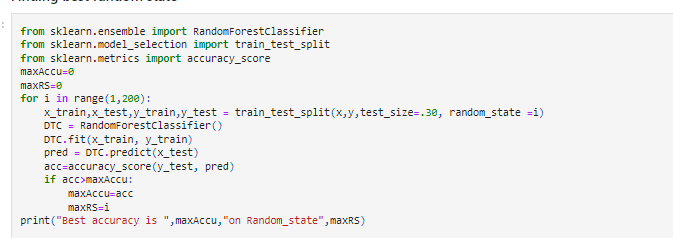
I had the above every method related to pre-processing pipeline which is checking skewness, removing skewness, encoding the categorical columns using ordinal encoding, correlation, feature scaling using standard scalarization, vif, and over sampling to balance data.

Since now we have done all the data analysis, EDA and pre-processing now it’s time to build our machine learning models.

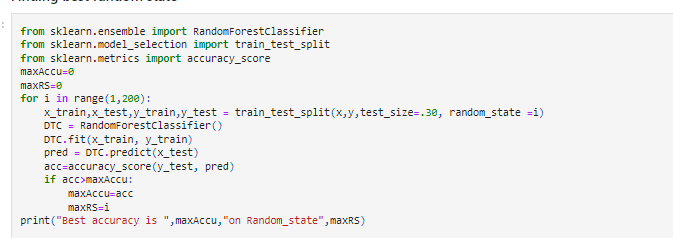
**Machine learning model**

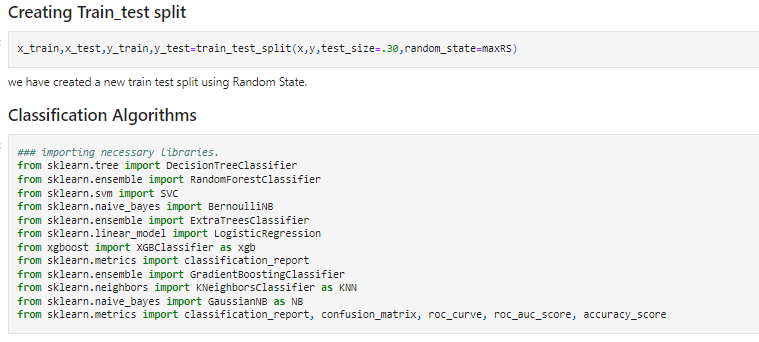
Before we build the models first, we need to find the first random state and accuracy using any one of the classification models.

# **Finding Best Random State**

**

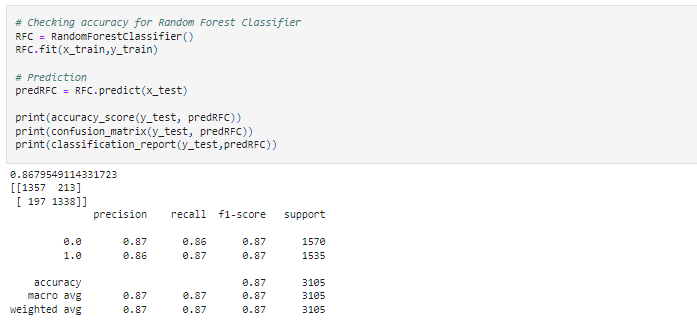
* Now I got best accuracy is 86.59 and Random state 102
* Now let’s create new train and test and fit them into the models to find our ideal model.

**

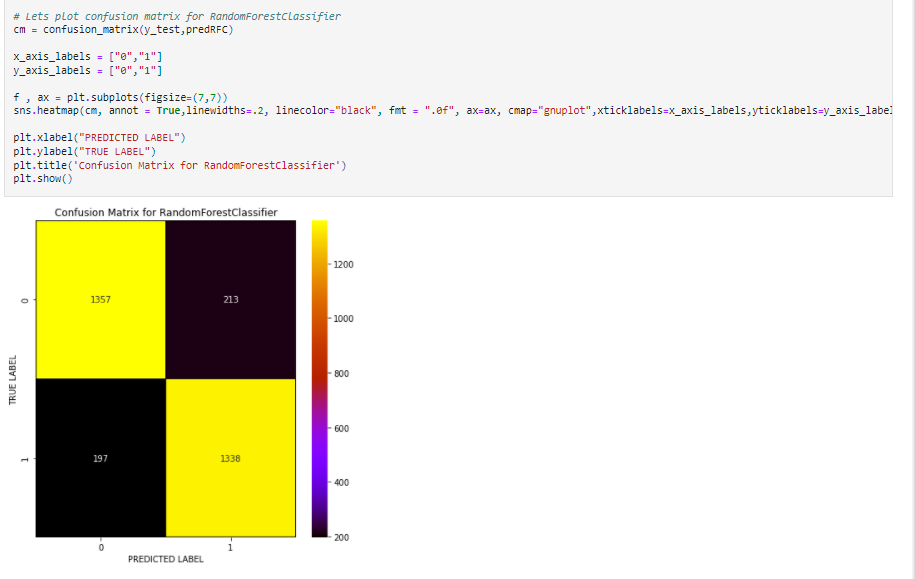
**

* All the machine learning algorithms are imported from sklearn library. In this project I have used 6 different algorithms to predict customer churn. The model which gives the best performance amongst them, we will be using that as best model for prediction.

# **Random Forest Classifier**

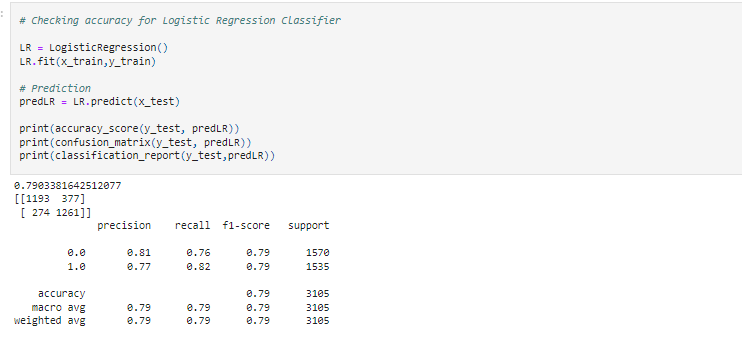


* Here we have created the Random Forest Classifier model and getting the accuracy for this model as 86.47%.



* With the help of confusion matrix, we can able observe the true positive rate, false positive rate, true negative rate and false negative rate and is plotted projected against true values.

# **Logistic Regression**

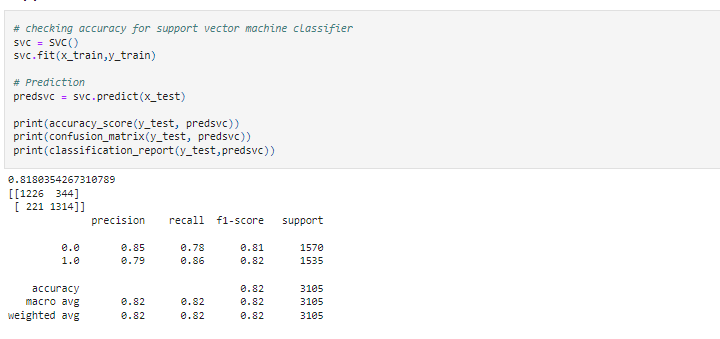


* We have created logistic regression model which is going the accuracy 78.84%.

**

* With the help of confusion matrix, we can be able to observe the true positive rate, false positive rate, true negative rate and false negative rate. And is plotted value against true values.

# **Support Vector Machine Classifier**

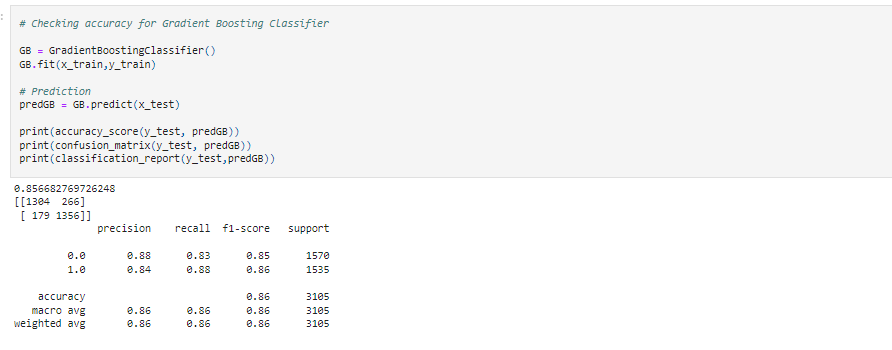


* Created Support Vector Machine Classifier model and it is giving the accuracy as 81.86%.

**

* With the help of confusion matrix, we can be able to observe the true positive rate, false positive rate, true negative rate and false negative rate. And is plotted predicted value against True values.

# **Gradient Boosting Classifier**

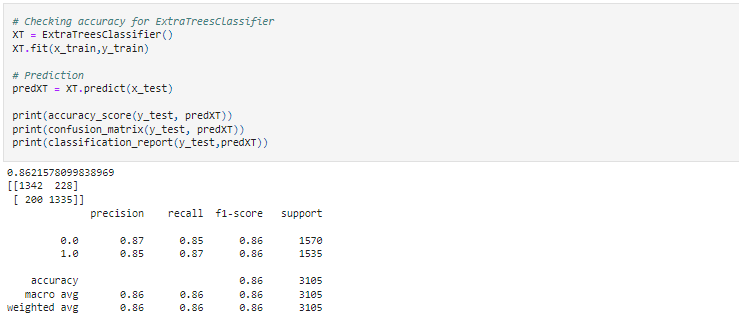


* We have created Gradient Boosting Classifier and the accuracy is 85.65%.

**

* With the help of confusion matrix, we can be able to observe the true positive rate, false positive rate, true negative rate and false negative rate. And is plotted value against true value.

Extra Trees Classifier:

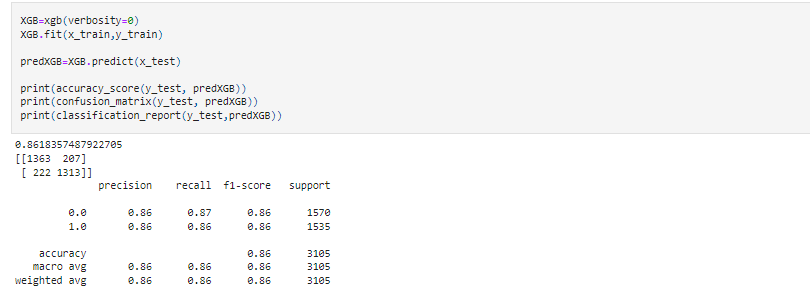
**

* Here we created the Extra Trees classifier model and getting accuracy for this model is 86.92%.

**

* With the help of confusion matrix, we can be able to observe the true positive rate, false positive rate true negative rate and false negative rate. And is plotted predicted value against True values.

# **XGB Classifier**



* Here we created XGB Classifier model and getting accuracy for this model is 86%.

**

* With the help of confusion matrix, we can be able to observe the true positive rate, false positive rate, true negative rate and false negative rate. And is plotted predicted value against True values.
* Above we predicted all models related to machine learning for customer churn from that model we got the best accuracy only, so let me know by the cross validation score more detail to conclude.

**Cross Validation Score**

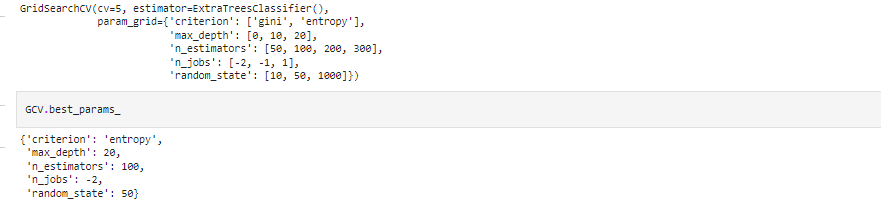
**

The difference between the accuracy score and cross validation score for the model used.

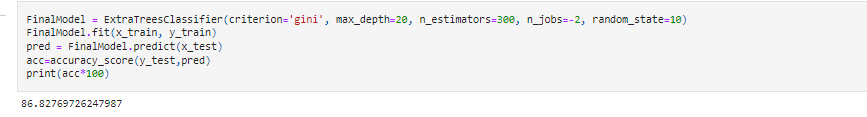
* Random Forest Classifier = 1.46%
* Logistic Regression = 1.03%
* Support Vector Machine Classifier = 1.97%
* Gradient Boosting Classifier = 2.34
* Extra Trees Classifier = 0.67%
* XGB Classifier = 2.68%
* From the difference between the accuracy score and the cross validation score we can conclude that Extra trees classifier as our best fitting model which is giving very less difference compared to other models.

# **HYPER PARAMETER TUNING**





* These are best parameters values for Extra trees classifier.

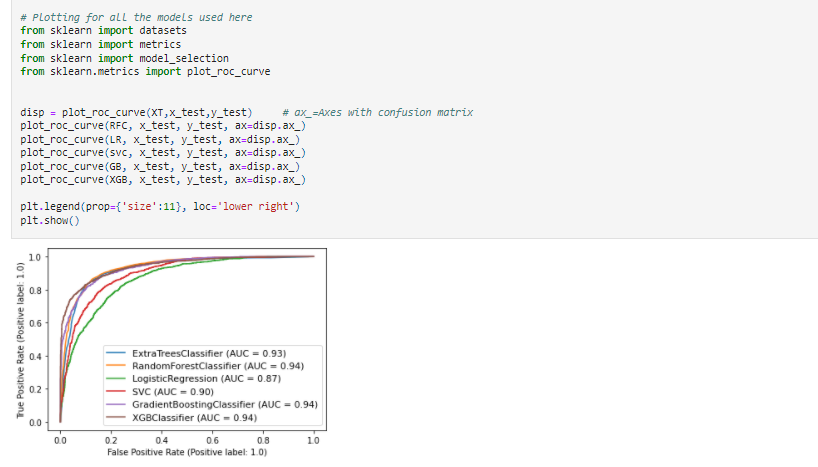
**

* The accuracy of best model increased after tuning and giving 86.82% which is very good.

**

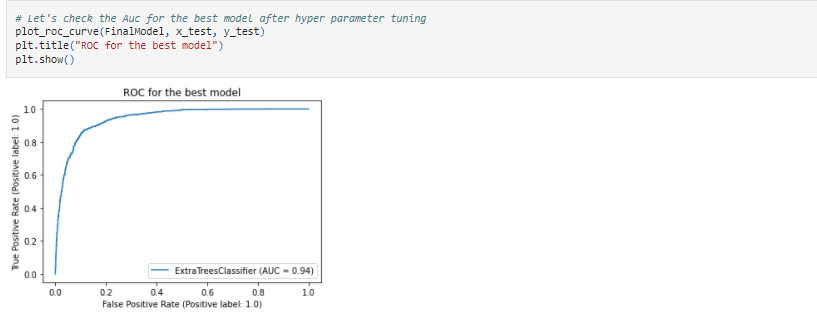
* With the help of confusion matrix, we can be able to see actual and predicted values.

# **Plotting ROC and Compare AUC for all the models used**

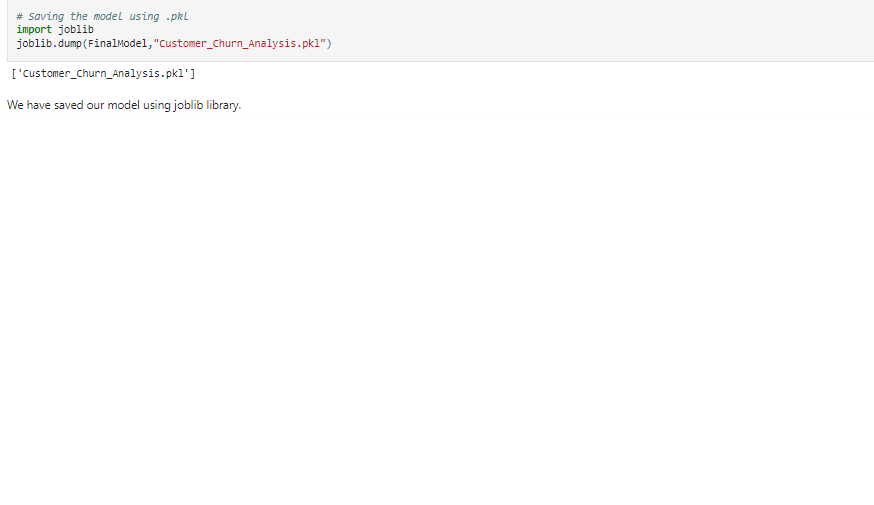


* Here we can see the area under the curve for each model used here.

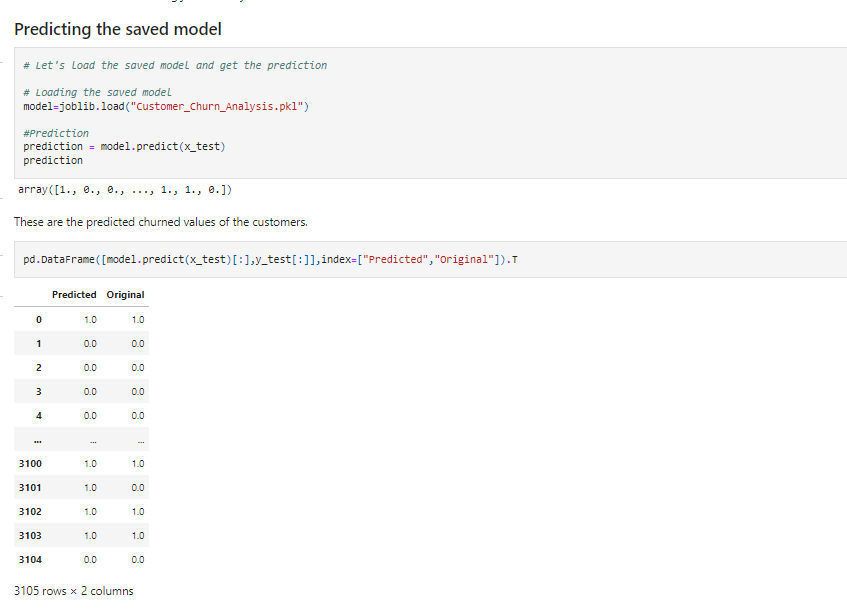
# **Plotting ROC and Compare AUC for the best model**



* Here we have plotted the ROC curve for the final model and the AUC value used for the best model is 94%.

**

* Above we have saved our model and now we have to save the predict model

**

**CONCLUSION**

* This problem needs a good vision on data, and in this problem “Feature Engineering” is the most crucial thing.
* You can see how we handled numerical and categorical data and how we build different machine learning models on the same dataset.
* I Think this process gives you easy to understand the “Customer Churn” for going further step.