Customer Churn Analysis

## Problem Definition:

Customer churn is a common problem across businesses in many sectors. If you want to grow as a company, you have to invest in acquiring new clients. Every time a client leaves, it represents a significant investment lost. Both time and effort need to be channelled into replacing them. Being able to predict when a client is likely to leave, and offer them incentives to stay, can offer huge savings to a business.

Here in this dataset customer churn (*target variable*) is provided with yes no type nominal data, which itself is clear that it is a Classification problem where we have to predict whether a customer will churn or not. To predict this we have been given various data due to which customers are more often to change the service provider. In this project we will be using many classification models that can help the companies to make decisions by predicting, how and why the customer churn will evolve in the future.

## Data Analysis:

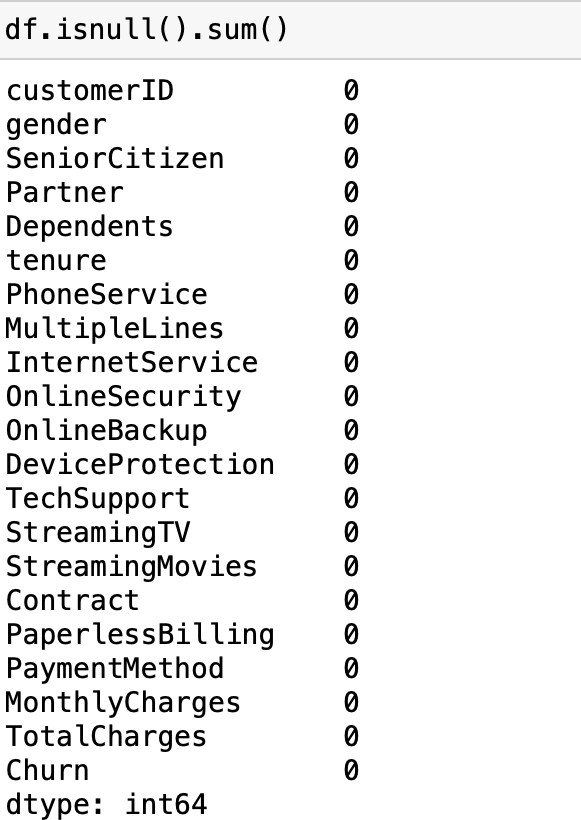
* Reading the CSV file and doing initial statistical analysis (shape, values etc).
* Solving missing values.
* Data Pre-processing: Reading the unique values for each column and removing those which won’t be significant in the analysis further.
* Model building and saving the best model with pickle.

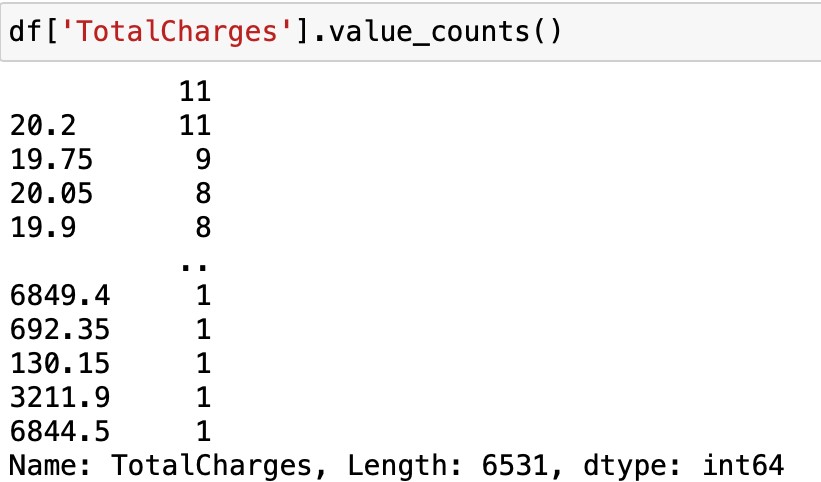
**Dataset contains:**

* customerID
* gender
* SeniorCitizen
* Partner
* Dependents
* tenure
* PhoneService
* MultipleLines
* InternetService
* OnlineSecurity
* OnlineBackup
* DeviceProtection
* TechSupport
* StreamingTV
* StreamingMovies
* Contract
* PaperlessBilling StreamingTV
* StreamingMovies
* Contract
* PaperlessBilling
* PaymentMethod
* MonthlyCharges
* TotalCharges
* Churn

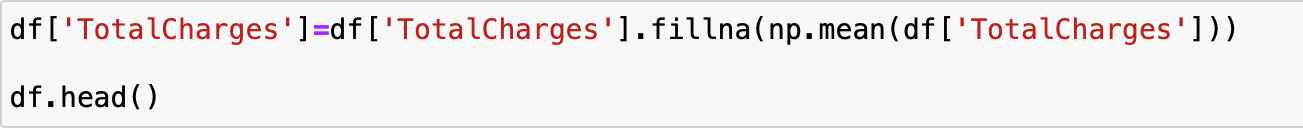
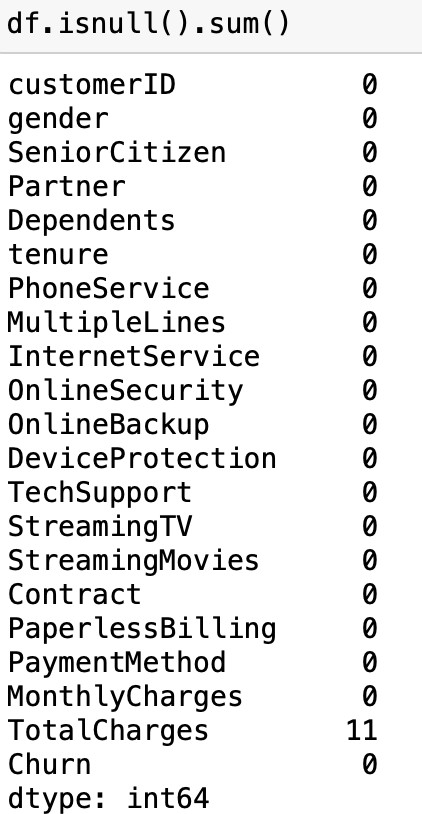
All given data describes its meaning itself and most columns are having nominal data.

Here, only 4 columns are Numerical, rest are Categorical.

At first I didn’t got any nan values in data but soon after certain analysis I found that there are values with not marked as nan.



Empty space later converted to mean of that column value

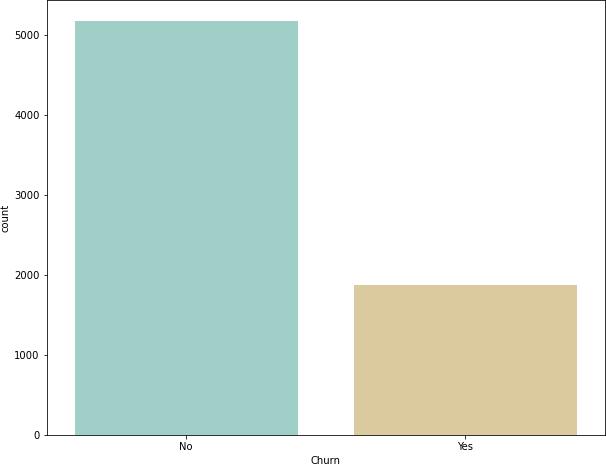
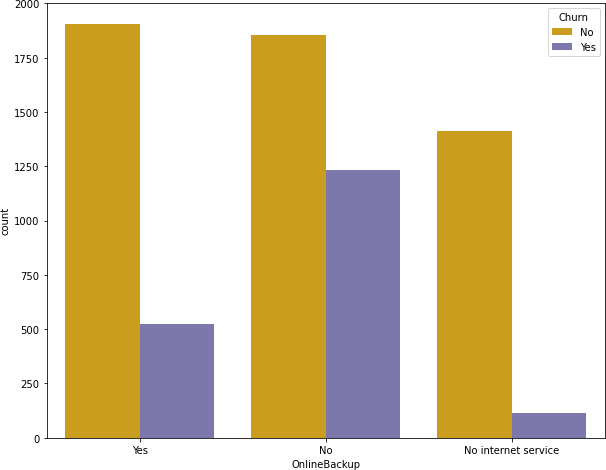


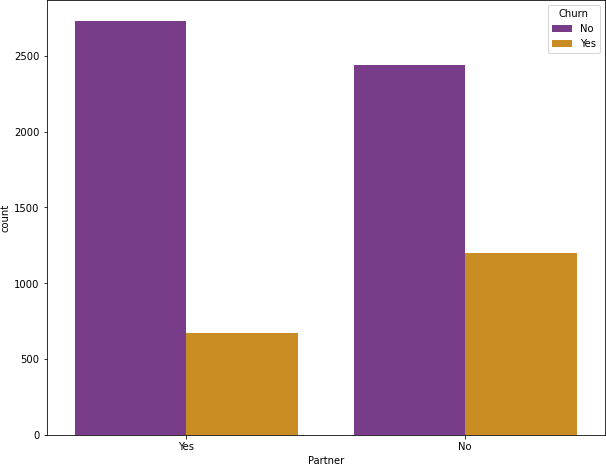
Converted the empty space to nan first with this above function than to its mean.

Finally there were no missing value then.

So then I filled that empty space with nan and again filled that nan values or missing values with mean. now my data is fulfilled with no missing values.

After all fixing missing values and reading all data I started visualization as to understand the graphical behaviour with churn rate for knowing what are the points related to customer churn.





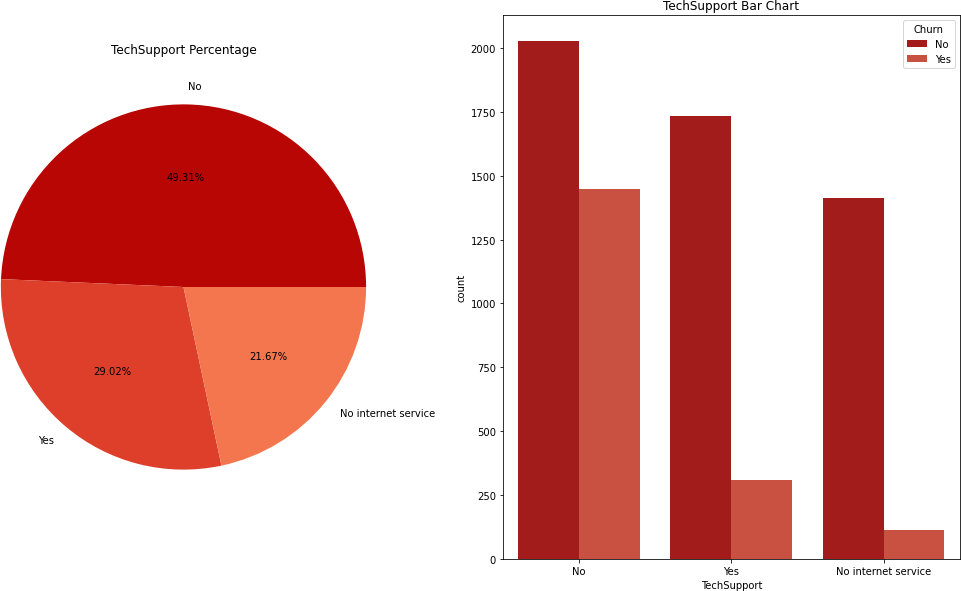
Plot relating churn with various aspects are shown here for better understanding.

Comments:

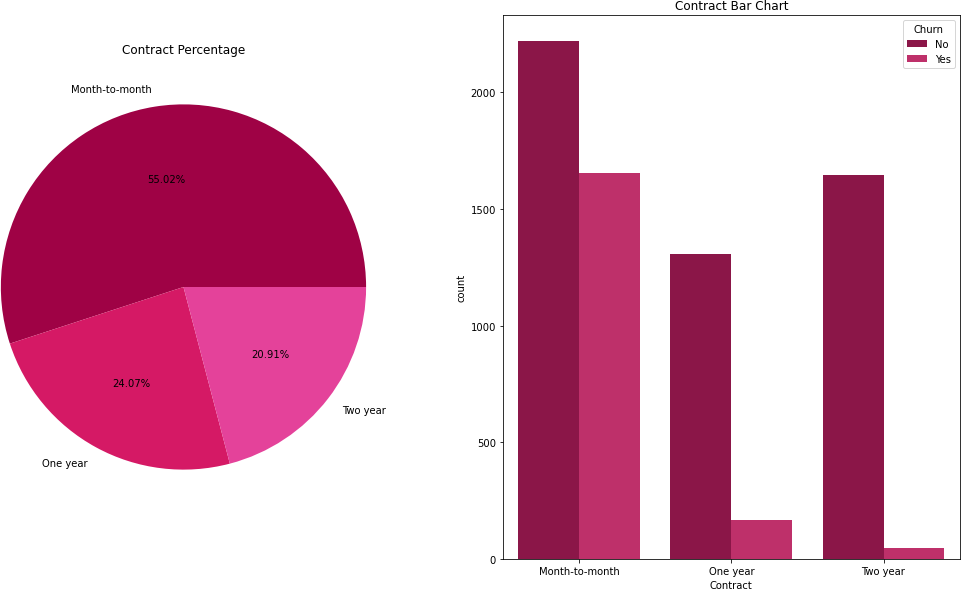
* Churning rate for the company are low but is still a problem.
* As we can see from above graphs company who has no online backups are having high churn rate. As customer need all benefits for their comfort and smooth process.
* Fiber optic service provider are also high which boom their company and attract new customers leading to customer churn for those company who don’t have this in present.

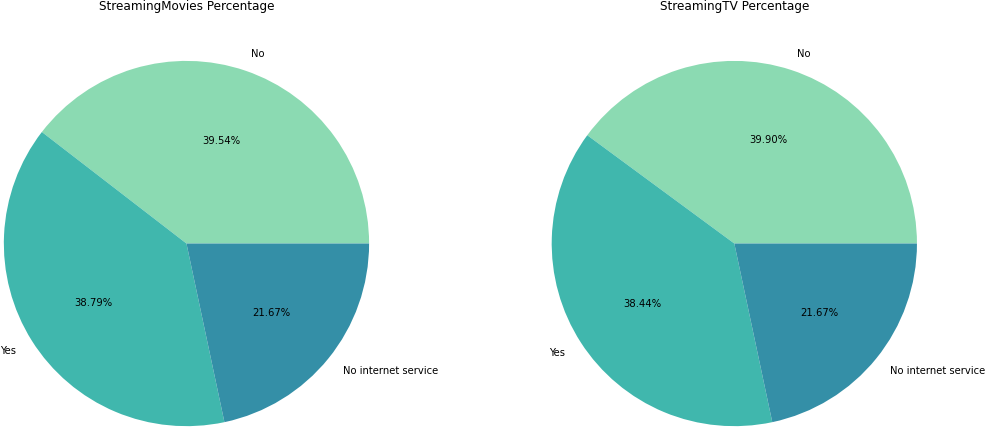
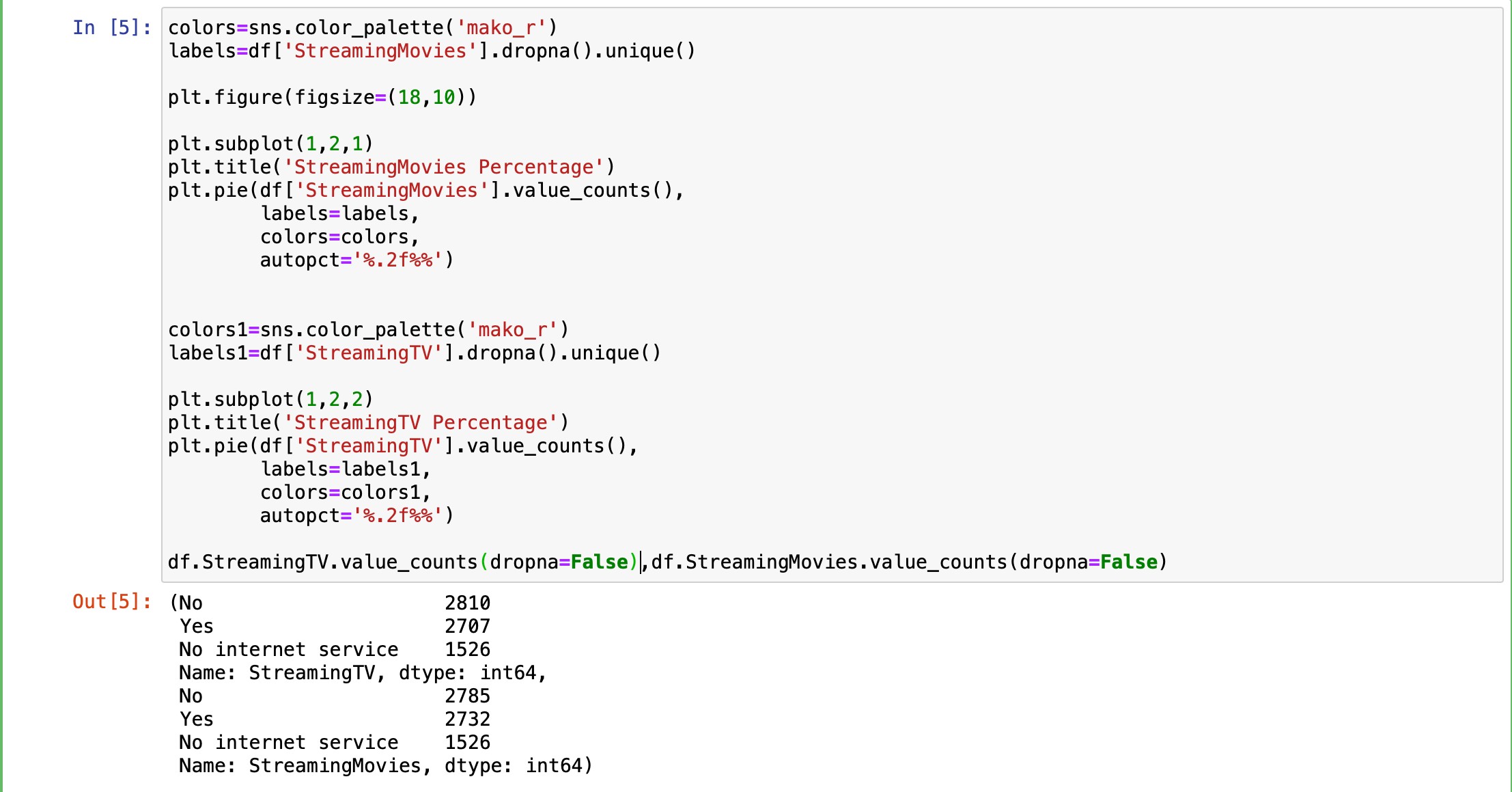
As the telecommunications sector is one of such sector where changing brand or moving to other company is relatively very easy and transparent market. Customers are more tend to attract to those company which gives more benefit and charge less.

During visualisation I found various factors which are same and can be dropped after correlation.



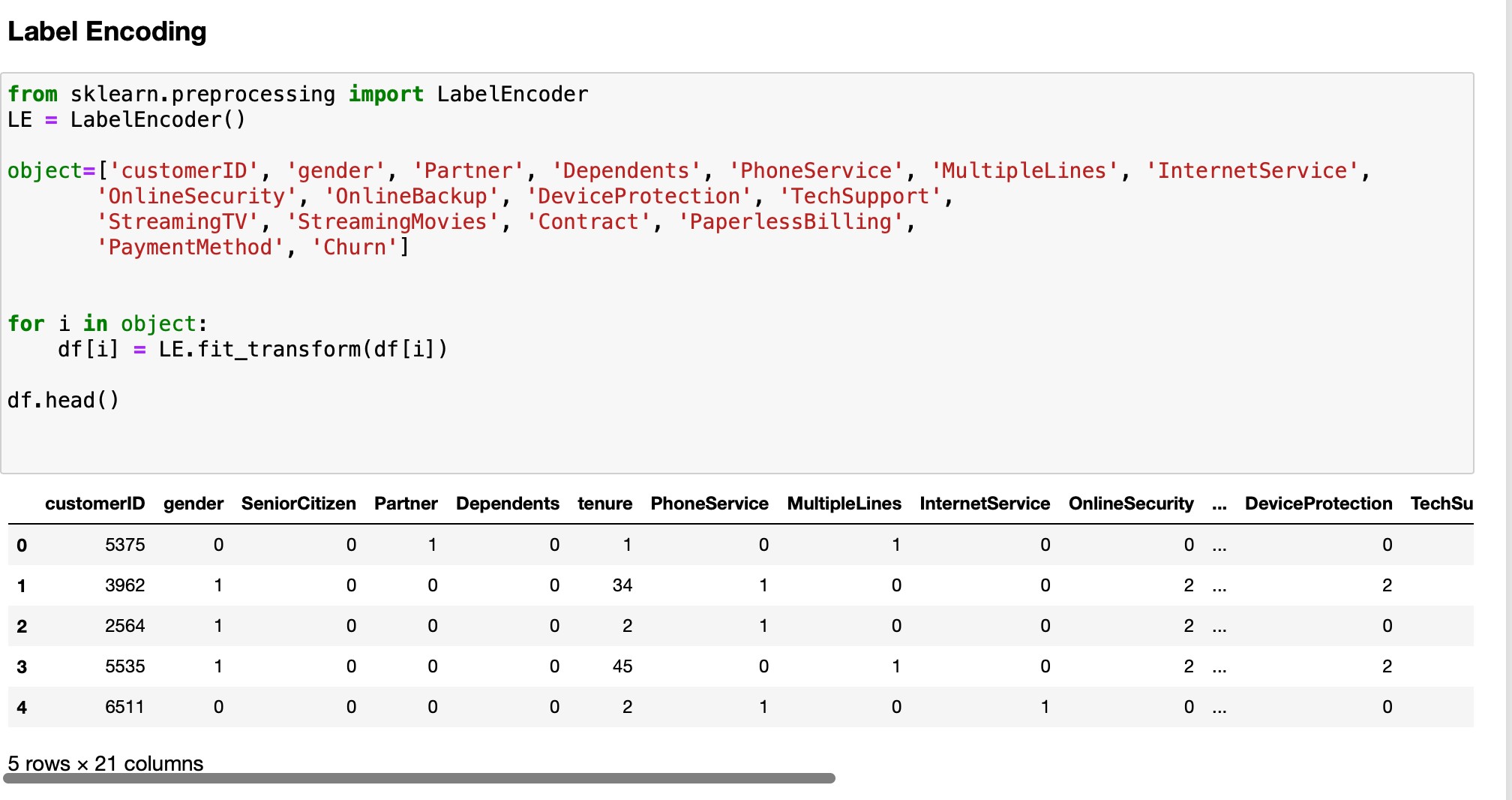
Company having tech support are having low churn rate.

C o n t a c t i n g o r subscription is the best way to reduce the churn rate. As month to month contract subscribers are more tending to churn as they can opt for other c o m p a n y a t a n y moment.

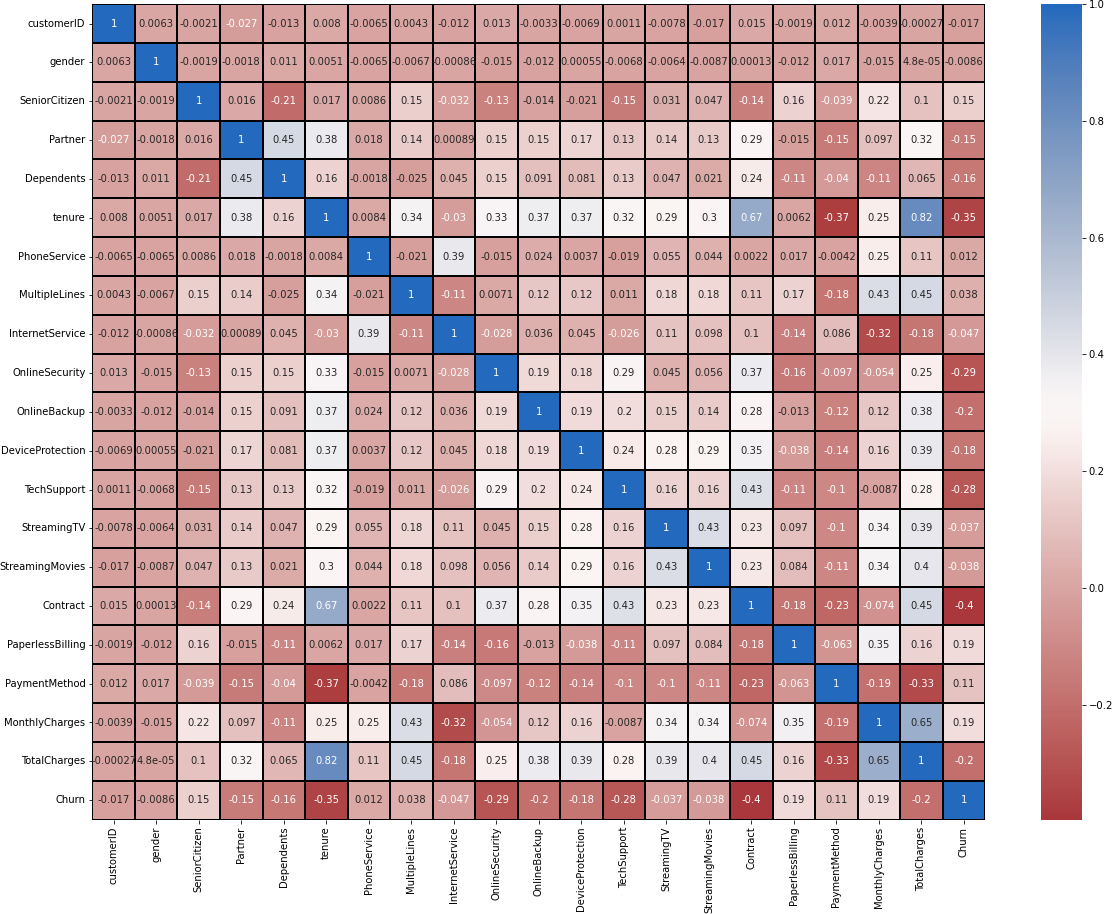


Streaming movies and streaming tv are having equivalent data so decided to drop one after correlation.also the code written for plotting used to compare those categorial data.We can also see value counts of them.

Now,As majority are categorial columns , so first applying **preprocessing** and converting all categorical data to numerical by **Label Encoder.**



Next, The correlation between different features of the dataset showed that customers with low satisfaction level left. The correlation heat map is shown below:



From graph we can see that there are normal correlation between some columns, hence will drop after checking VIF.

* Streaming tv and streaming movies are having correlation near 0.2.
* Device protection and tech support are having correlation near 0.2.
* Online security and online backup are also having almost same values.
* Since cases of multicollinearity is seen, later by applying VIF it we be solved.

## EDA Concluding Remark:

* After EDA I came to conclusion that customers tend to shift from one company to other only because of some extra benefits or low pricing or easy and smooth process.
* Technical service to be provided to user.
* Relationship could be seen between many features by matplotlib and seaborn.

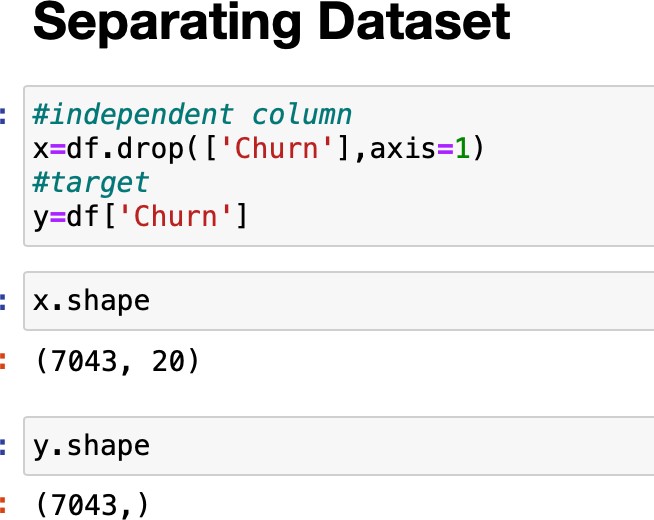
We have made all the data transformation by changing the categorical columns to numerical for further process . Now it is time to check the outliers and skewness in the dataset.

## Pre-Processing Pipeline:

I used box-plots to identify the outliers and found no outliers as some were present in categorical datas which was due to label encoding so moving further with preparing dataset into training and testing set, separating features and target columns and then checking Skewness.

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed. Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution. A normal distribution has a skew of zero, while a lognormal distribution, for example, would exhibit some degree of right-skew.

## Separating Feature and Target column:

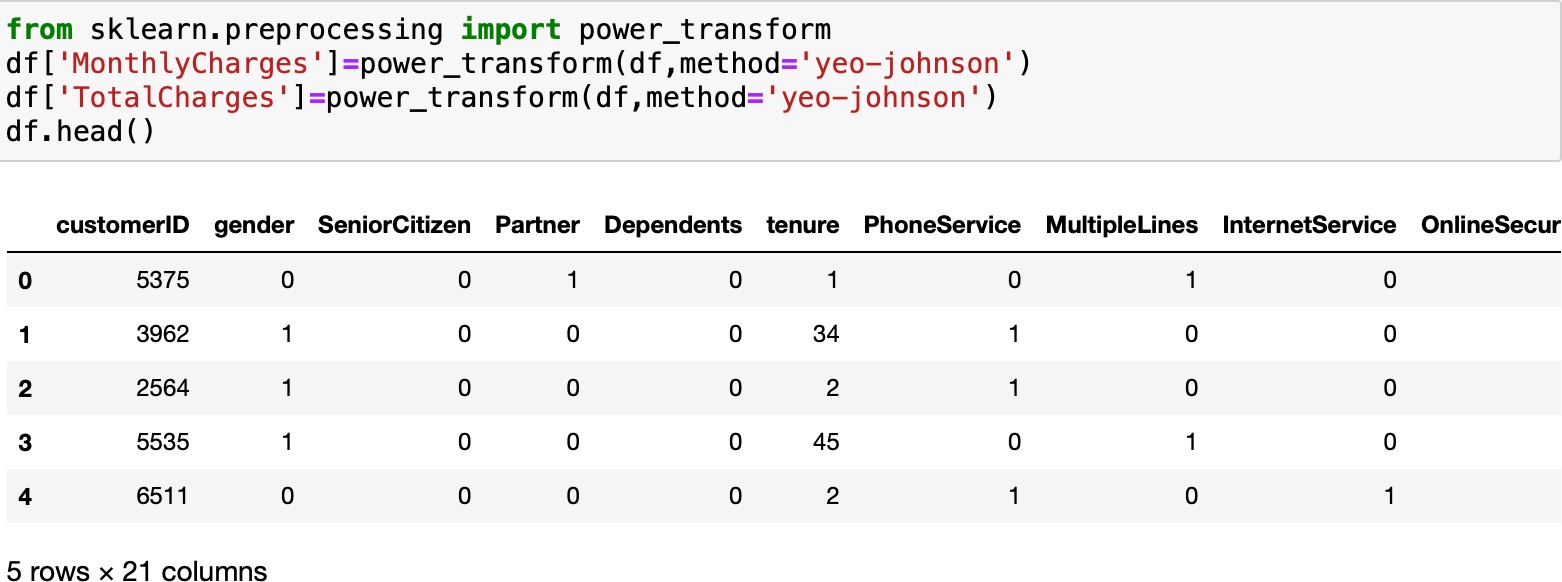


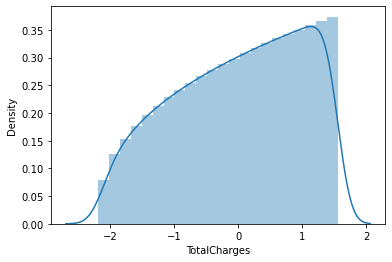
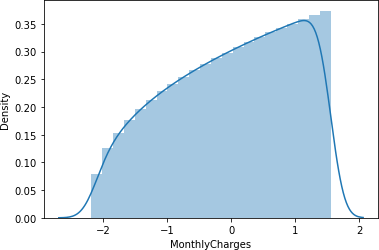
**Skewness Check**

I had used dist-plot to show skewness in columns ,and observed skewness in two columns :

* Total charges
* Monthly charges

Hence removing skewness by power transform.



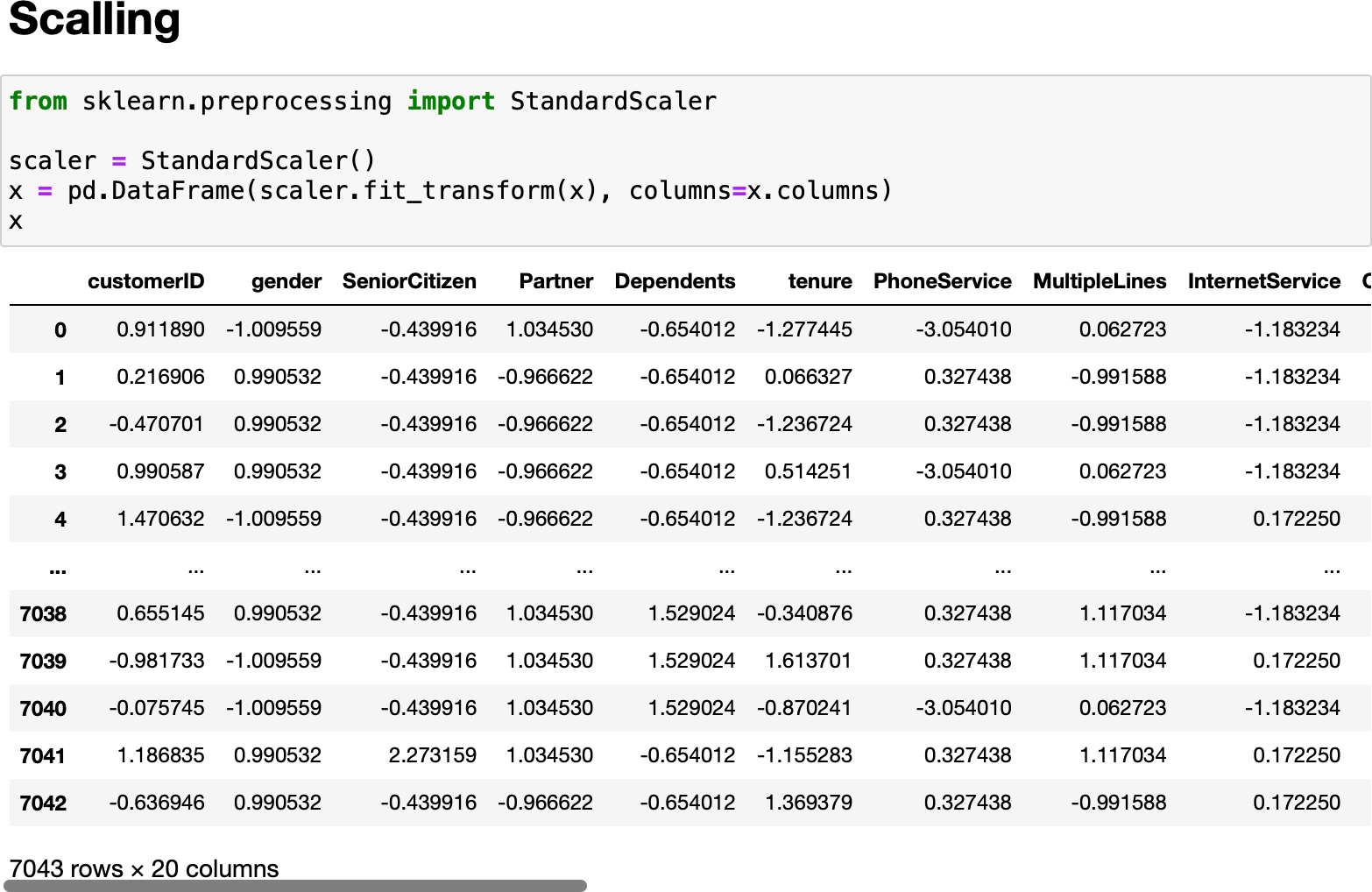


Since the skewness of the data is in the acceptable range and the data is also normally distributed in the columns, in such case we can make use of Standard Scaler method else we can make use of Min Max scaler method.

Here we used standard scaler method.

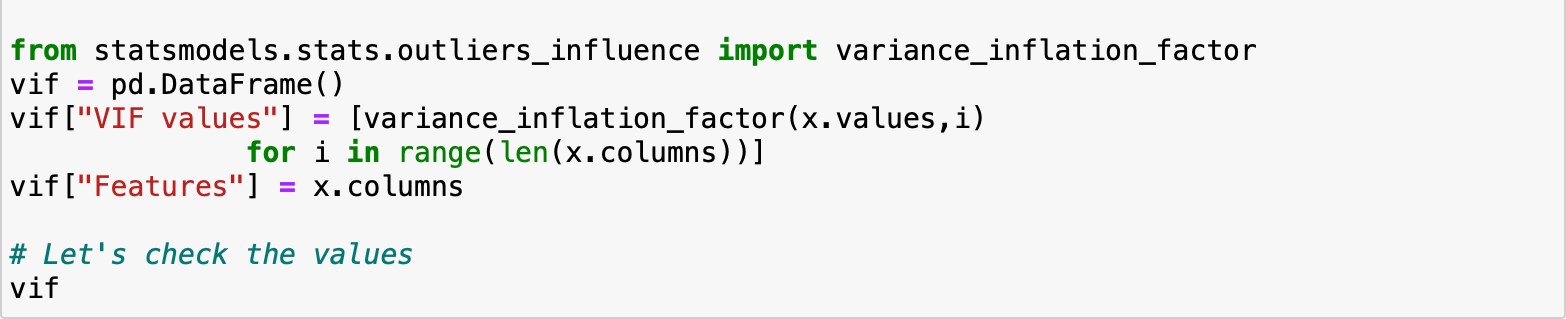
## Standard Scaler method:

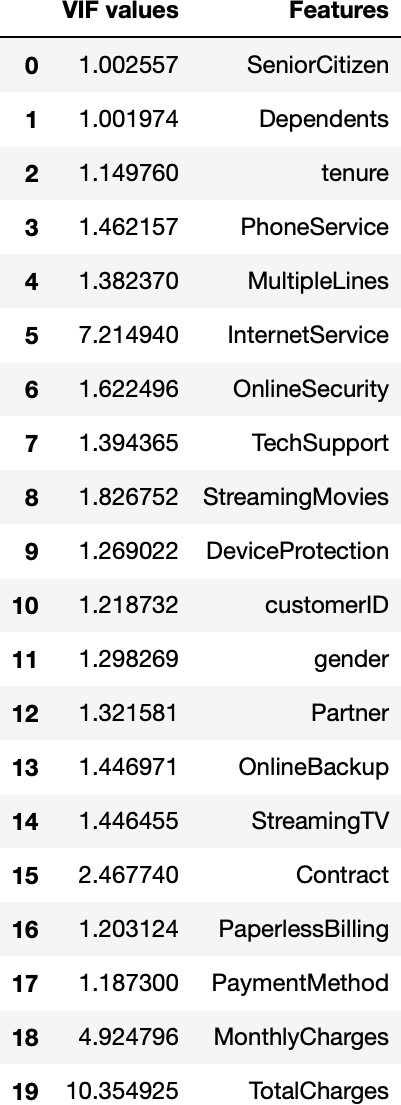
Standard Scaler helps to get standardised distribution, which makes mean = 0 and scales the data to unit variance. It helps in improving our model accuracy also solve the issue of data bias.



In the heat map we have found some features having high correlation between each other which leads to multicollinearity problem. In order to solve multicollinearity problem, we will check VIF values. If we find VIF values greater than 10 in any features that means the features causing multicollinearity issue in the data. To overcome with this problem, we need to drop that particular feature column.

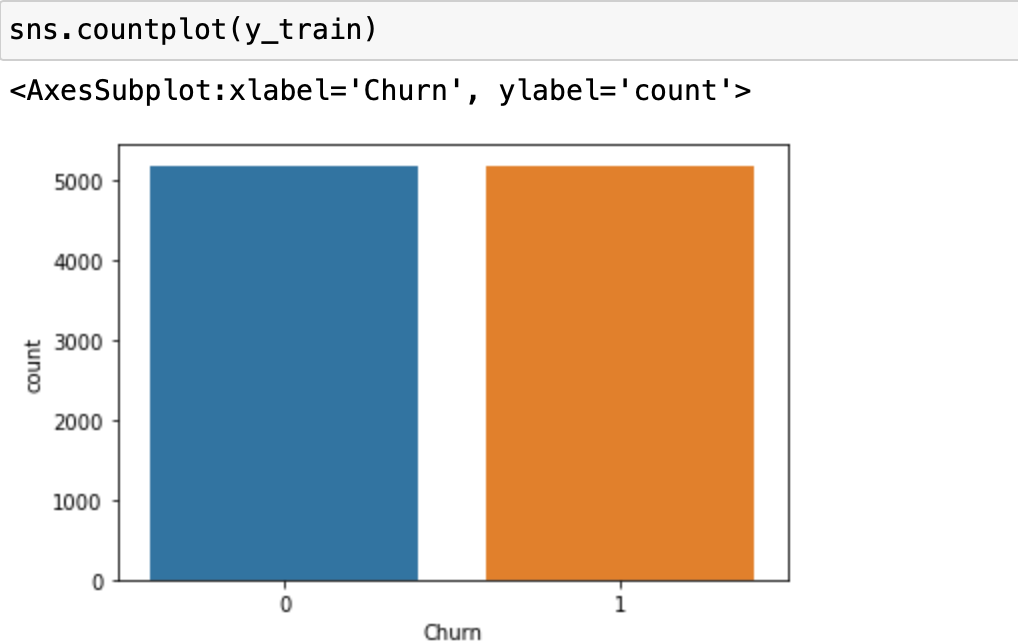
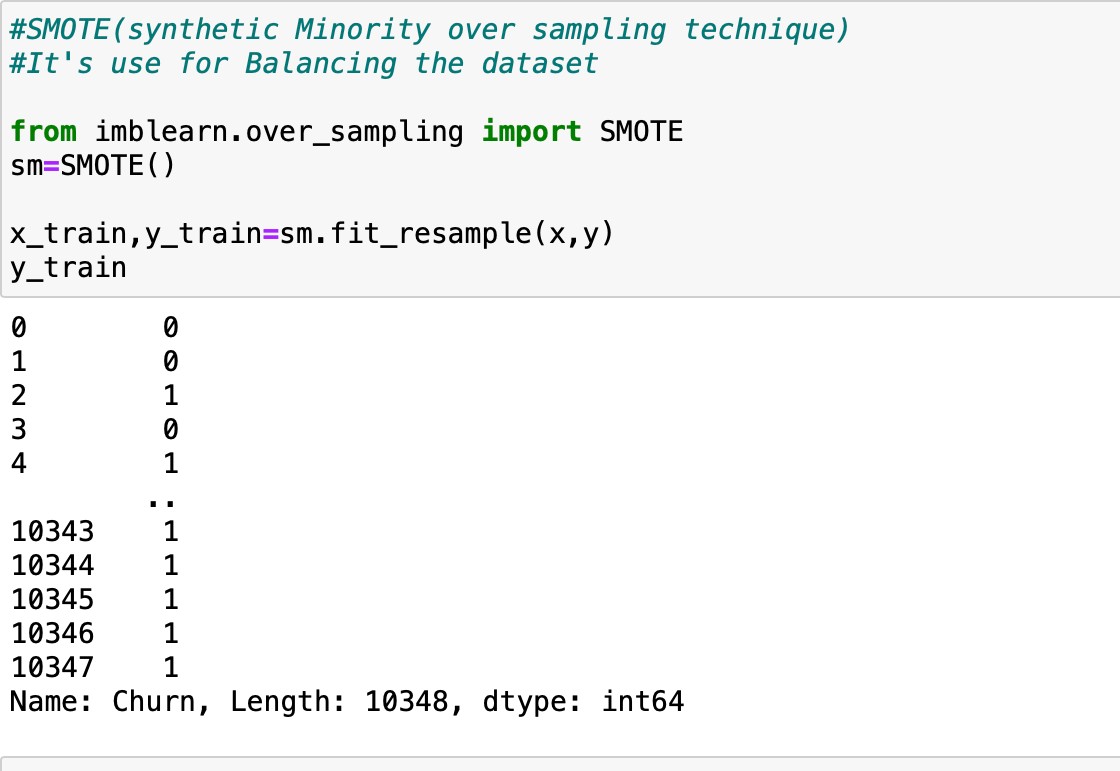
## Checking Variance Influence Factor



A variance inflation factor is a tool to help identify the degree of multicollinearity. The dependent variable is the outcome that is being acted upon by the independent variables—the inputs into the model. Multicollinearity exists when there is a linear relationship, or correlation, between one or more of the independent variables or inputs.

Here total charges was having via more than 10 so dropped it and moved further for balancing the dataset as our target column was bias.

## Smote for oversampling



Now when our data set is balanced, so moving forward for model building and choosing the best model.

## Building Machine Learning Models:

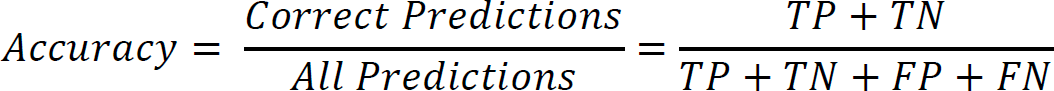
After making sure our data is good and ready we can continue to building our model. In this notebook I have build 7 model but will show best 4 different models with different algorithm. In this step we will create a baseline model for each algorithm using the default parameters set by sklearn and after building all 4 of our models we will compare them to see which works best for our case.

## Confusion matrix to evaluate our model

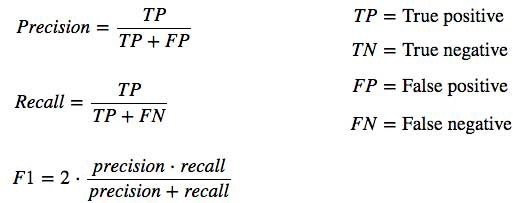
We will use 5 metrics below to evaluate models:

* + Accuracy: the proportion of true results among the total number

of cases examined.

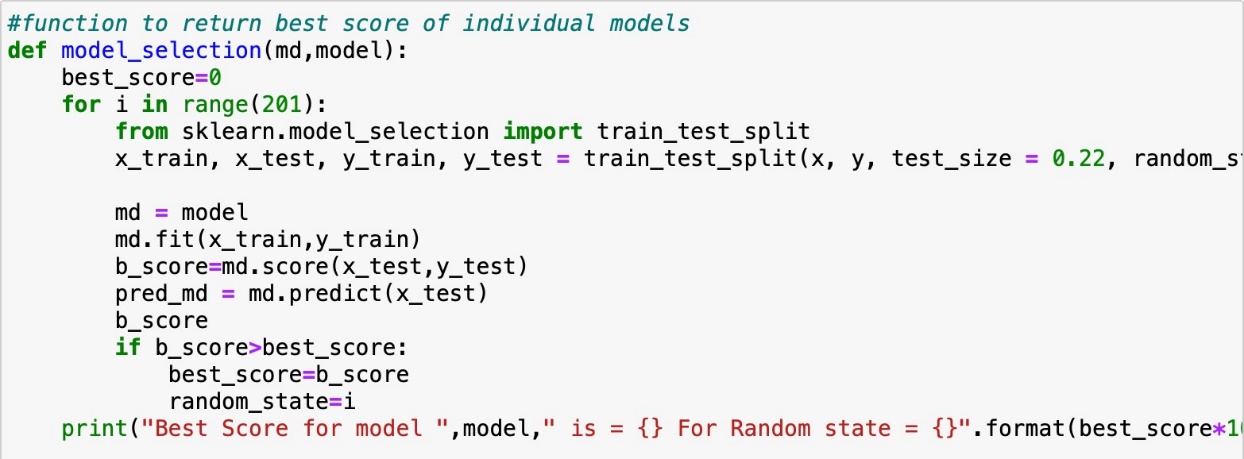
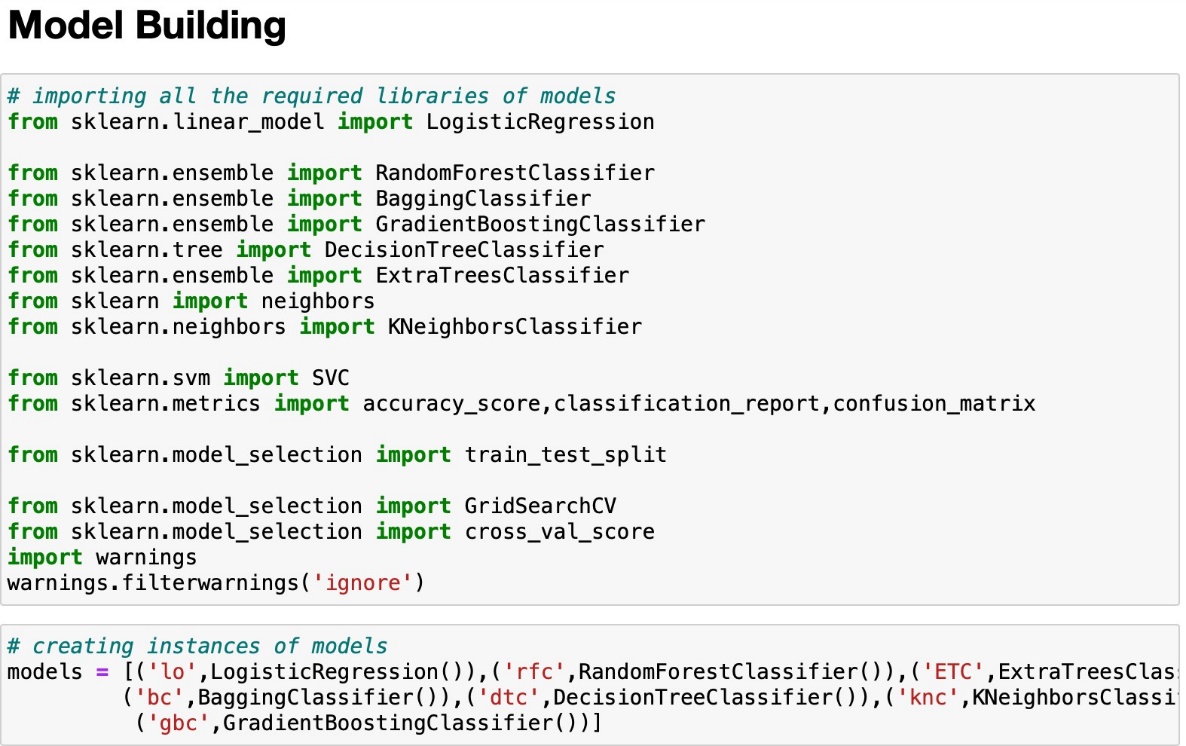


* + Precision: used to calculate how much proportion of all data that was predicted positive was actually positive.
  + Recall: used to calculate how much proportion of actual positives is correctly classified.
  + F1 score: a number between 0 and 1 and is the harmonic mean of precision and recall.



* + Area Under Curve (AUC): indicates how well the probabilities from the positive classes are separated from the negative classes.

Now the dataset is ready for next step which is splitting dataset at a test size of 0.22 for best accuracy score, for which a code is set to find best random state.



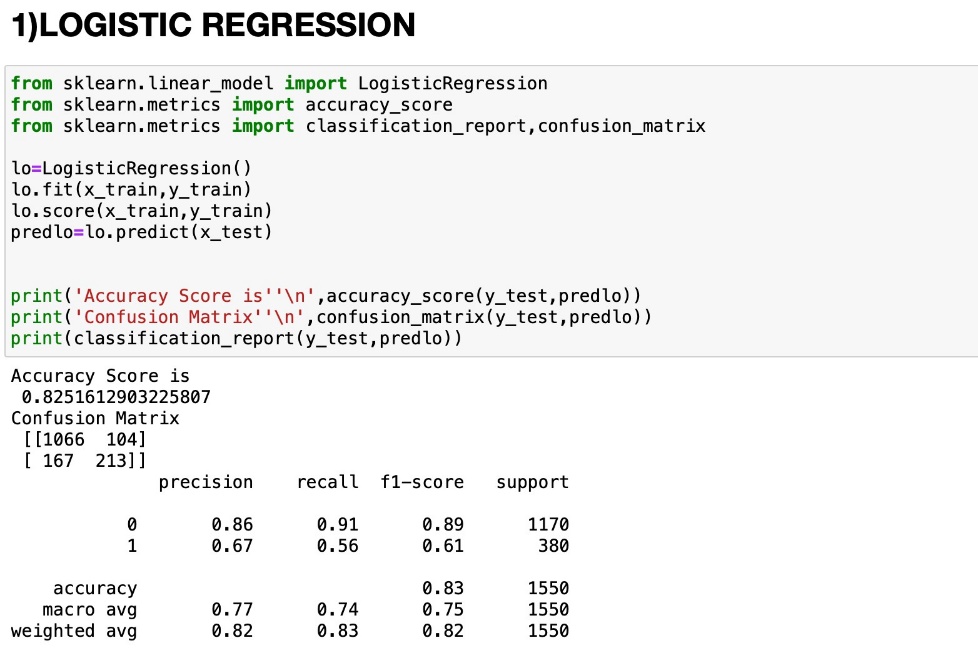
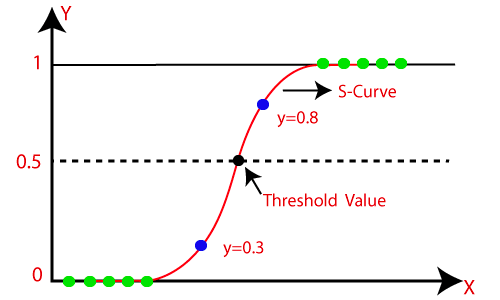


By running this code we get best random state for different models.

Now ,As I said in earlier par of this article I have build 7 model but will show only 4 model which gave best results :Logistic Regression, Random Forest, Support vector classifier and Gradient boosting .Before we start below are a simple definition of each algorithms and how they work. If you aren’t familiar with any of the said algorithm you should definitely try to read more in- depth explanation about them before you continue.

**Logistic Regression**

In statistics, the Logistic model is a statistical model that models the probability of one event taking place by having the log-odds for the event be a linear combination of one or more independent variables. In regression analysis, logistic regression is estimating the parameters of a logistic model.



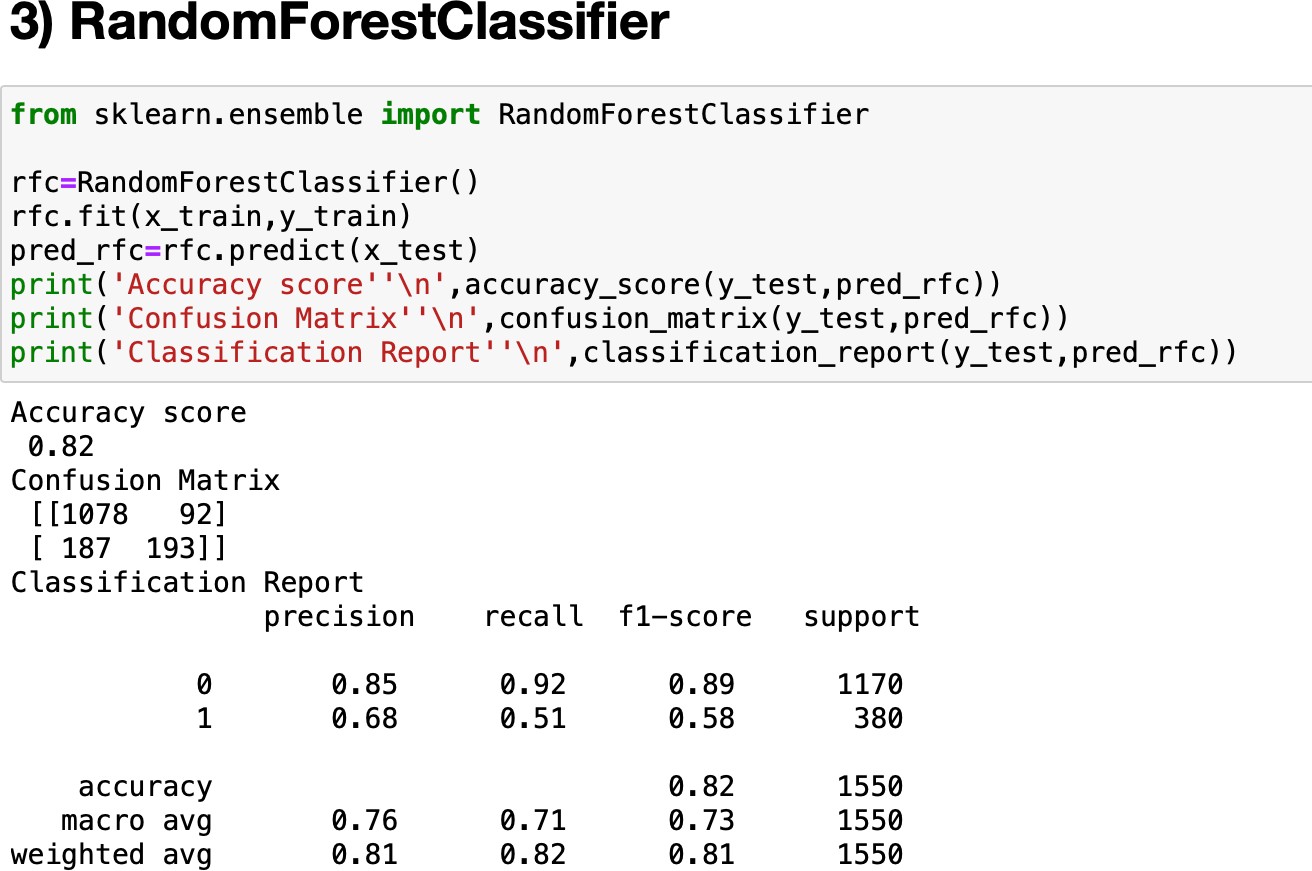
Accuracy score for Logistic Regression is 82.5%.

## RandomForest Classifier

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and e a s y t o u s e algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is.

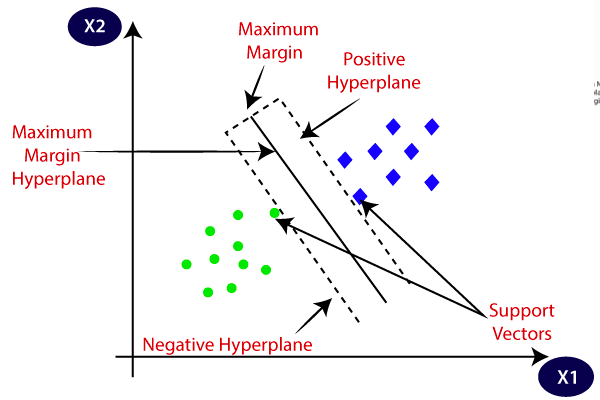
Random forests

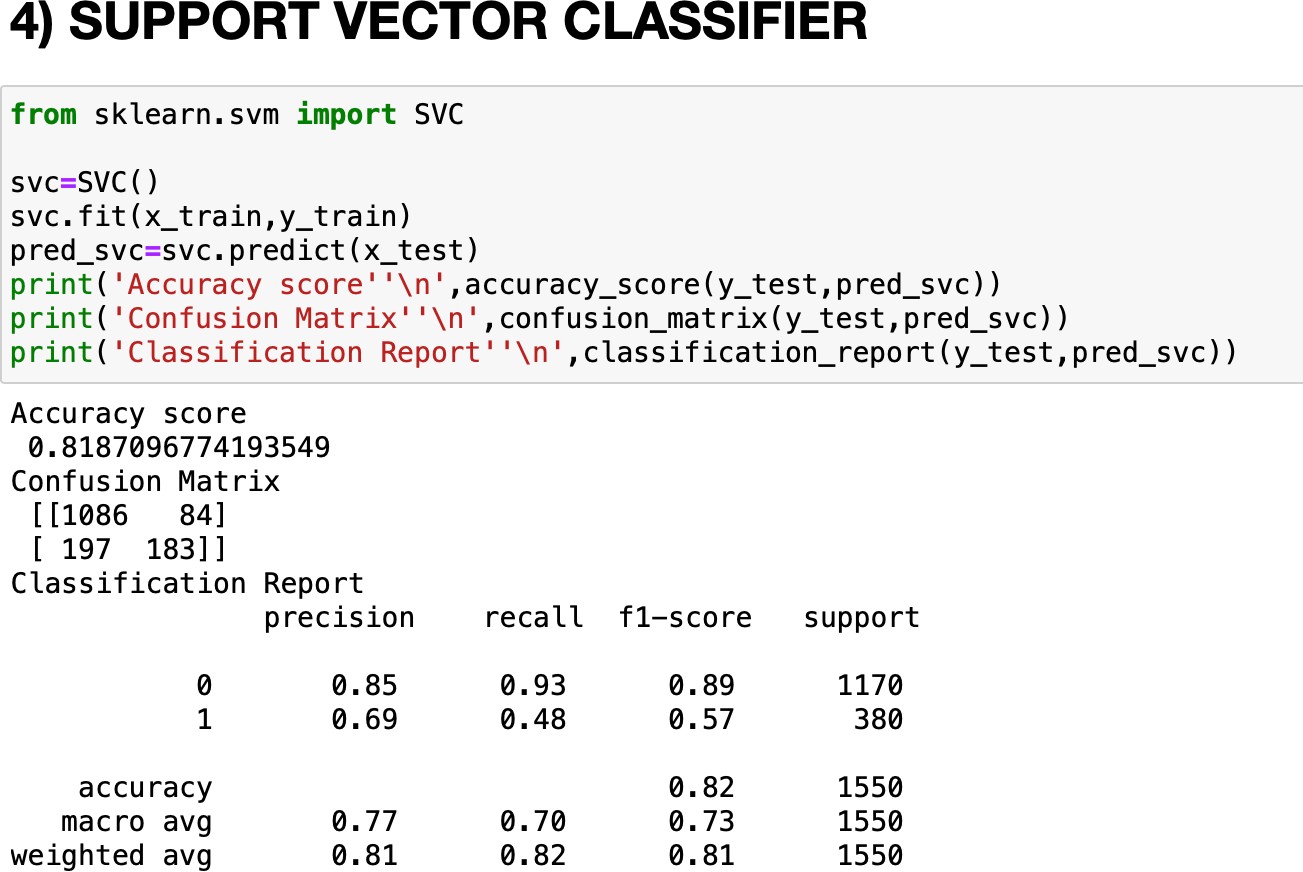
creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.



It also have good accuracy score of 82%.

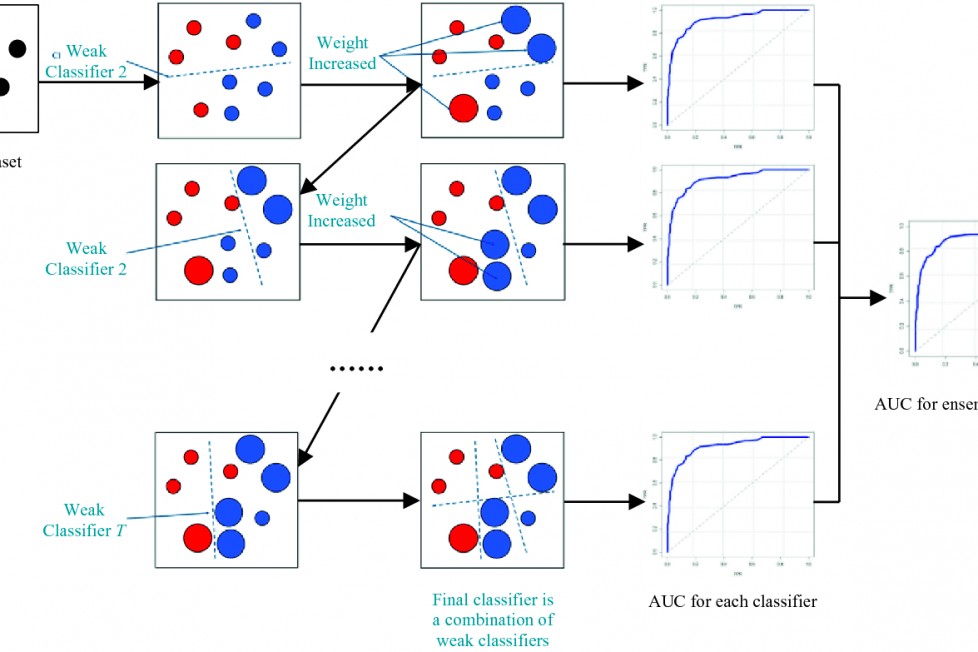
# SUPPORT VECTOR CLASSIFIER

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification or regression. It is mostly used in Classification and is called support vector classifier. In SVM the data is plotted in n (Number of features.) dimensional space where a hyperplane differentiates between classes very well. Then a maximum space is created for differentiating two classes this space is called margin.

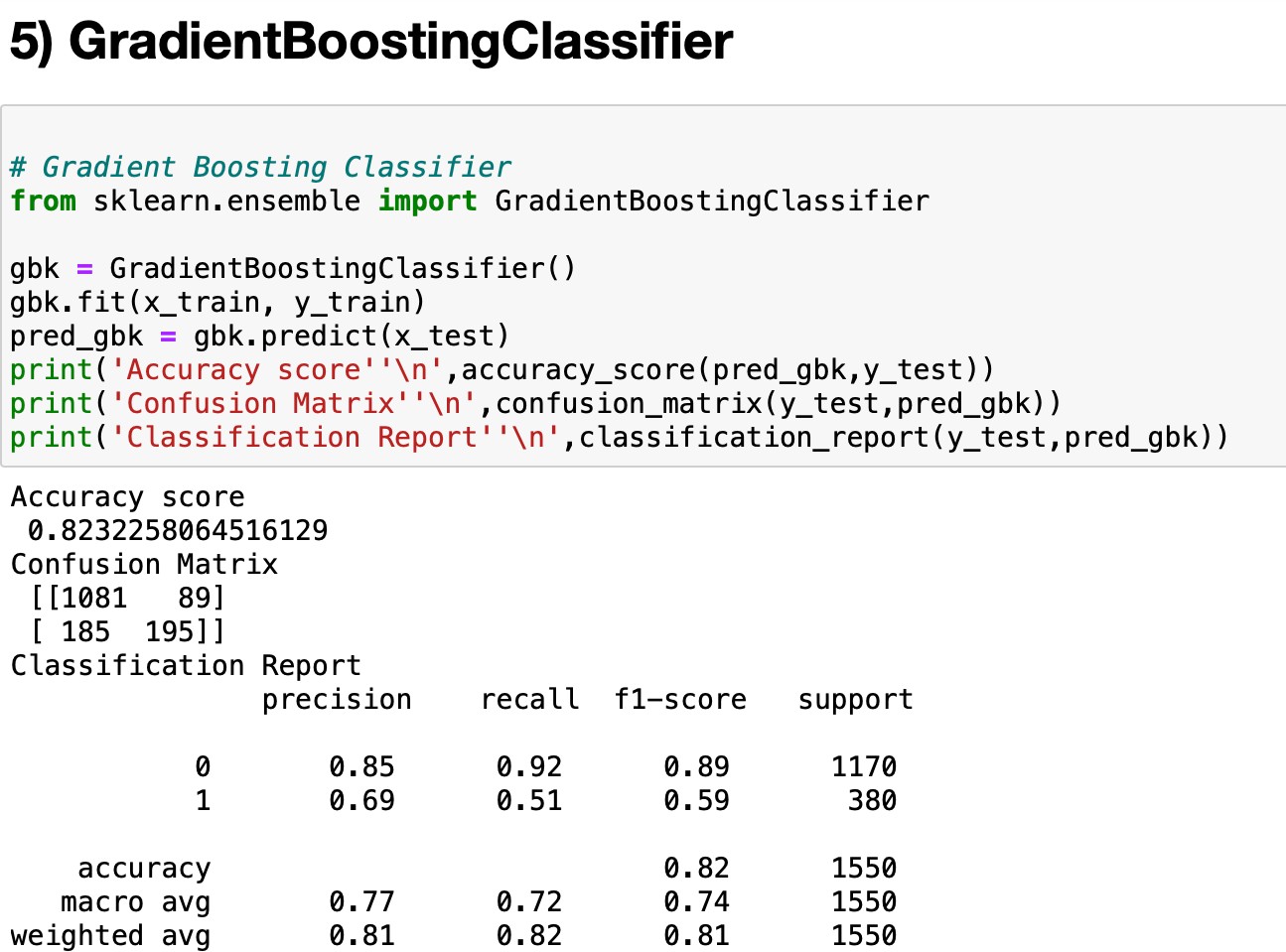


Thus SVM classifier gives us accuracy score of 81.87%

# GradientBoostingClassifier

Gradient boosting are the supervised machine learning algorithm that can be used for both regression and classification. It basically uses various weak learning models together and gives the best output. It has two necessary parts:1)a weak learner and 2) an additive component. The weak learner are decision trees and additive component are the gradient boosting model comes from the fact that trees are added to the

model over time with manipulating existing trees and also minimizes errors.



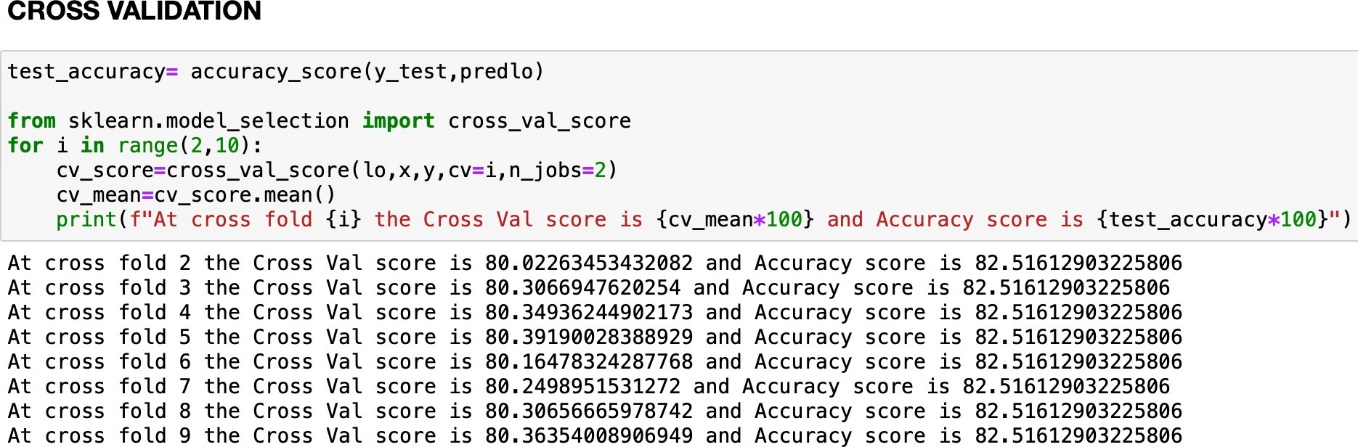
Here, accuracy score is 82.32%

## Cross-validation

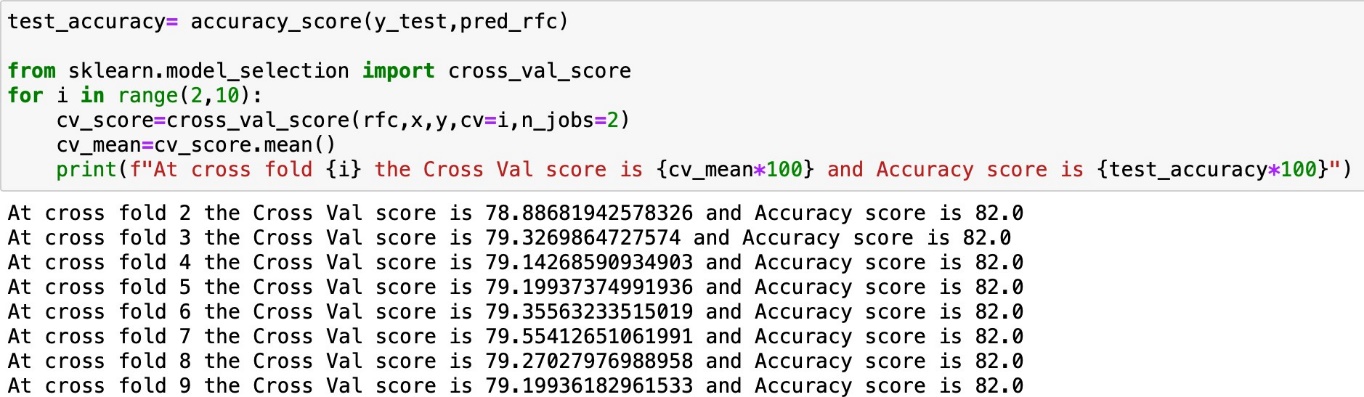
Cross-validation is a statistical method used to estimate the performance (or accuracy) of machine learning models. It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited.

Here we could check the best cv score and accuracy score which gives us the best model knowledge.

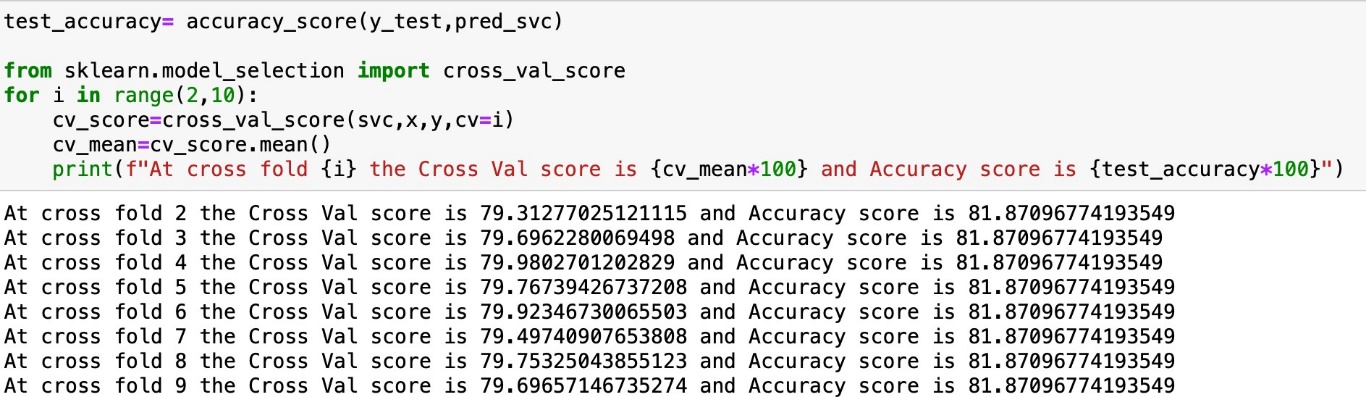
As best model is the one which has the least difference between accuracy score and cv score. Below are the cross validation results for all above model with best cross fold value:



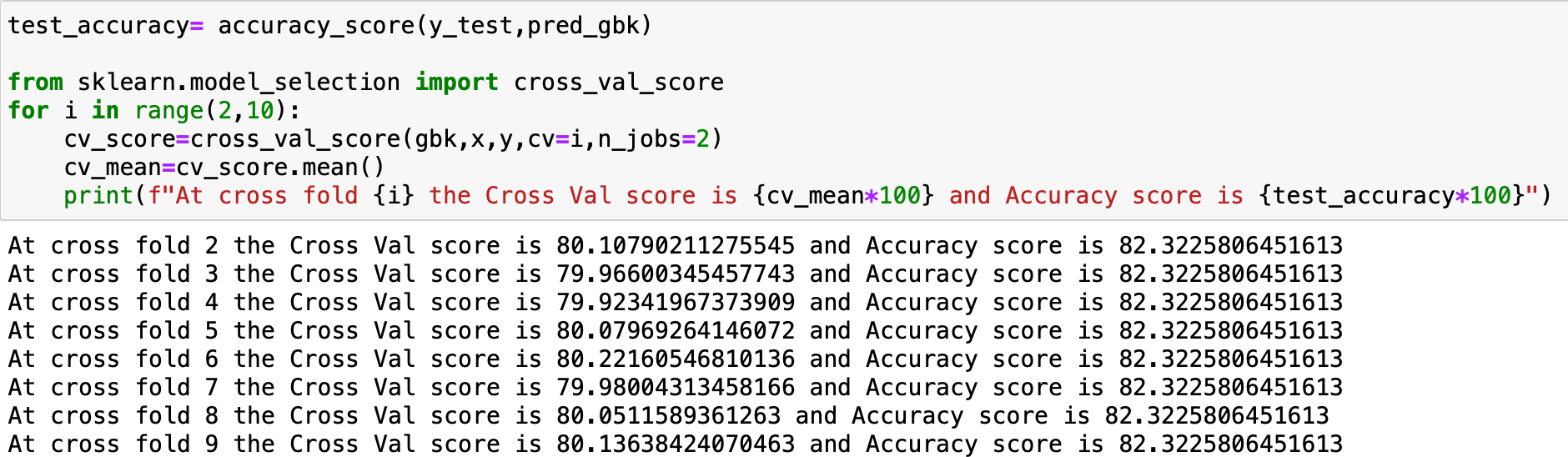
# Logistic Regression



# Random Forest



# Support Vector Classifier

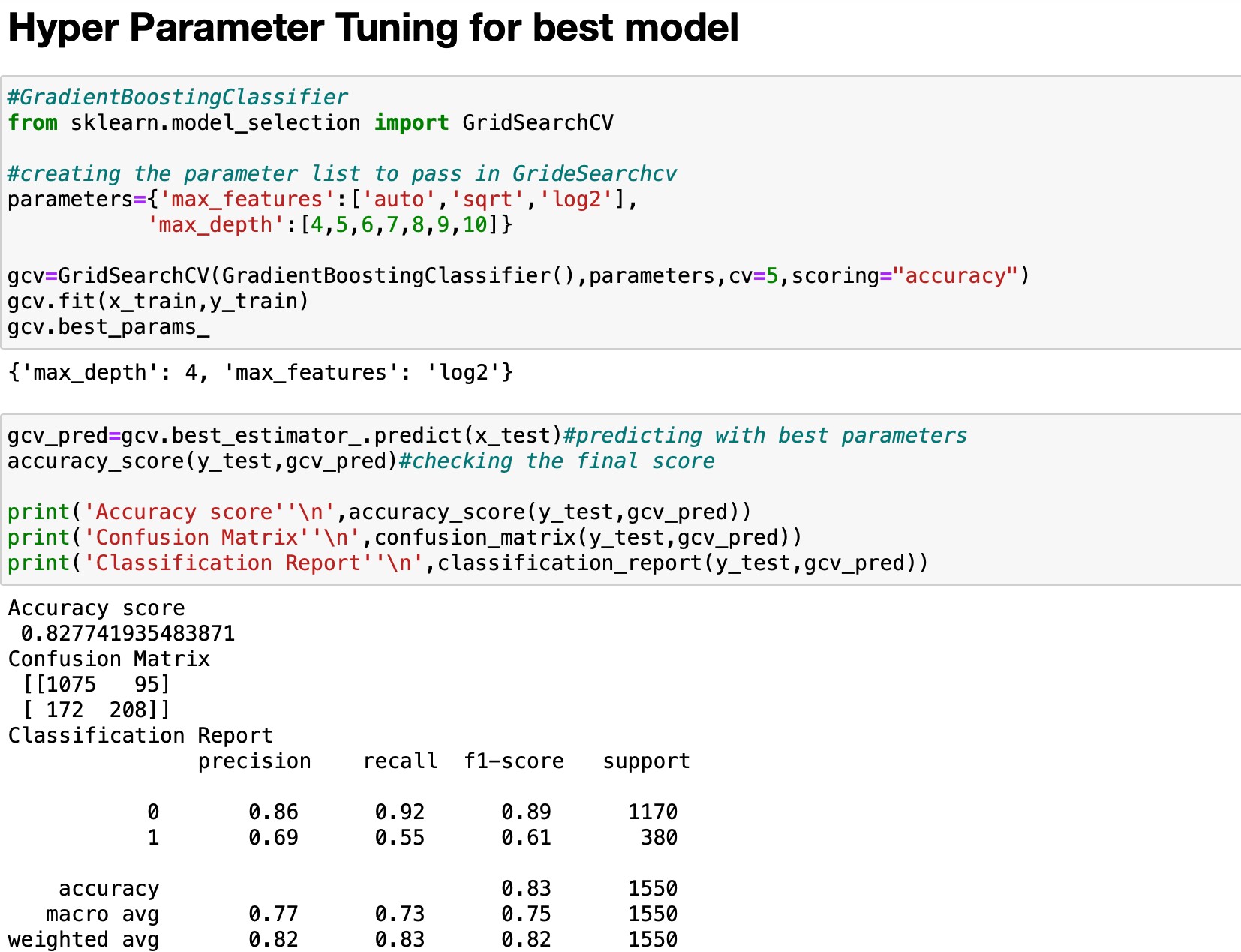


# Gradient Boosting Classifier

Now all models are performing good ,picking the best for hyper- parameter.i.e, **Gradient Boosting Classifier** as it has the least difference between cv score and accuracy score.

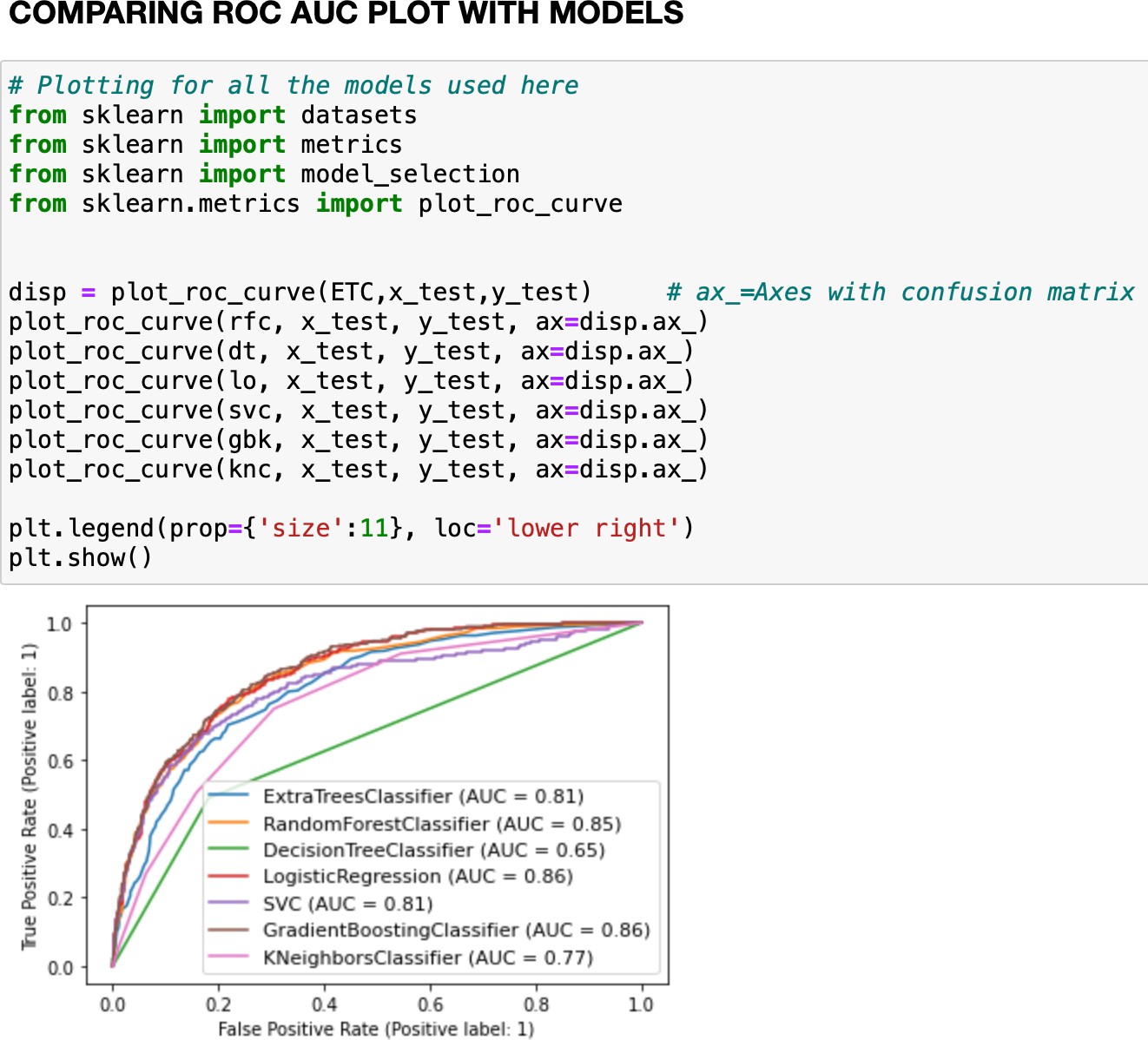
# Hyper Parameter Tuning for best model

Thus Gradient Boosting showed the best result after hyper parameter and so plotting arc roc curve for all models.



Thus after hyperparmeter tuning accuracy score for gradient boosting classifier is increased to 82.72%.

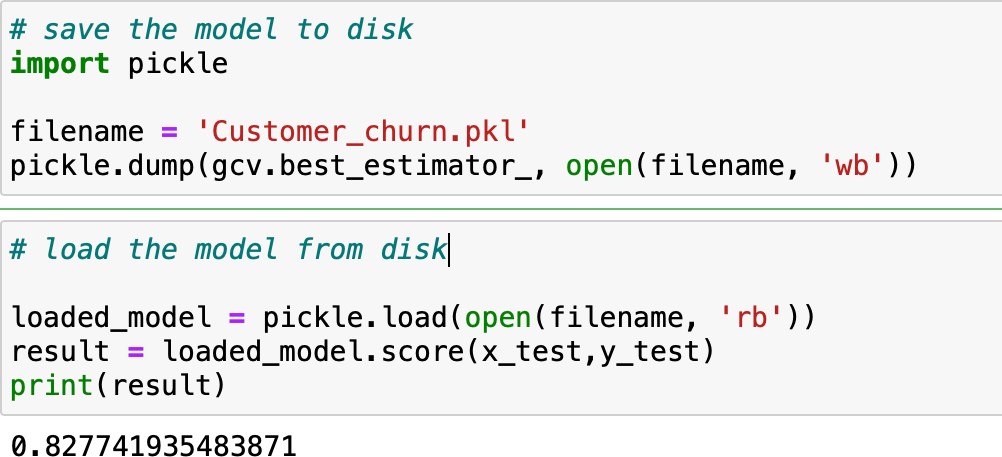
## Arc roc curve:



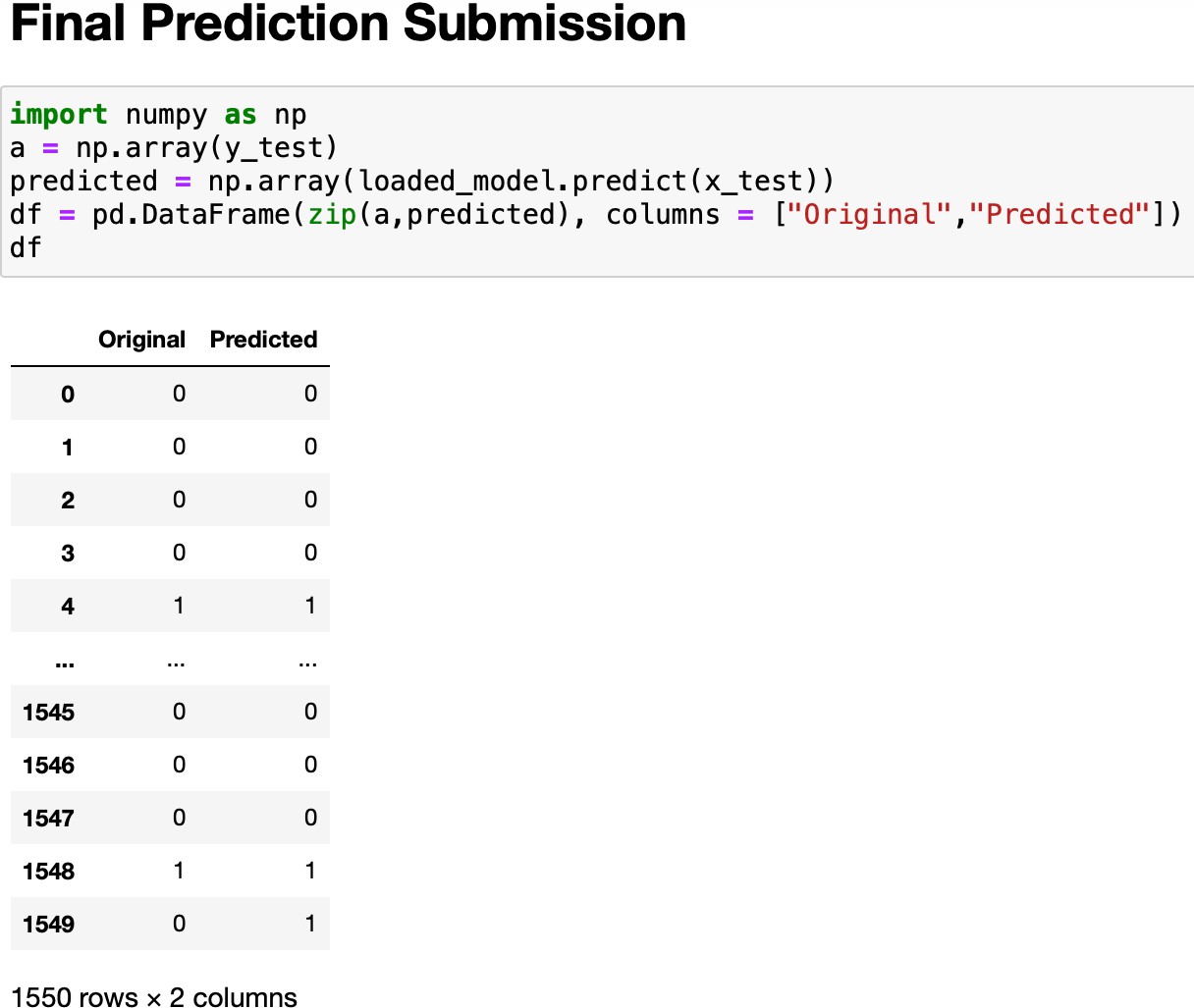
Thus in arc roc curve also Gradient Boosting is having highest score. So now saving the model with it.

**S**

# aving the best Model



Saving the model with pickle and then loading again for prediction.



Final submission is showing the comparison between original and production for customer churn.

## Concluding Remarks:

* The accuracy of the technique is critical to the success of any proactive retention efforts. If the Marketer is unaware of a customer about to churn, no action will be taken to retain that customer.
* To succeed at retaining customers who are ready to abandon your business, Marketers & Customer Success experts must be able to predict in advance which customers are going to churn and set up a plan of marketing actions that will have the greatest retention impact on each customer. The key here is to to be proactive and engage with these customers.
* Special retention-focused offers or incentives may be provided to happy, active customers, resulting in reduced revenues for no good reason
* Bringing it all together, predicting customer churn is important. Effective action can be taken to retain the customer before it is too late. The ability to predict that a customer is at high risk of churning while there is still time to do something about it represents a huge additional potential revenue source for every online business.

Thank you