Customer Churn Analysis

Problem Statement:

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.

When start the project first of all import the necessary libraries for visualizing reading the whole data some plotting graphs libraries are pandas,numpy,matplotlib,seaborn these are the packages used of datacleaning visualizing , EDA in a dataset.

First import necessary libraries and read csv file from github

Pandas is open-source library tool which provides high performance data analysis tool by its powerful data structures.

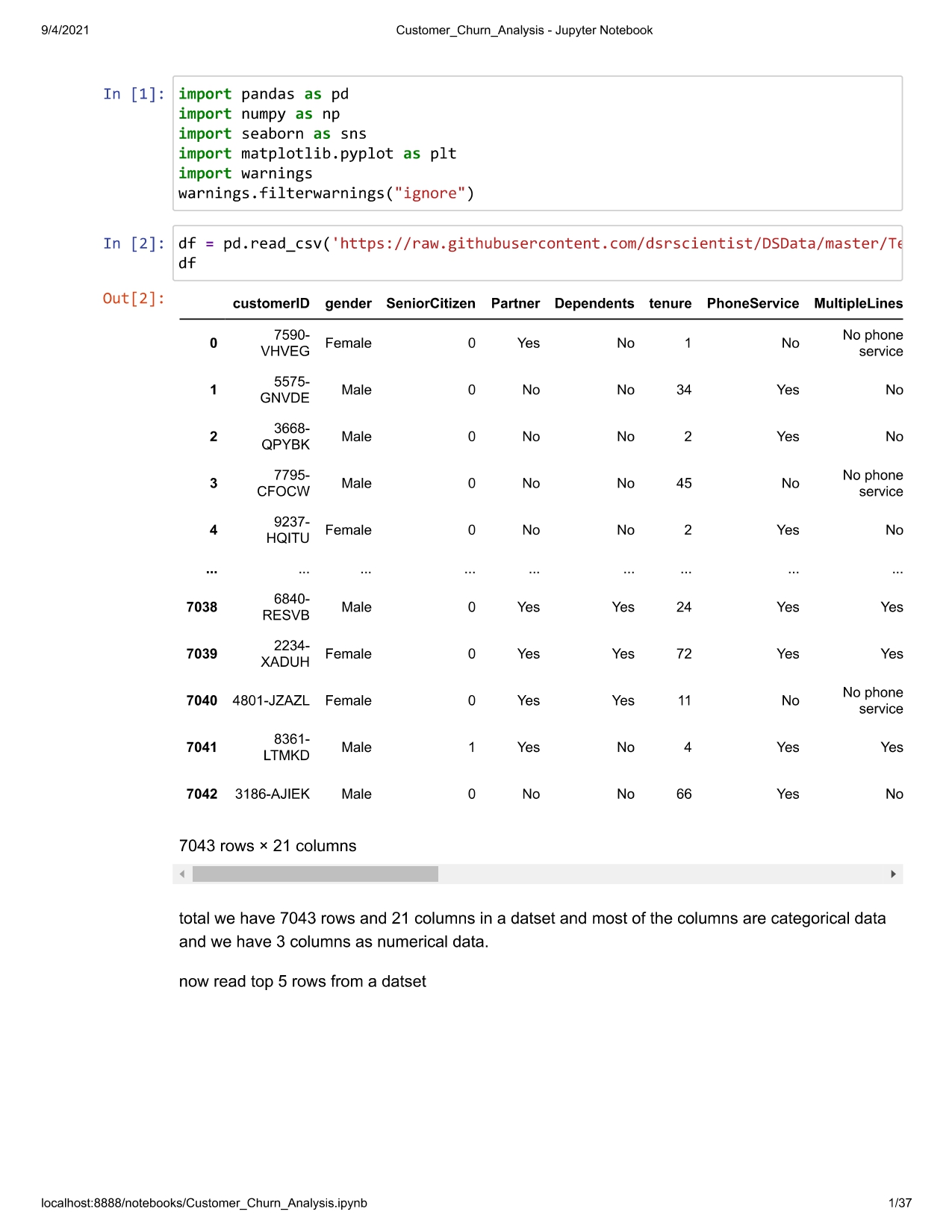
It helps to shorten the procedure of handling the data with extensive set of features.

NumPy is most used package for scientific computing for multi-dimensional array of objects.

Seaborn and Matplotlib is used here as a visualization library for the stunning plot to understand the data in a better way.

warning occurs when there is some obsolete of certain programming elements, such as keyword, function or class, etc

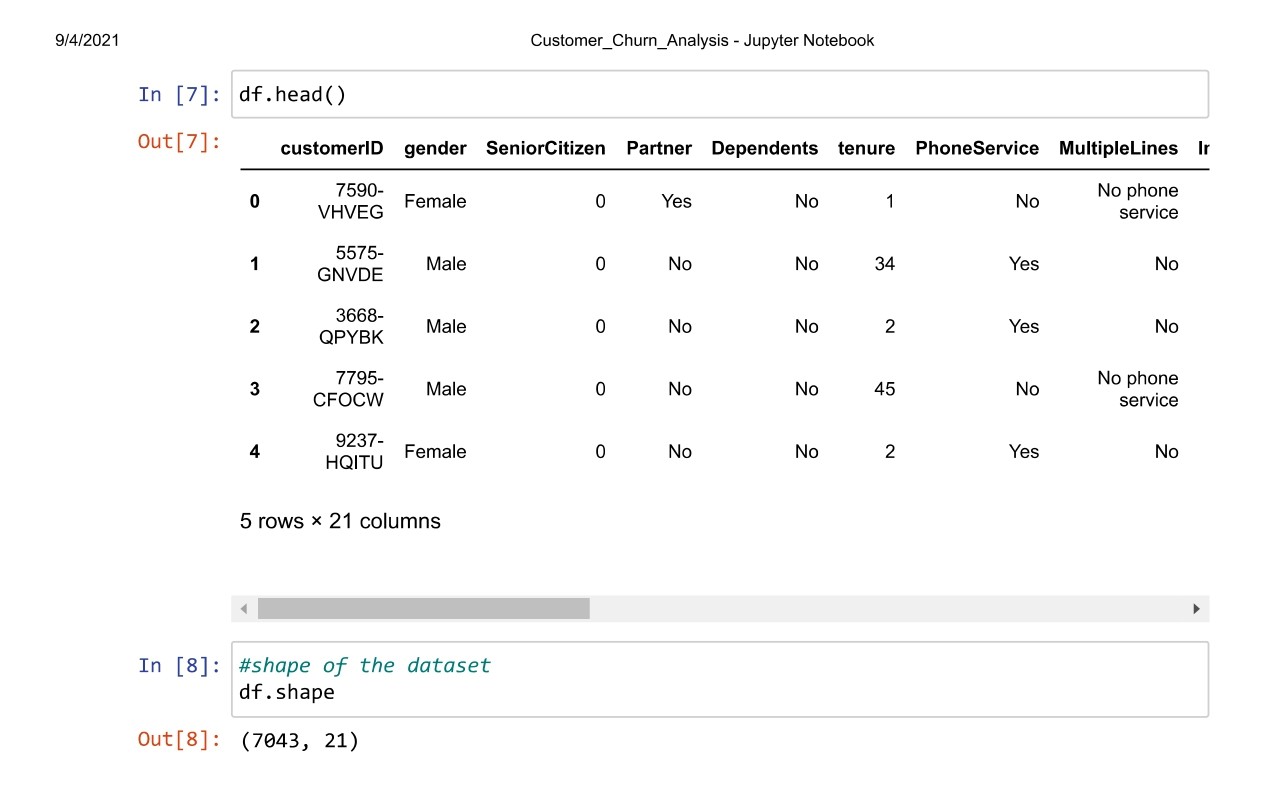
lets import all the libraries in above discussed and read csv file too



total we have 7043 rows and 21 columns in a datset and most of the columns are categorical data and we have 3 columns as numerical data.

Loading the dataset and head () method will display the first 5 data from the dataset whereas tail () method displays the last 5 data from the dataset.

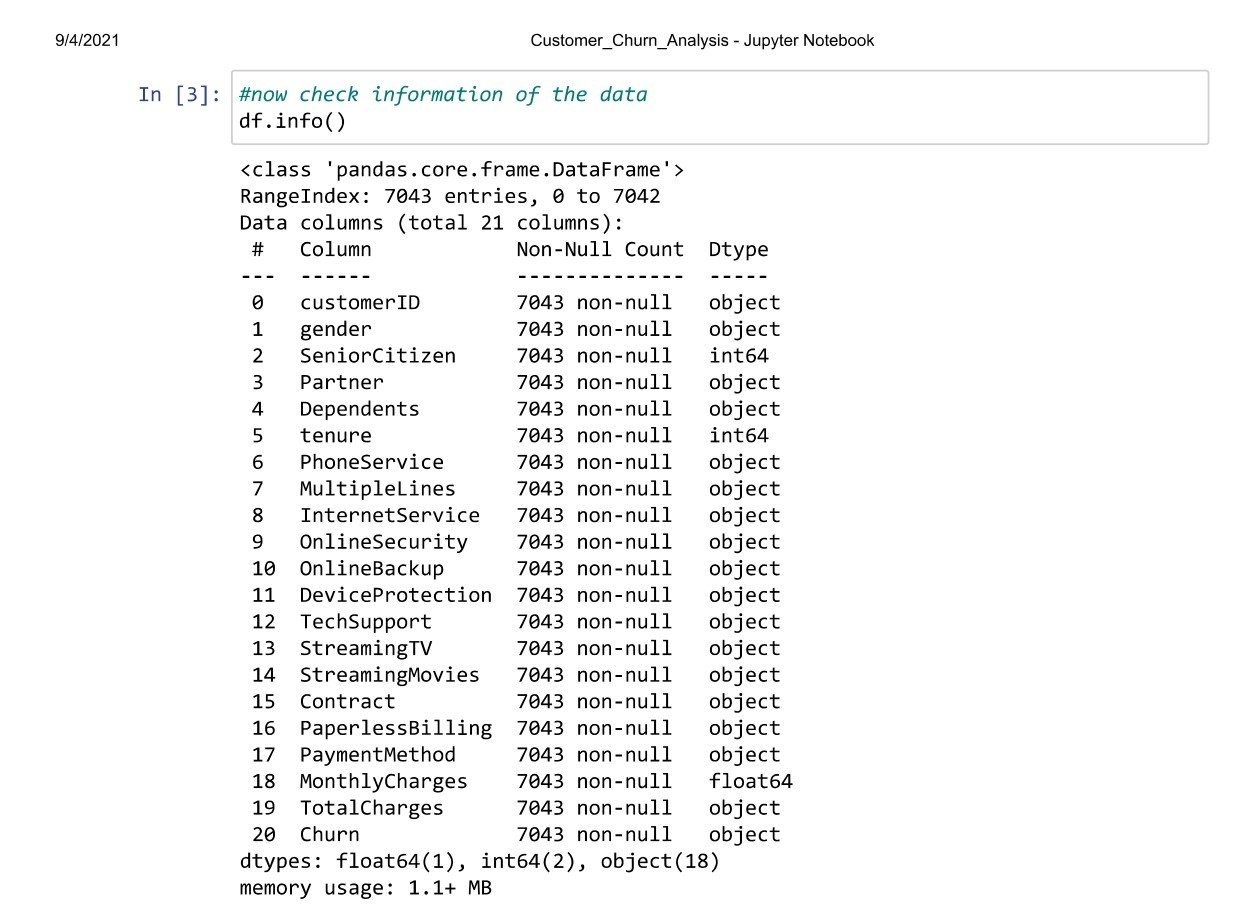
And check shape of the dataset



Shape of the dataset is 7043 rows and 21 columns

We will check whether dataset is having null values or not.

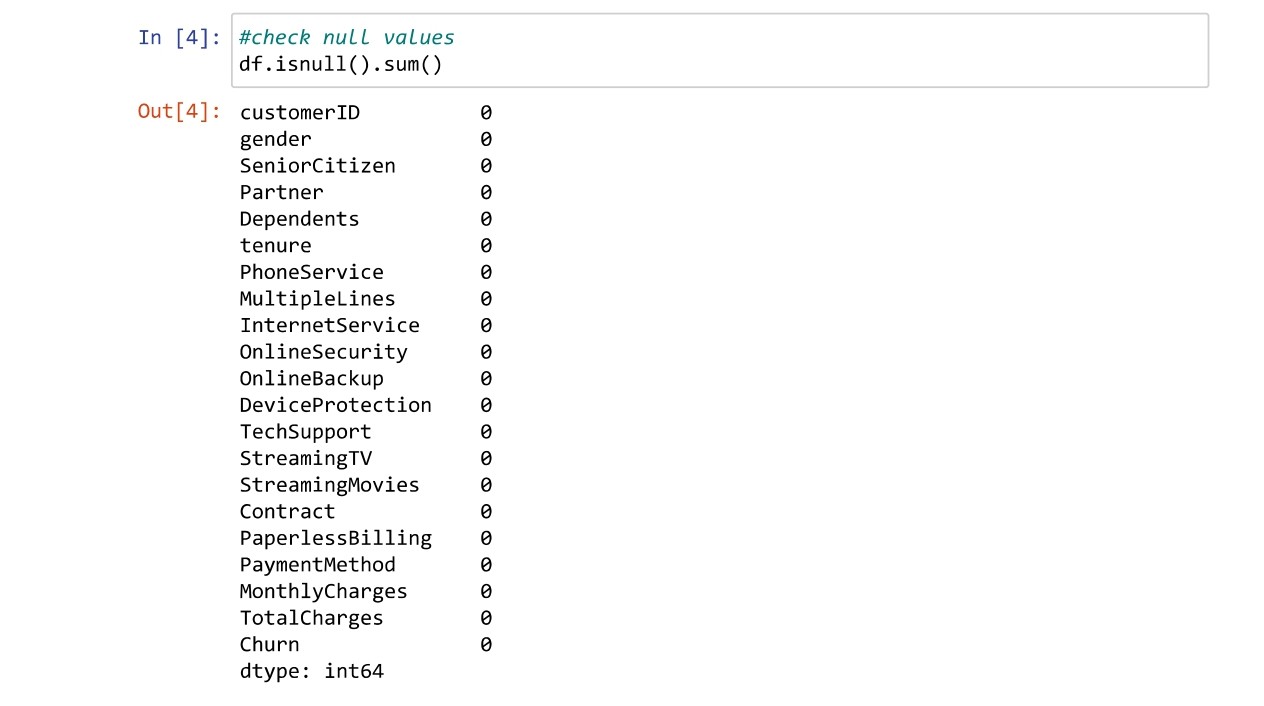
The below given info() method describes the each column in a dataset and gives information about the non null count and index entries which is rows and column values,memory usage and thier datatypes.



in above information tells here we have good sounds to hear there are no null values in a dataset and one float dtype,2 integer dtype and 18 objects,it tells also shape of the data we have total 7043 rows and 21 columns

As we have categorical data's in our given dataset and our target variable is also a category variable (Yes / No) ,So our problem is Classification Problem(Binary).

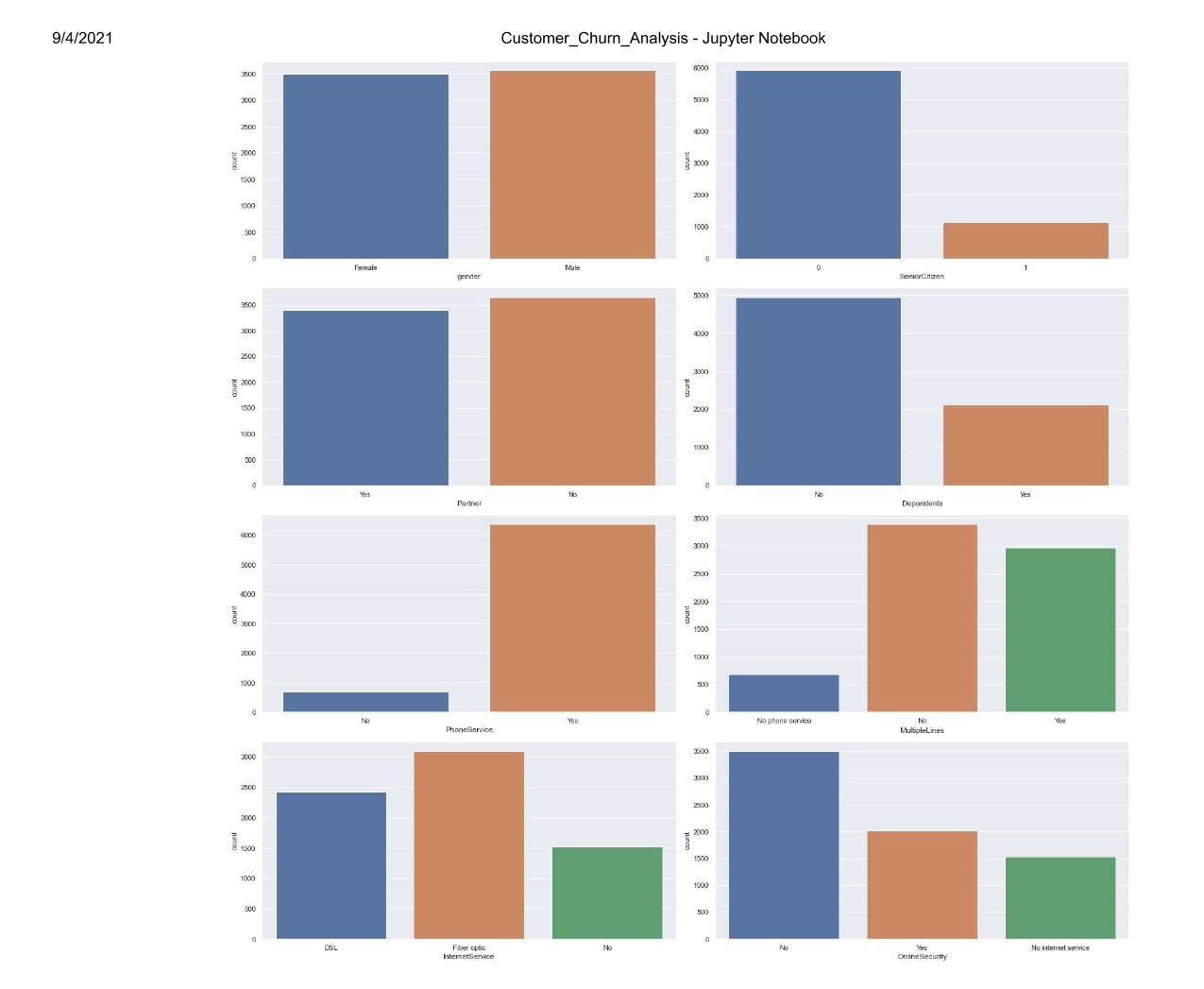
Check null values with isnull()

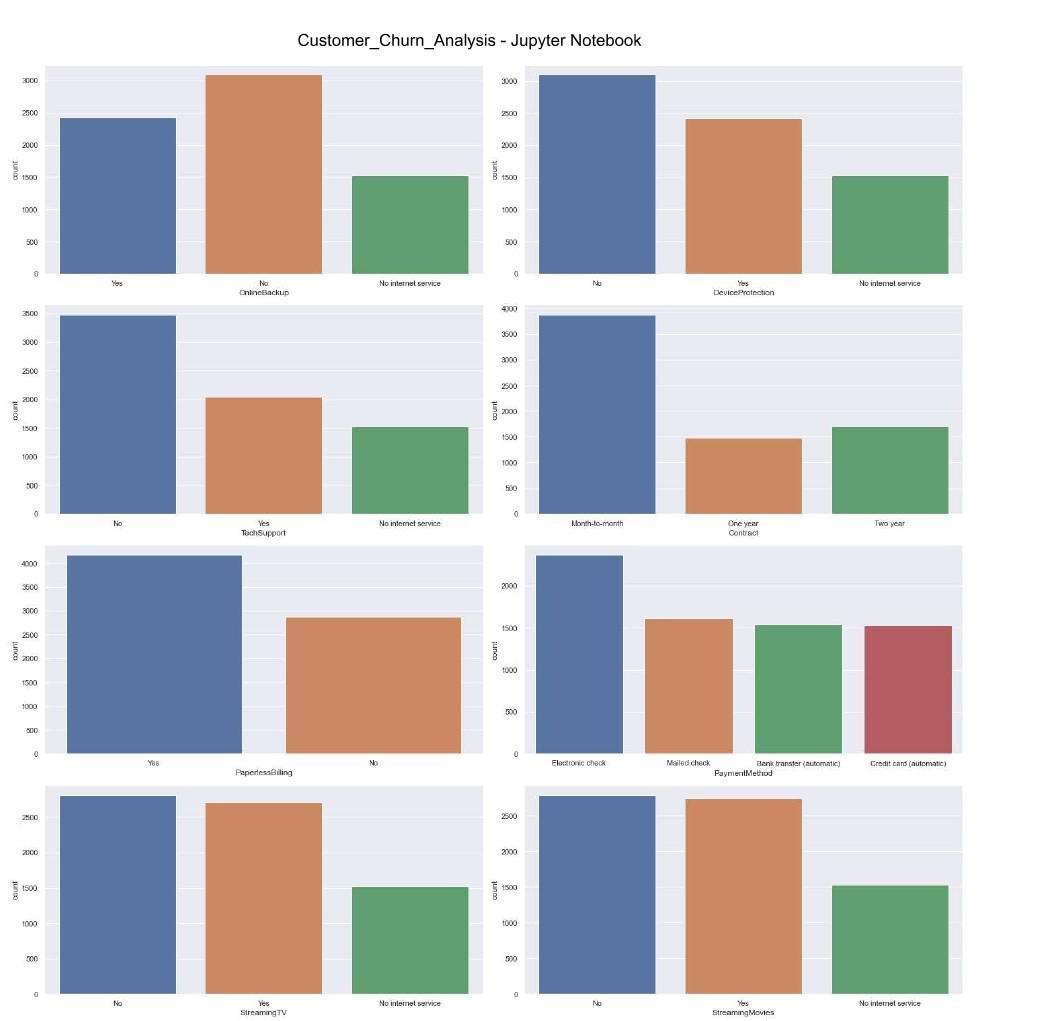


**Data Analysis.**

**EDA**







We can see that GENDER has two values Yes or No and that shows that customer is a male /female.

SENIOR CITIZEN already has numerical value which is 1 or 0 that explains whether the customer is a sr.citizen or not

Column PARTNER has Yes or No which describes whether customer has partner or not.

DEPENDANTS - explains whether customer has any dependent or not and the values are Yes or No.

Customer has any PHONE SERVICVE or not that shows Yes or No.

Customer has MULTIPLE LINES or not and has values as Yes or No or No Internet services.

Same goes for PHONE SERVICE as well as it has DSL, Fiber Optic, No

Whether Customer has online security or not and it has Yes or No or No Internet services.

OnlineBackup, StreamingMovies, StreamingTV,TechSupport , DeviceProtection has values of Yes, No, No Internet Service and the customers has either one of the values.

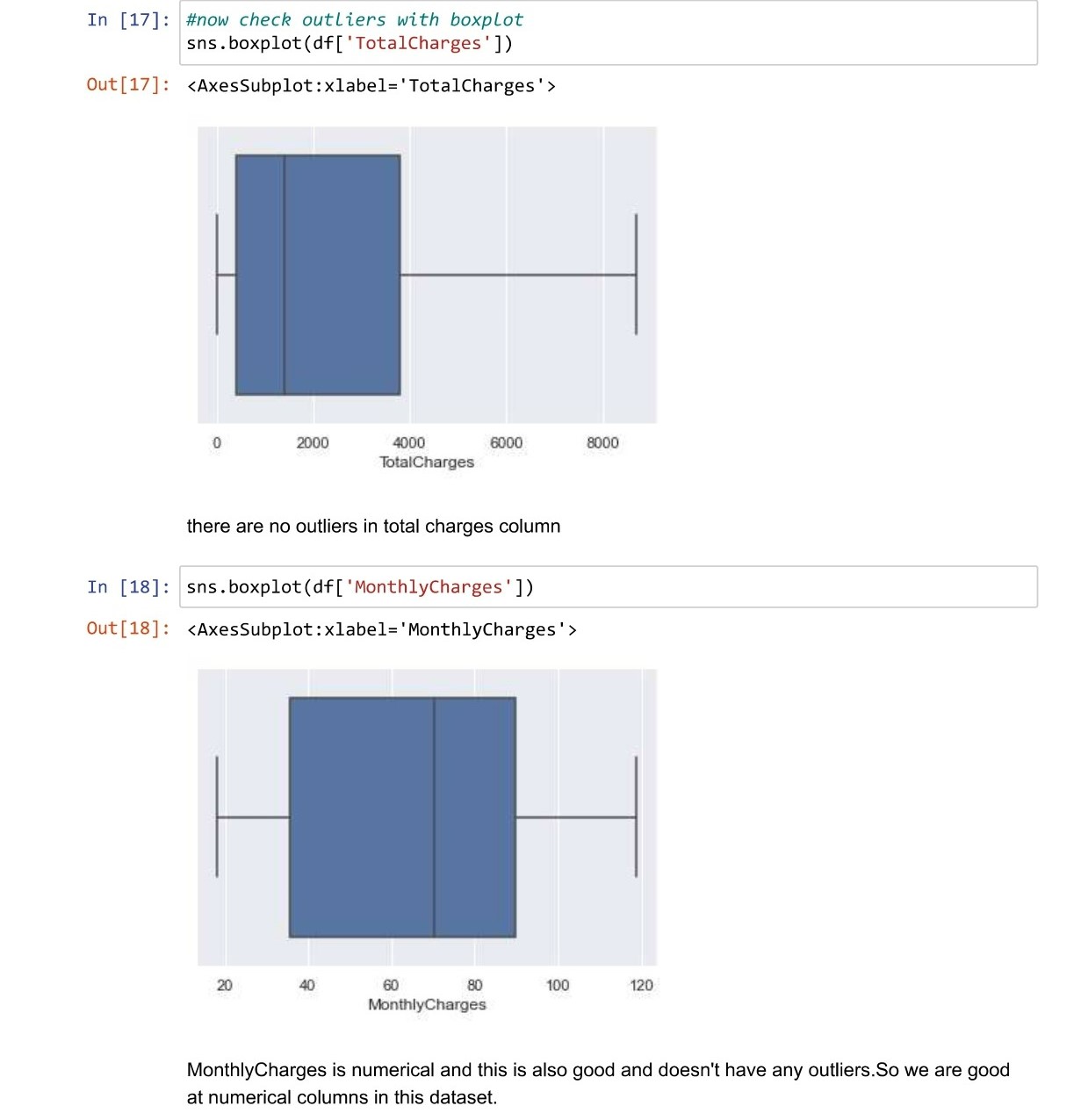
Other than that, PaperlessBilling explains that whether customer has paperless billing or not which is yes / No.

Also we have PaymentMethod which has options fopr customers to pay their bill through any one of the ways - Electronic check, mailed check, bank Transfer (automatic), Credit card(automatic)

Contract explains the time period for the customers with the company - Month-to-Month, One Year, Two Year.

**Pre-Processing Pipeline**.

From box plot,there were no outliers in dataset as well



Checking target variable is balanced or imbalaced and above discussed about binary values cobert into 1&0

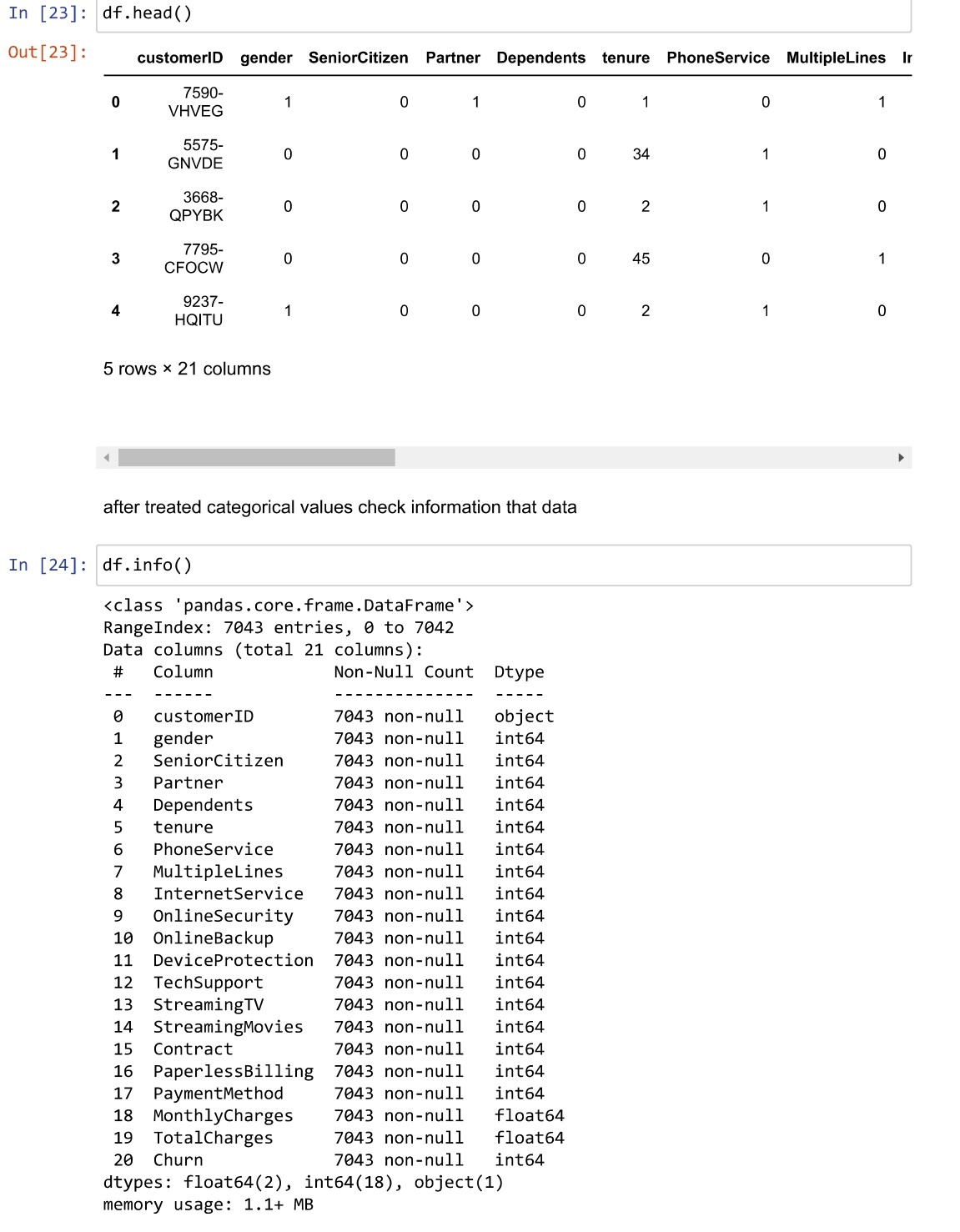


Our target varaiable imbalced data and do balance to use some teccniques as well

Applying label Encoder technique to convert categorical into numerical values

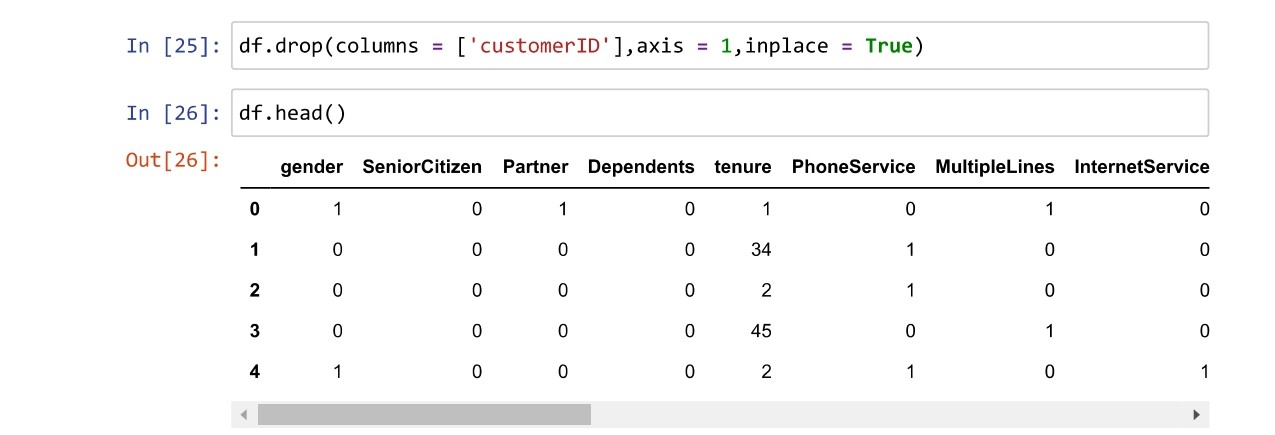


Check the data set after prepeocessing anf information of the dataset



Splitting x and y variables,

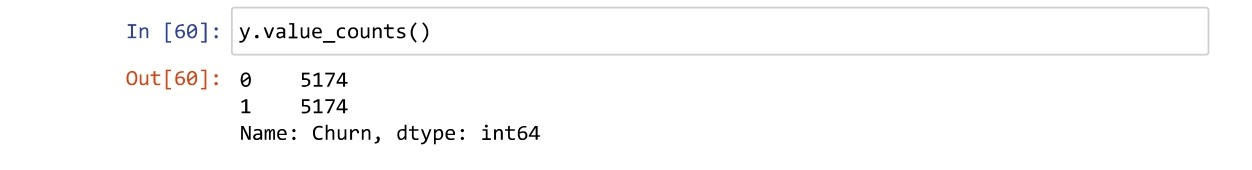
I am dropping CustomerID as it won't have any effect on the dataset.



**Imbalanced Data**

As our target variables are im-balanced as discussed earlier, Using SMOTE () method to balance the classes.

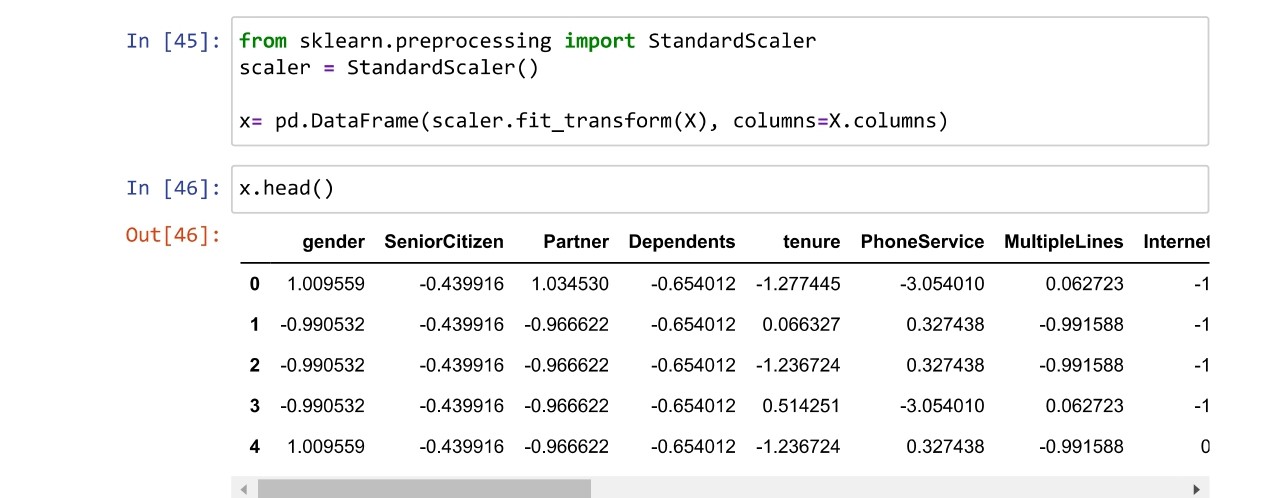




Dataset is balanced and counts are equal

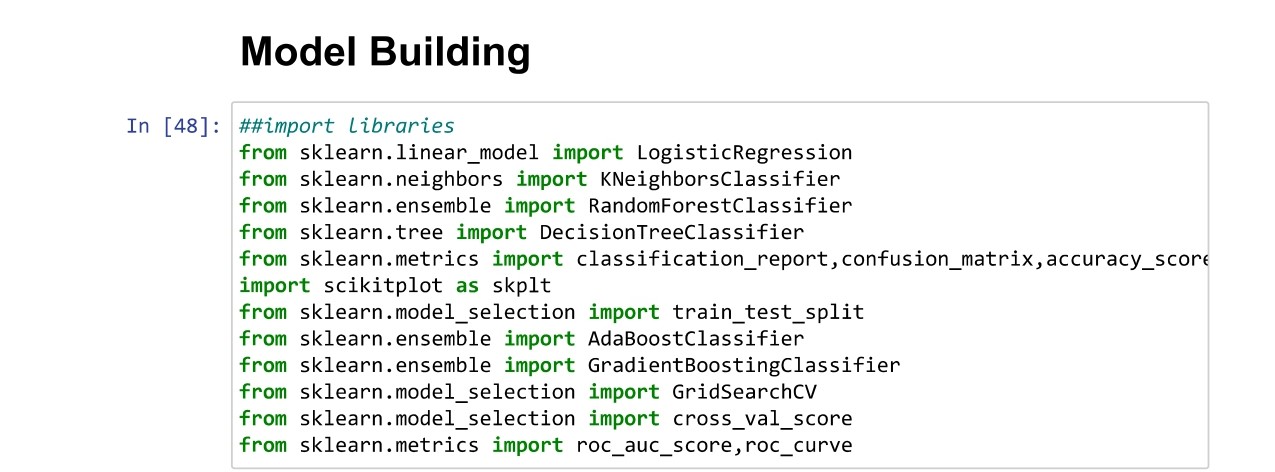
Scaling

As dataset need to get scaled to get max accuracy ,standard scaler was taken help to get out dataset properly scaled before model building



**Building Machine Learning Models**.

Importing necessary libraries for model building



For train test split 75% data for training and remaining 25% for testing purpose

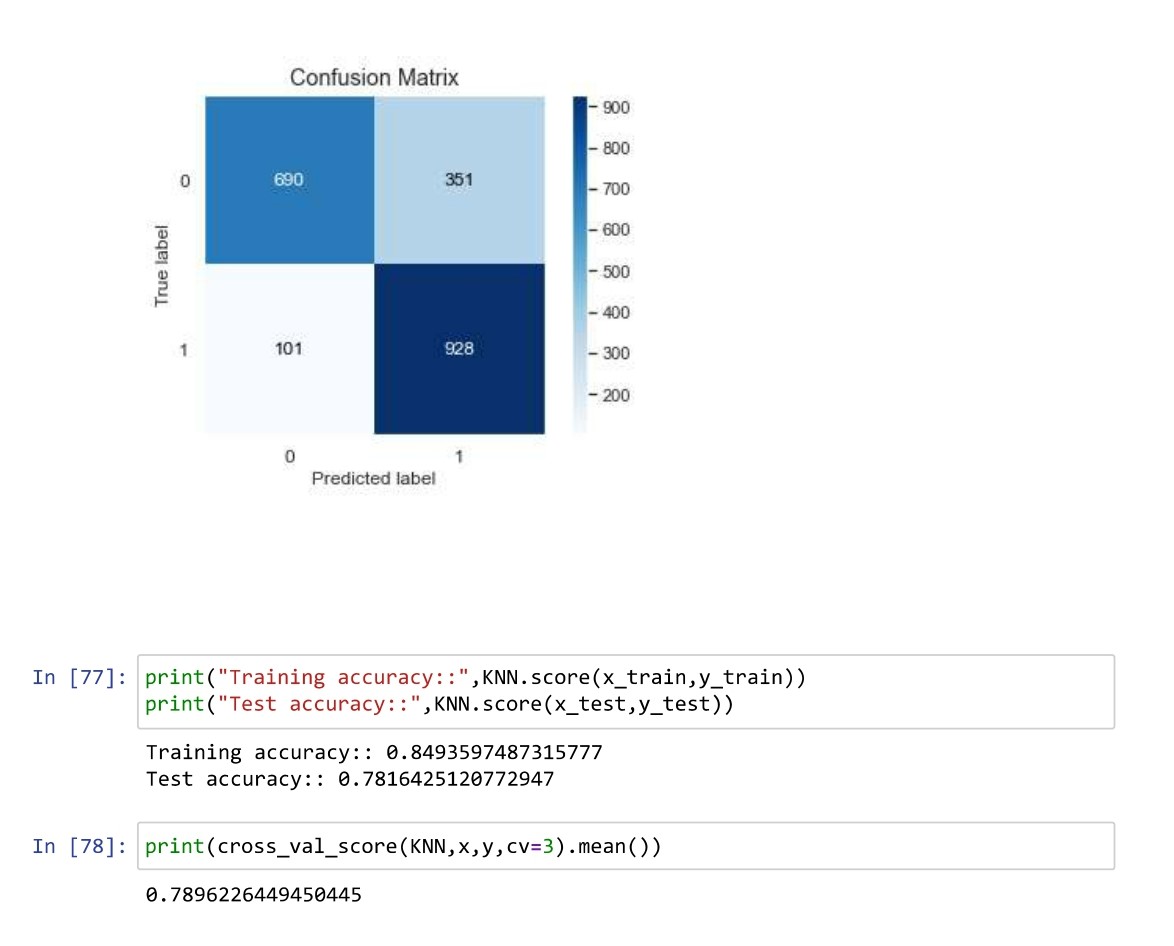


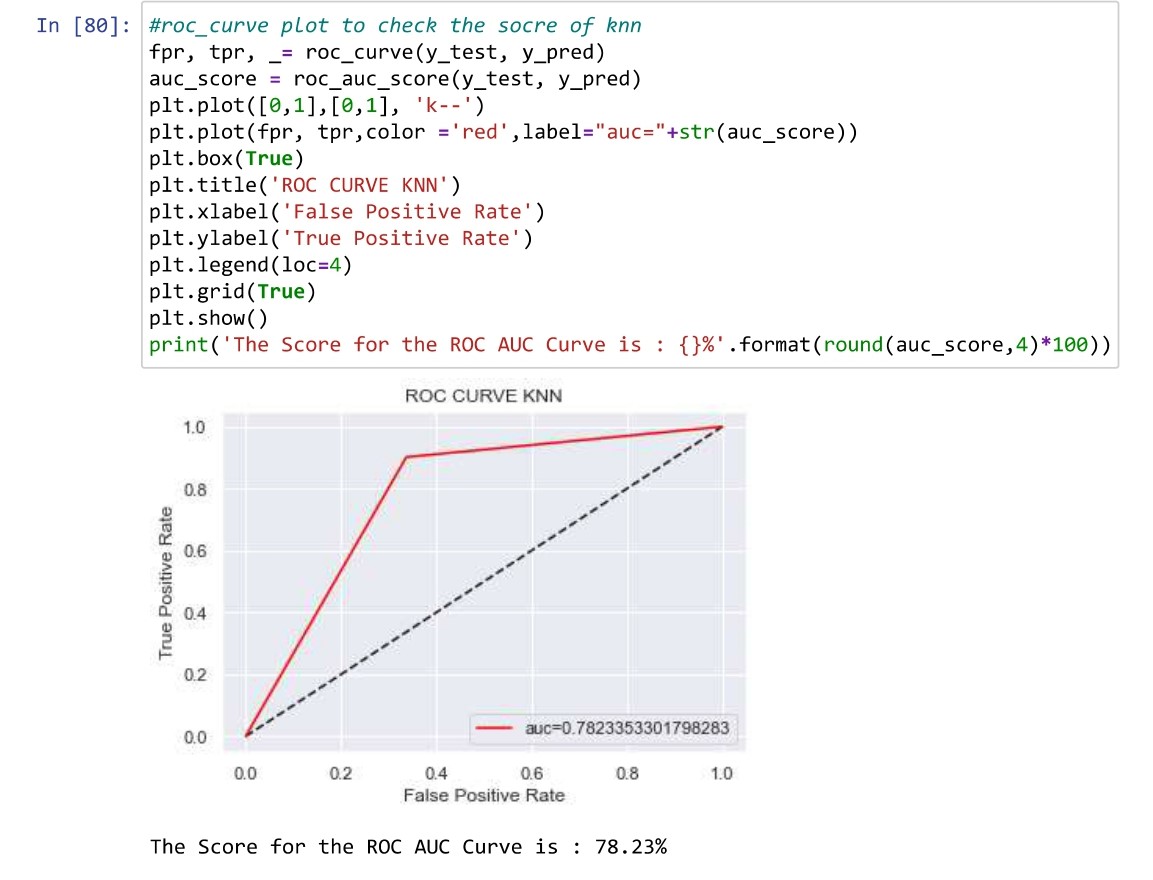
I tried different models

**KNeighborsClassifier-**

KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).

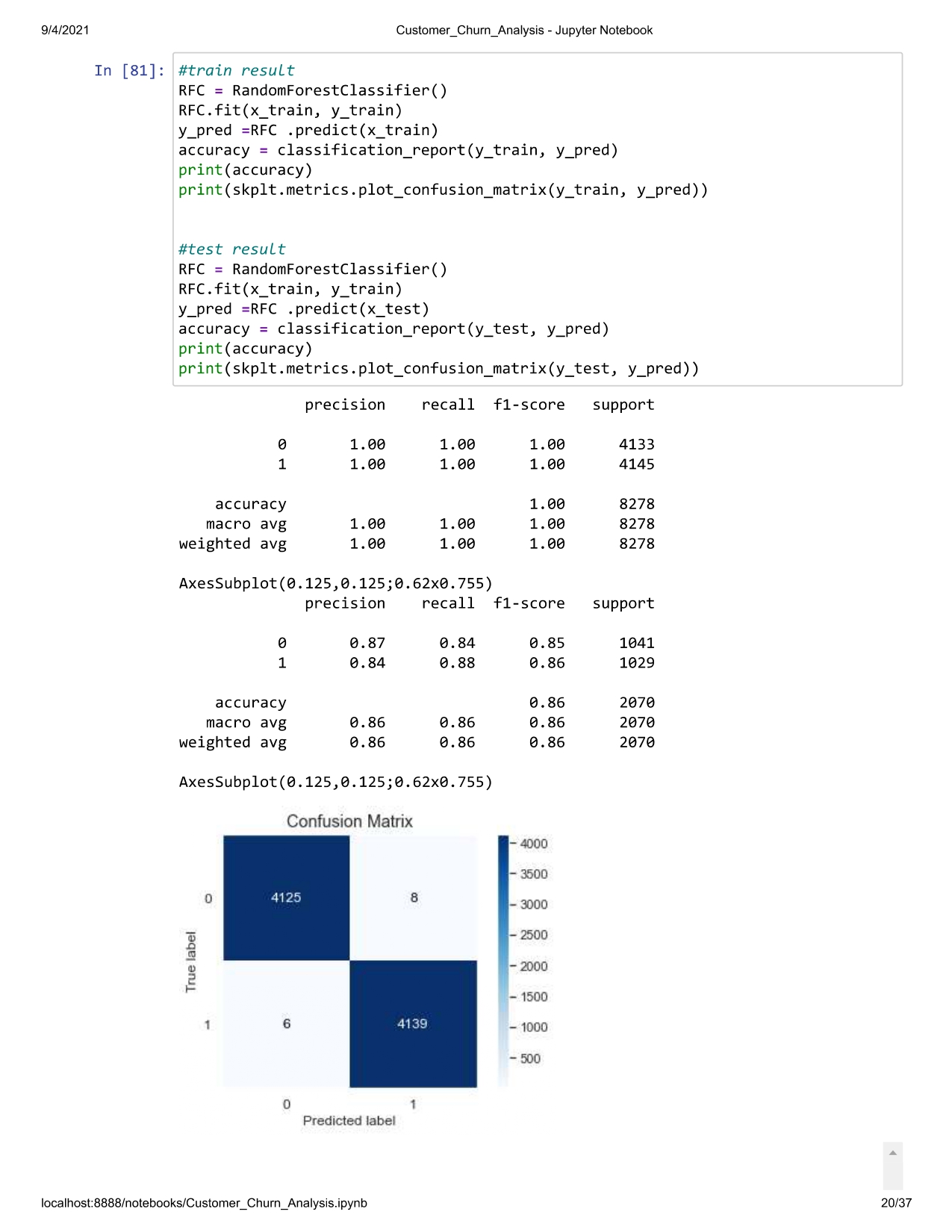


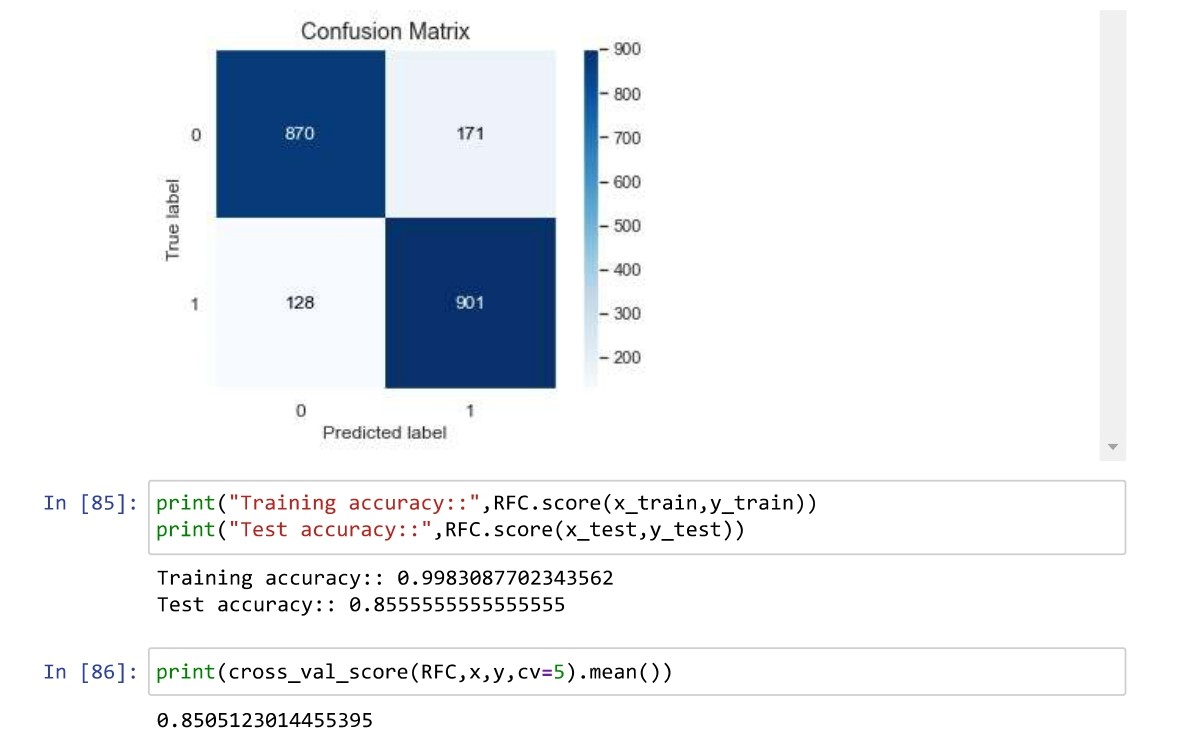


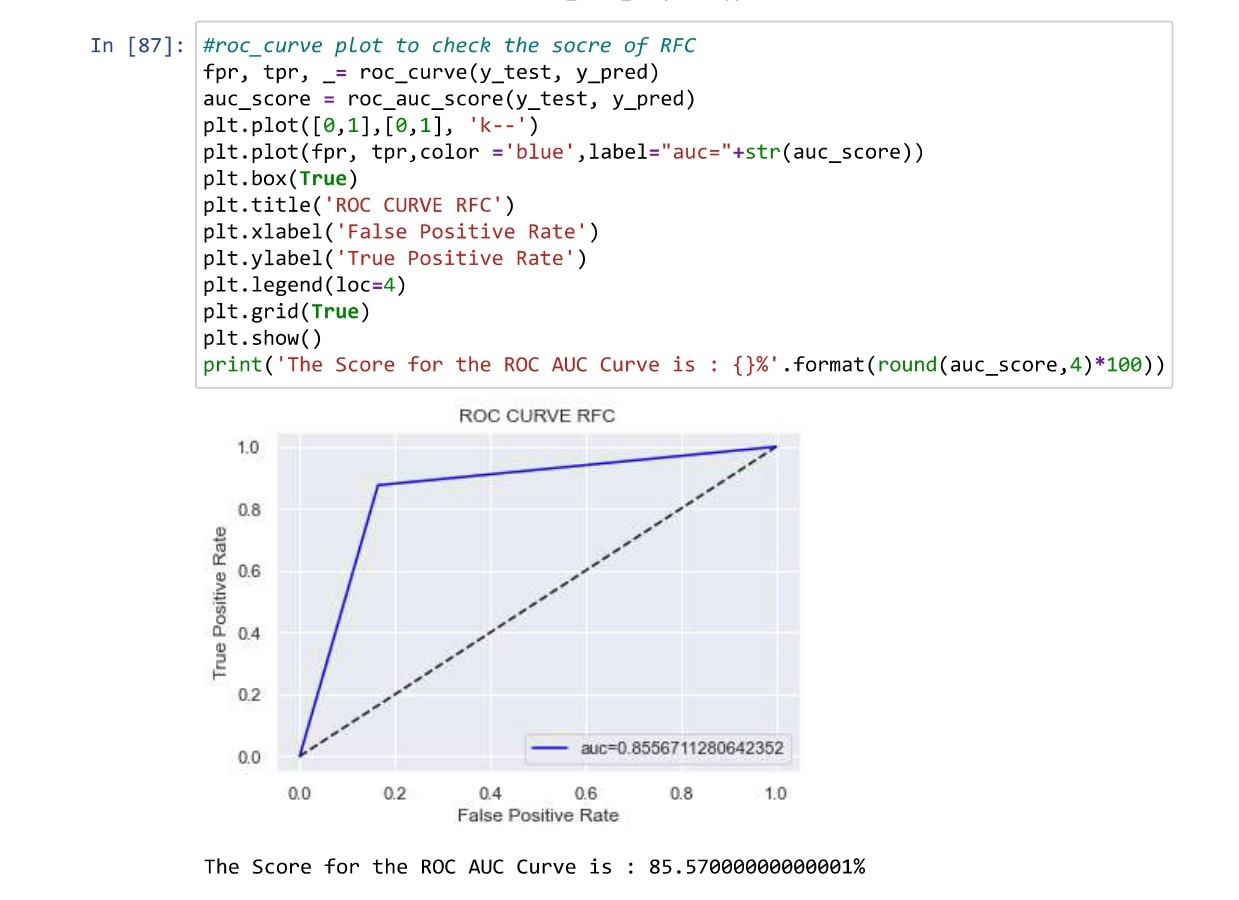


RandomForestClassifier

 Random forest is a supervised learning algorithm. ... The general idea of the bagging method is that a combination of learning models increases the overall result. Put simply: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

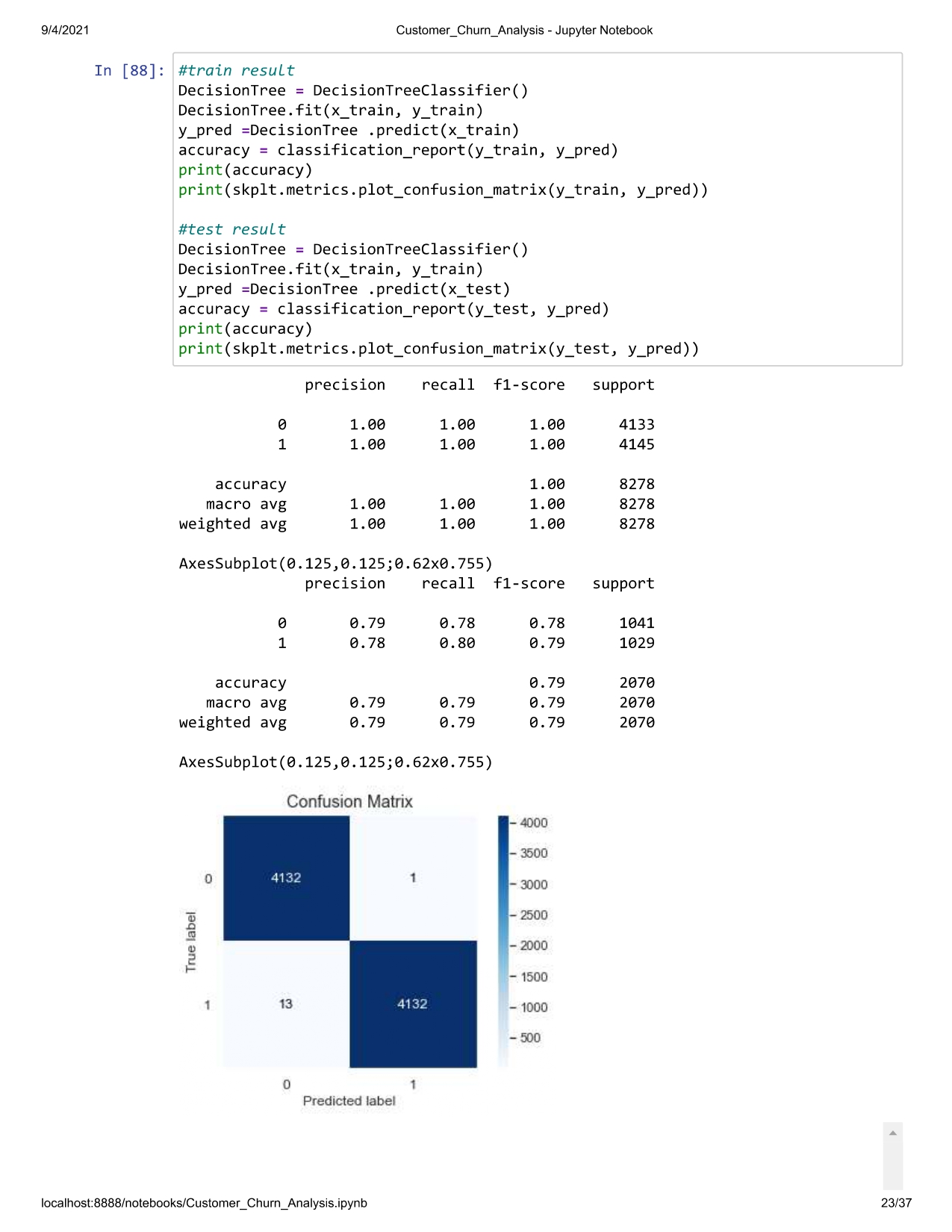


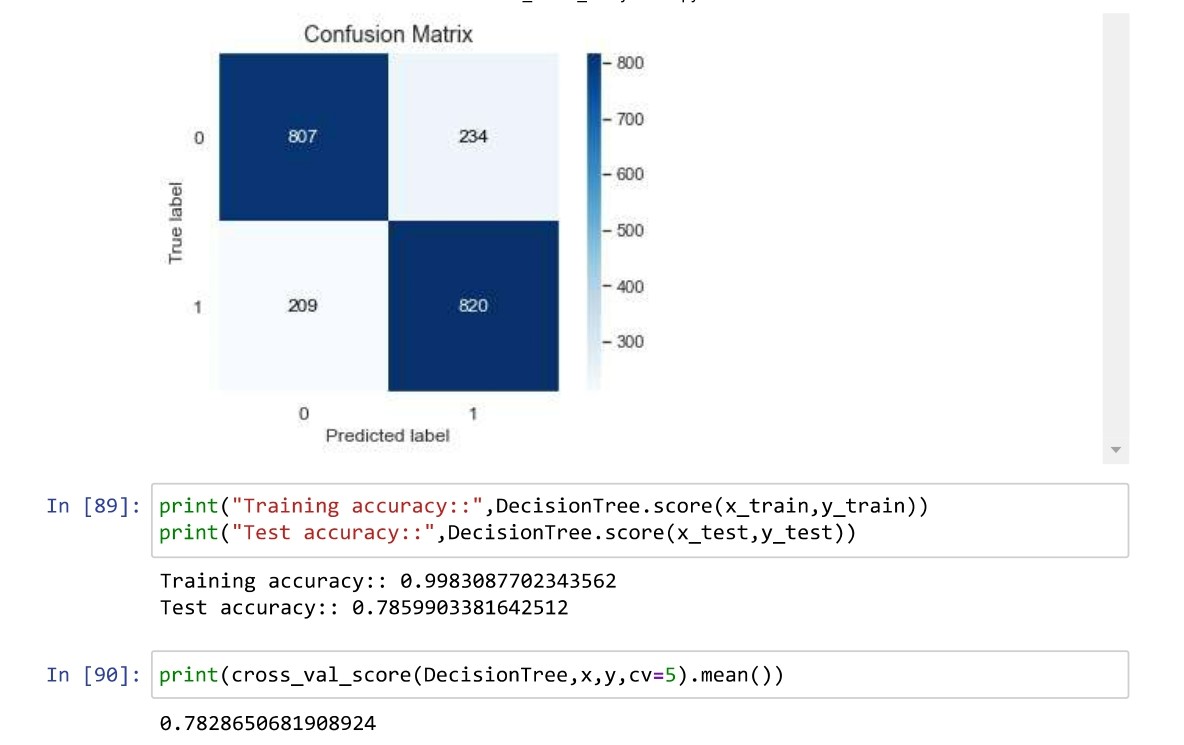


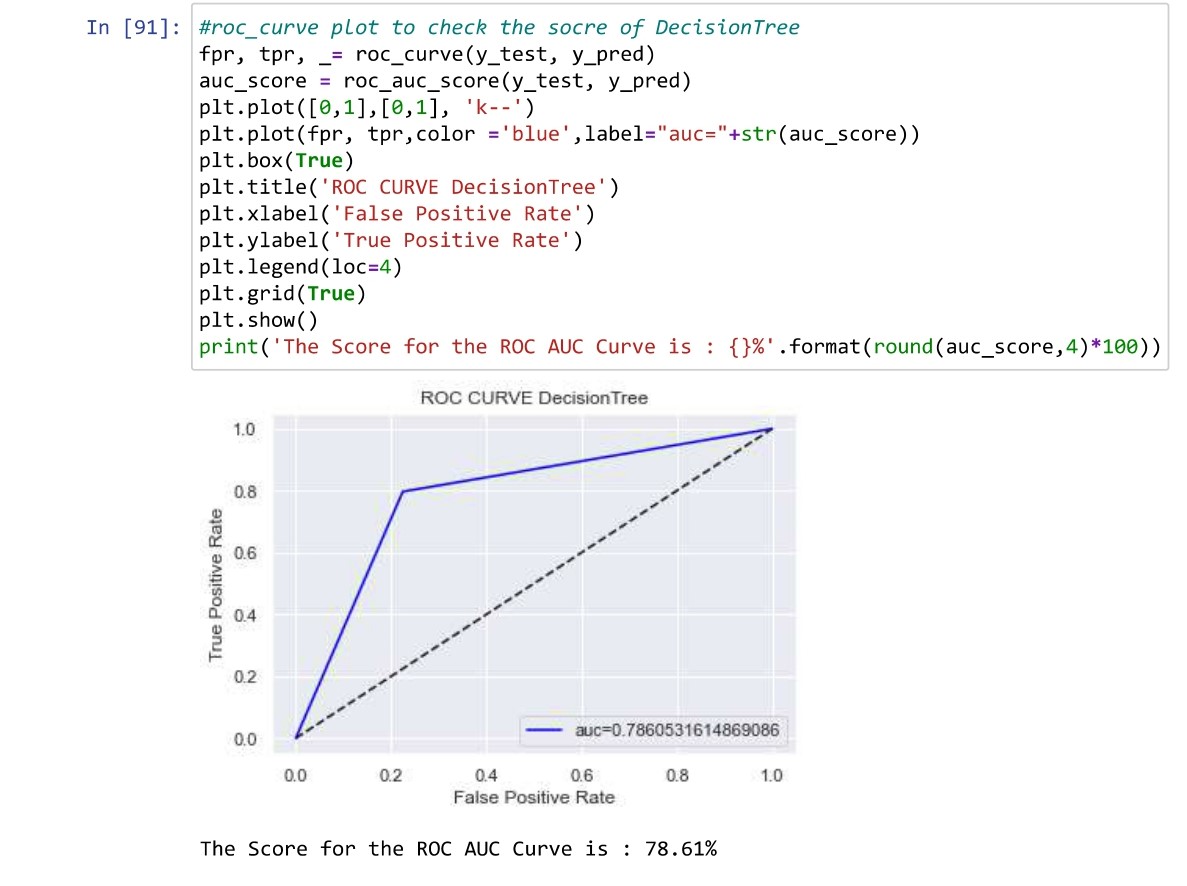


DecisionTreeClassifier

The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

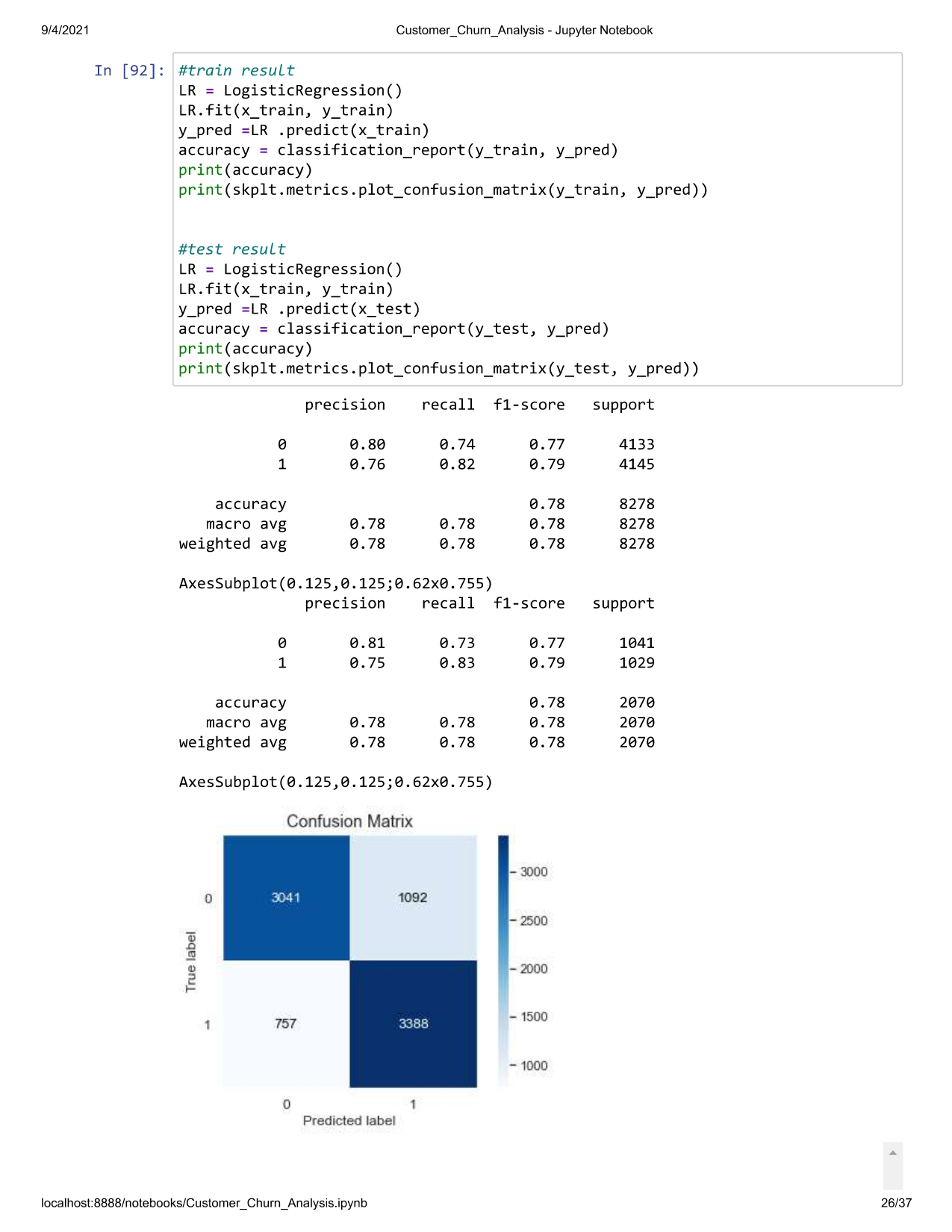


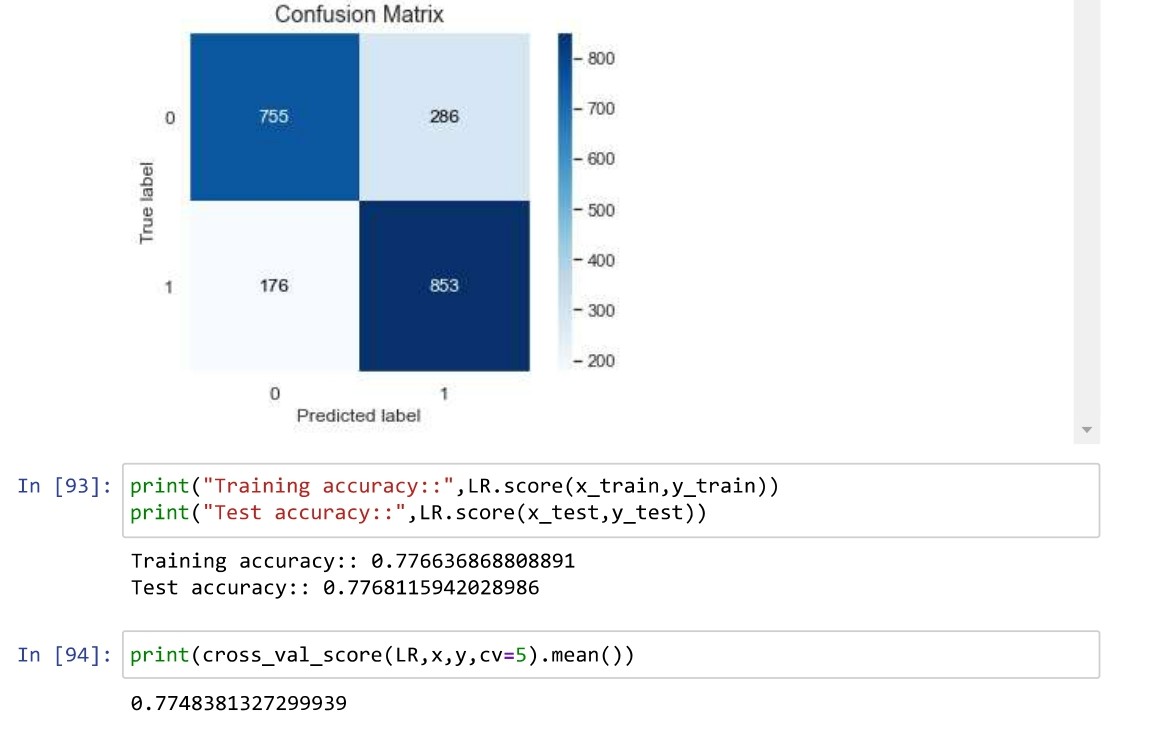


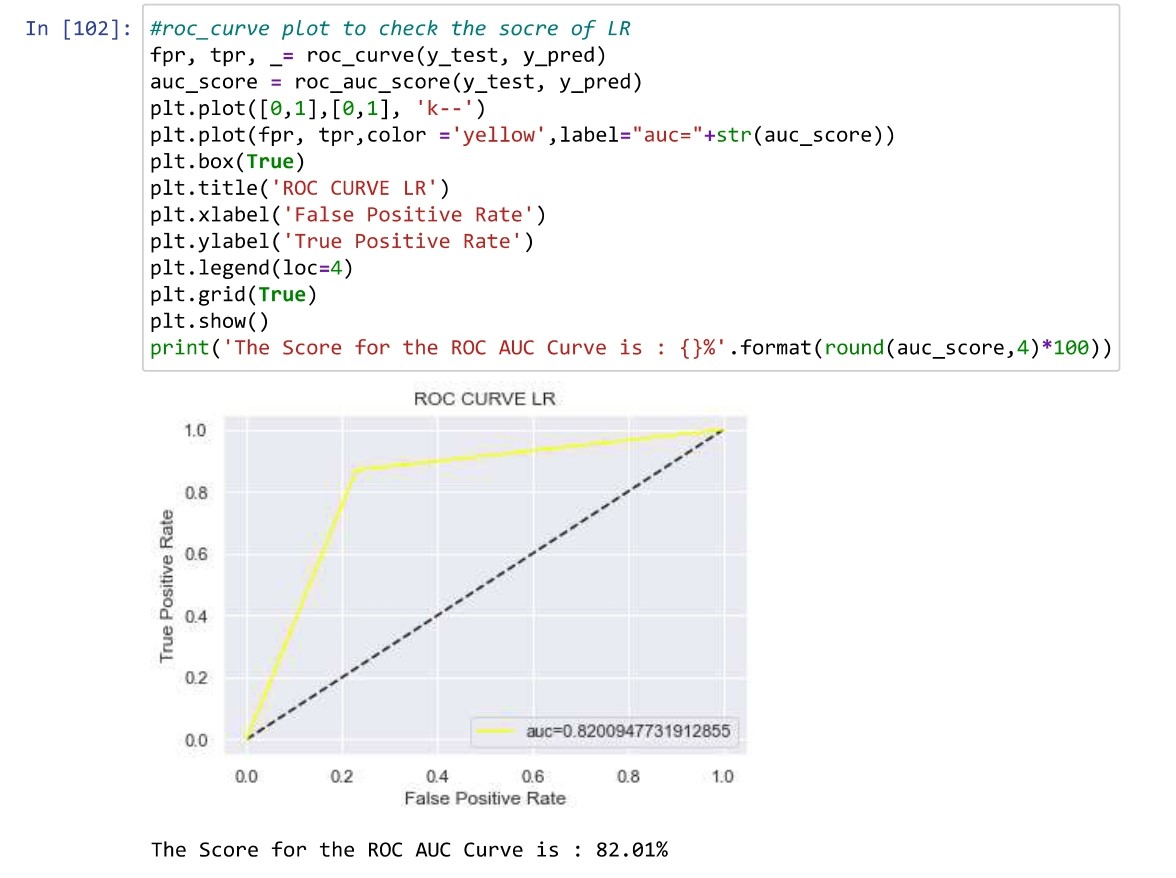


LogisticRegression

Logistic regression is a linear algorithm (with a non-linear transform on output). It does assume a linear relationship between the input variables with the output. Data transforms of your input variables that better expose this linear relationship can result in a more accurate model.



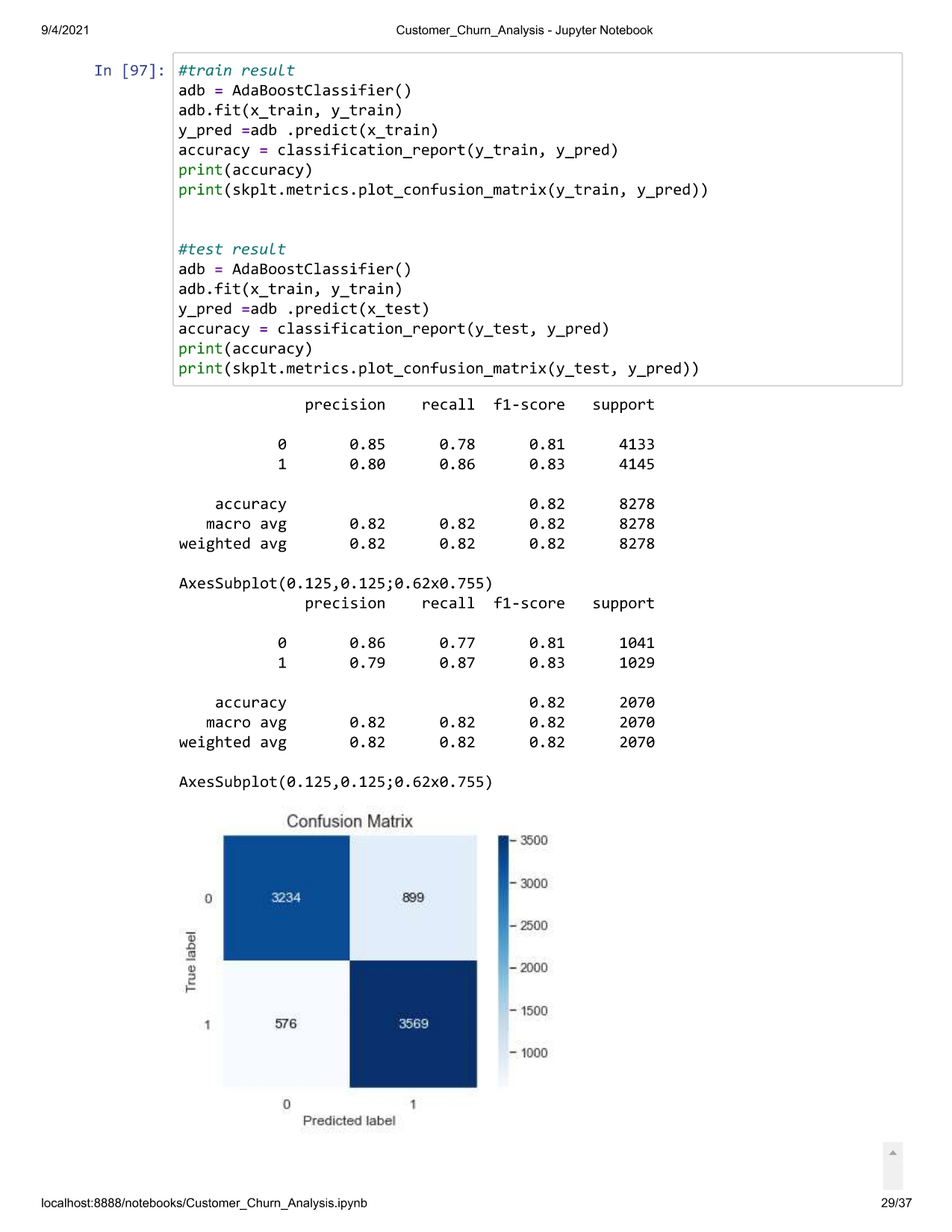


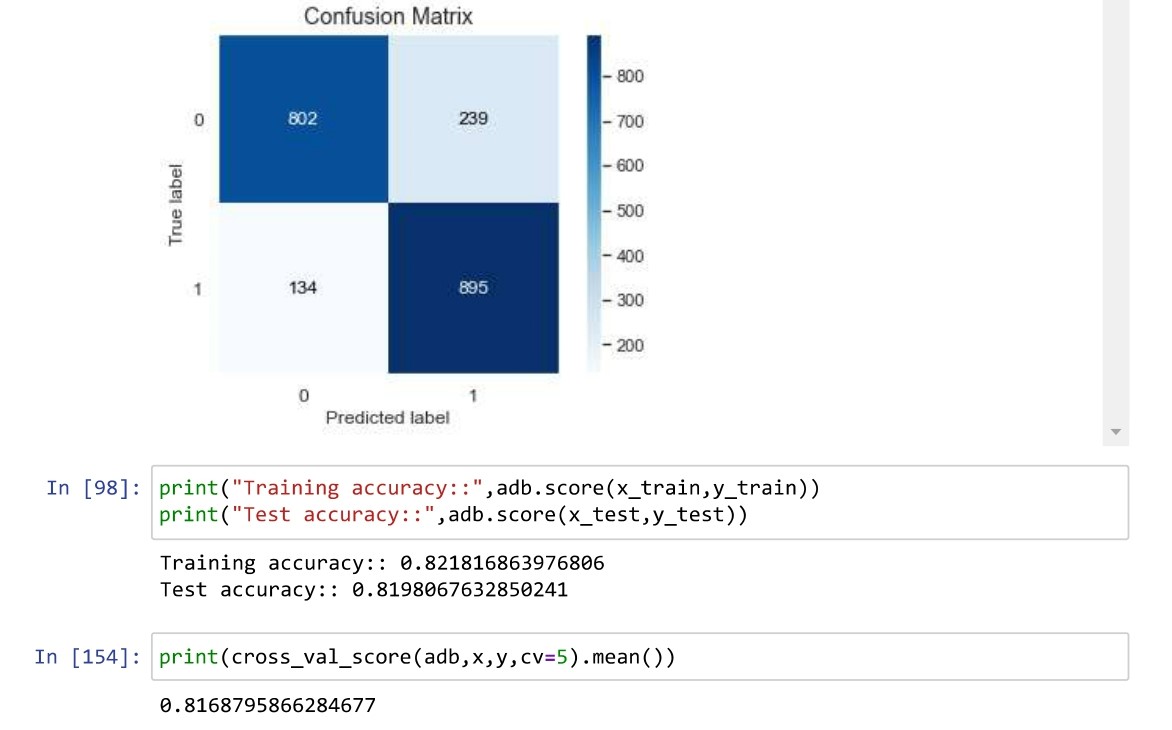


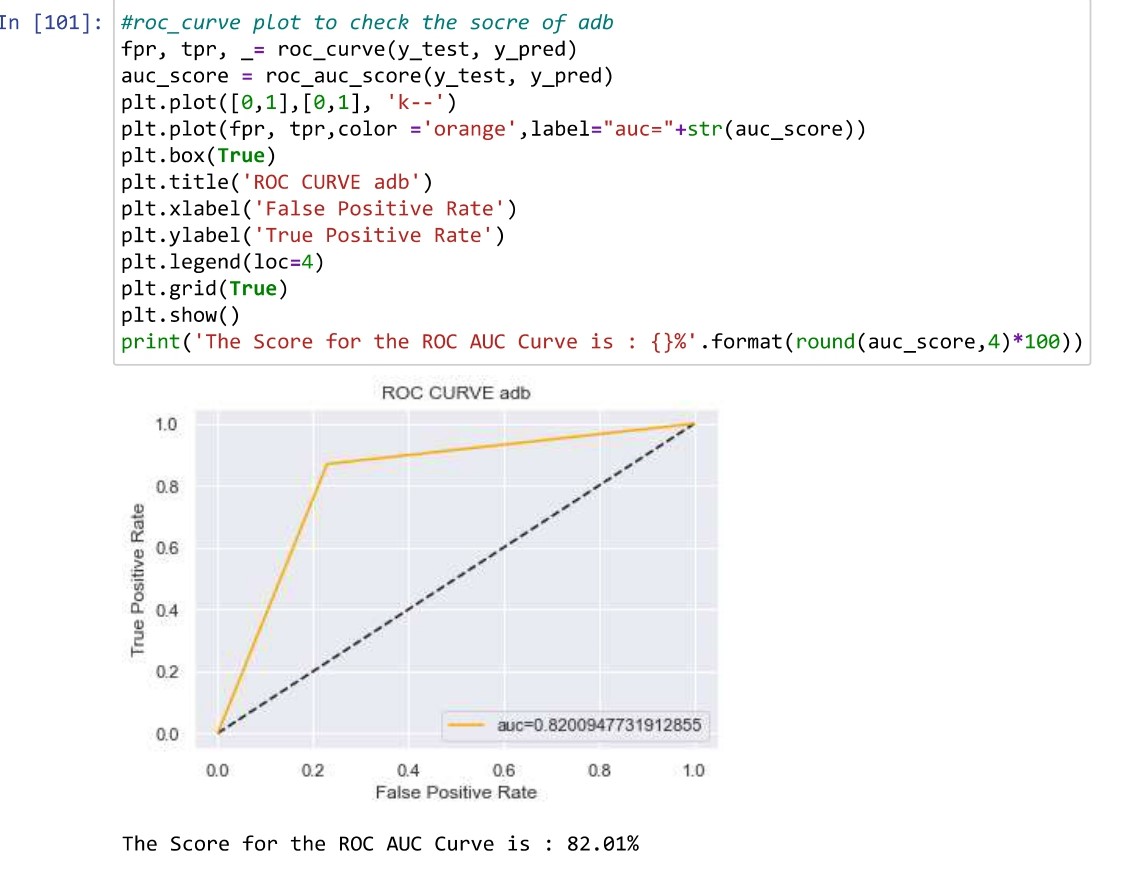
**ENSEMBLE TECHNIQUES**

**AdaBoostClassifier**

It combines multiple classifiers to increase the accuracy of classifiers. AdaBoost is an iterative ensemble method. ... The basic concept behind Adaboost is to set the weights of classifiers and training the data sample in each iteration such that it ensures the accurate predictions of unusual observations

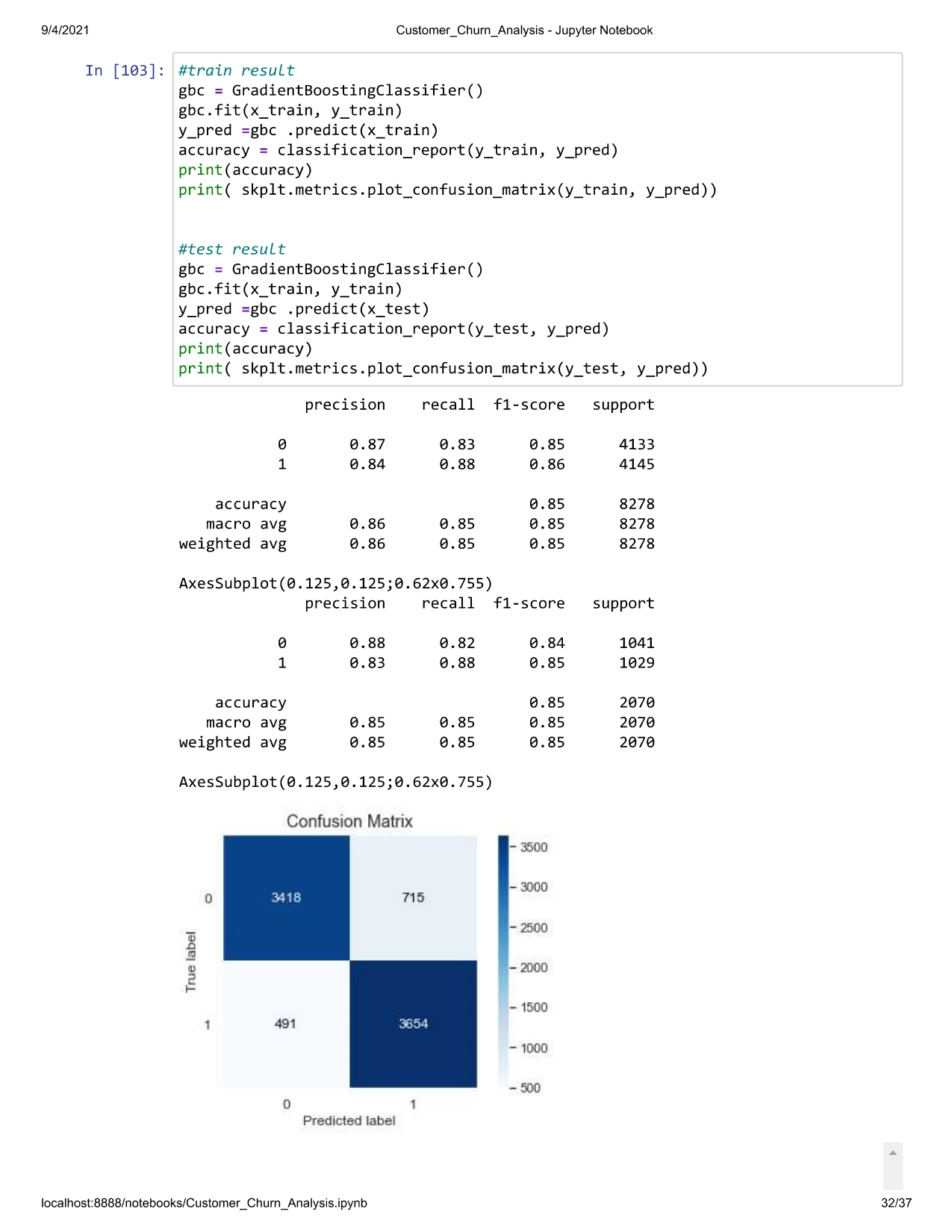


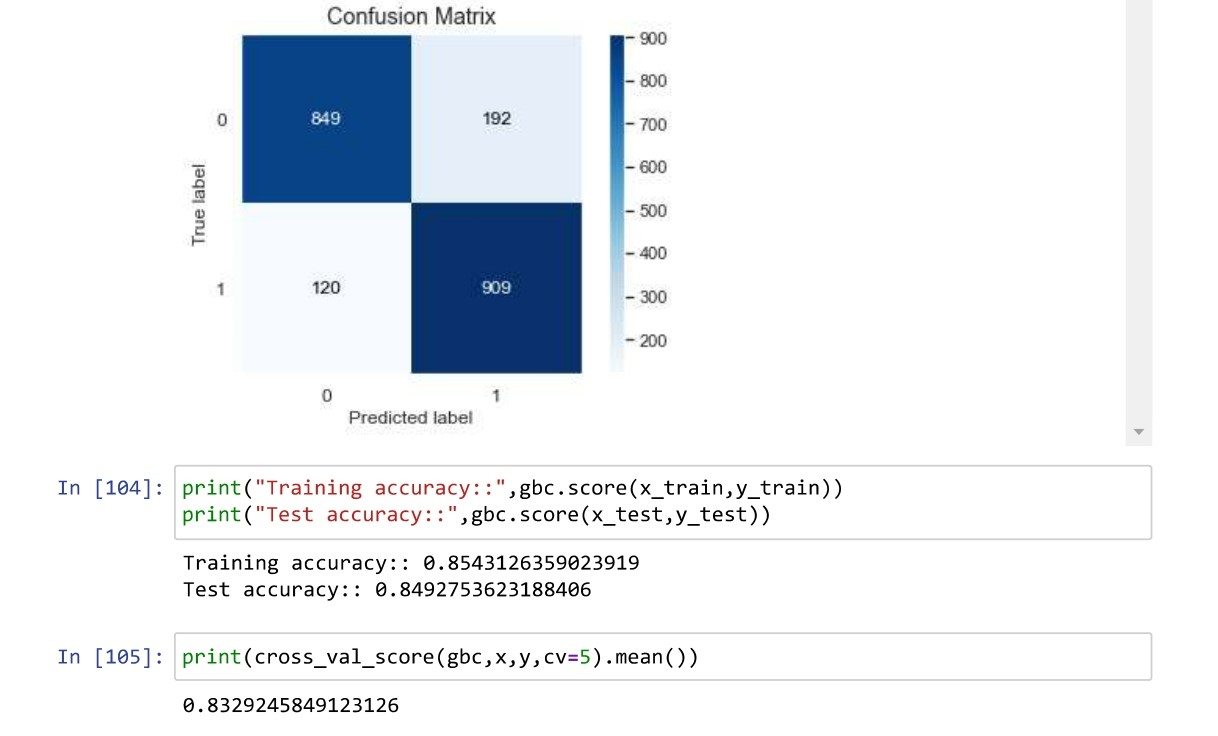


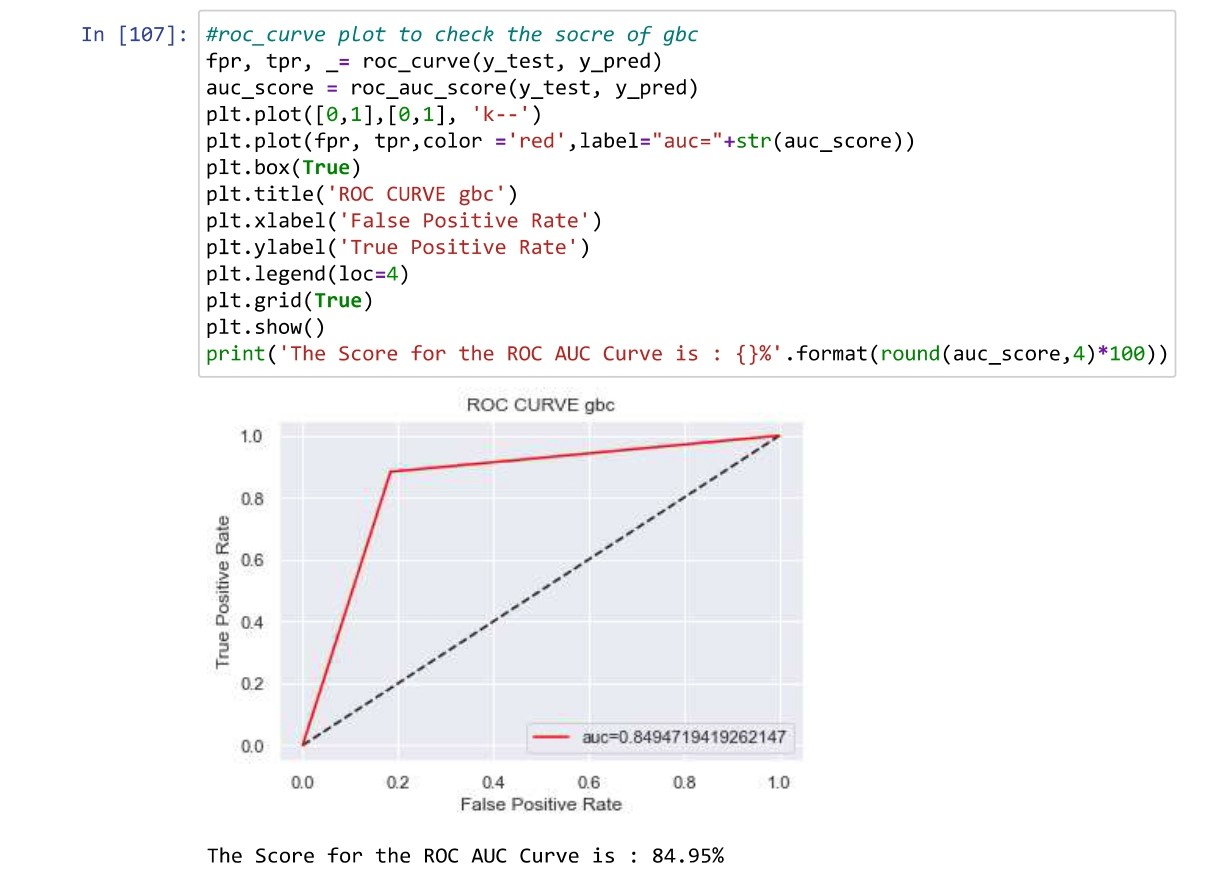


**GradientBoostingClassifier**

Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. ... Gradient boosting algorithm can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier).



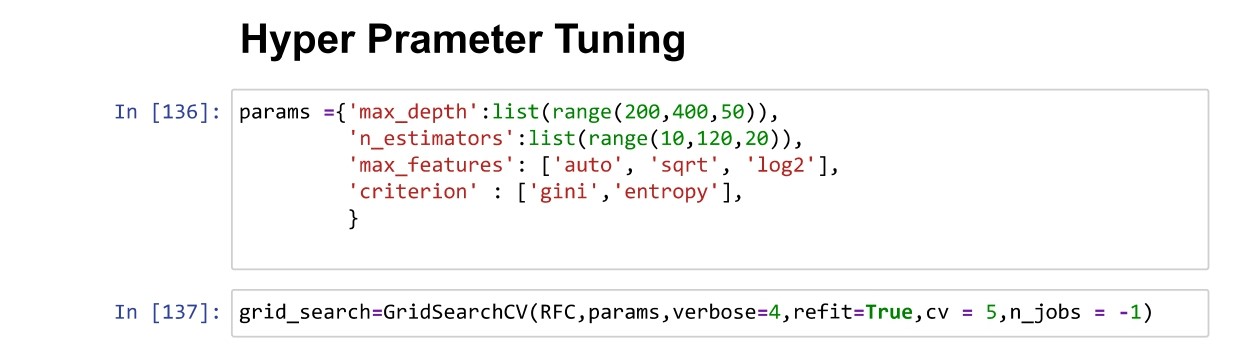


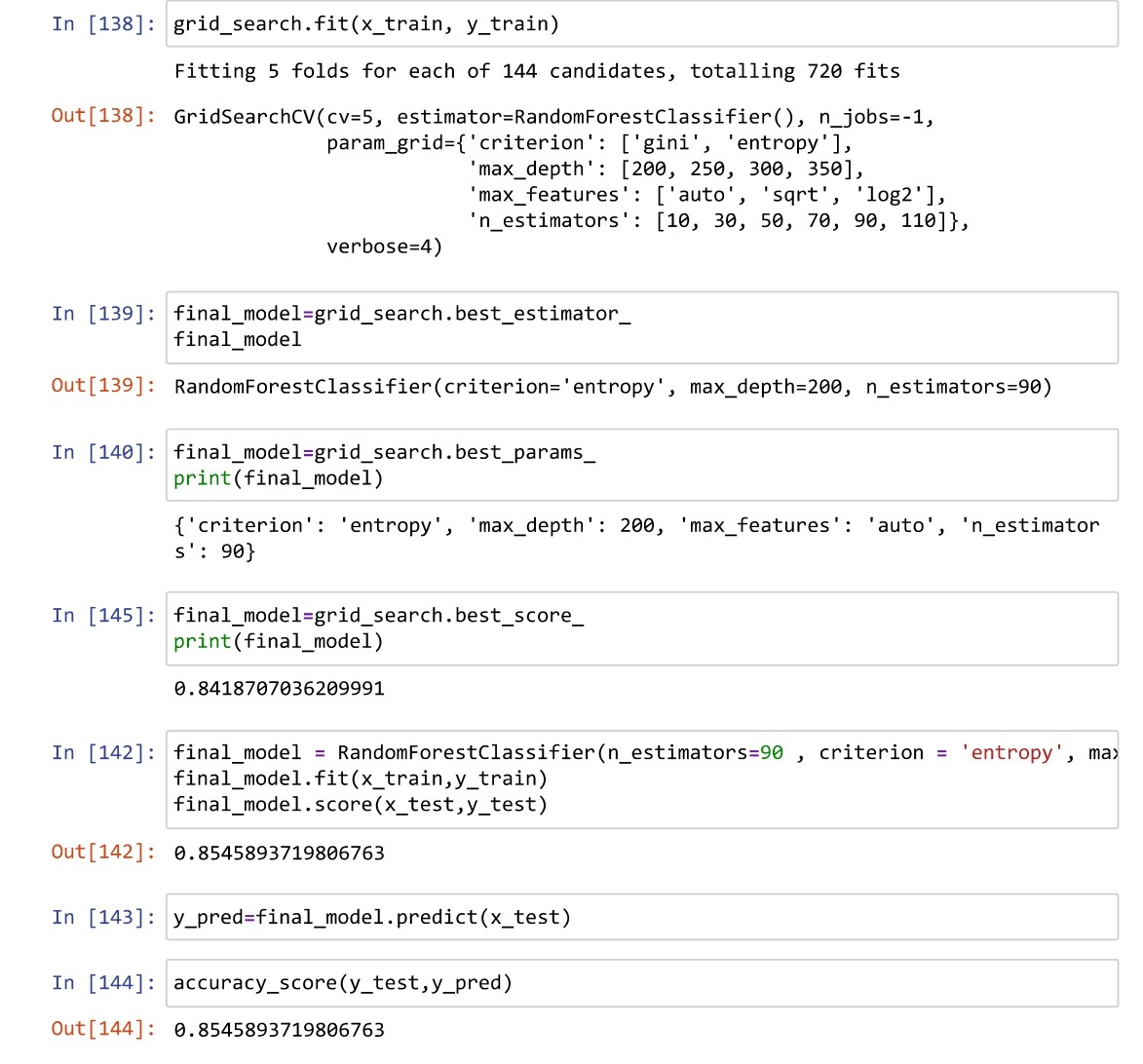


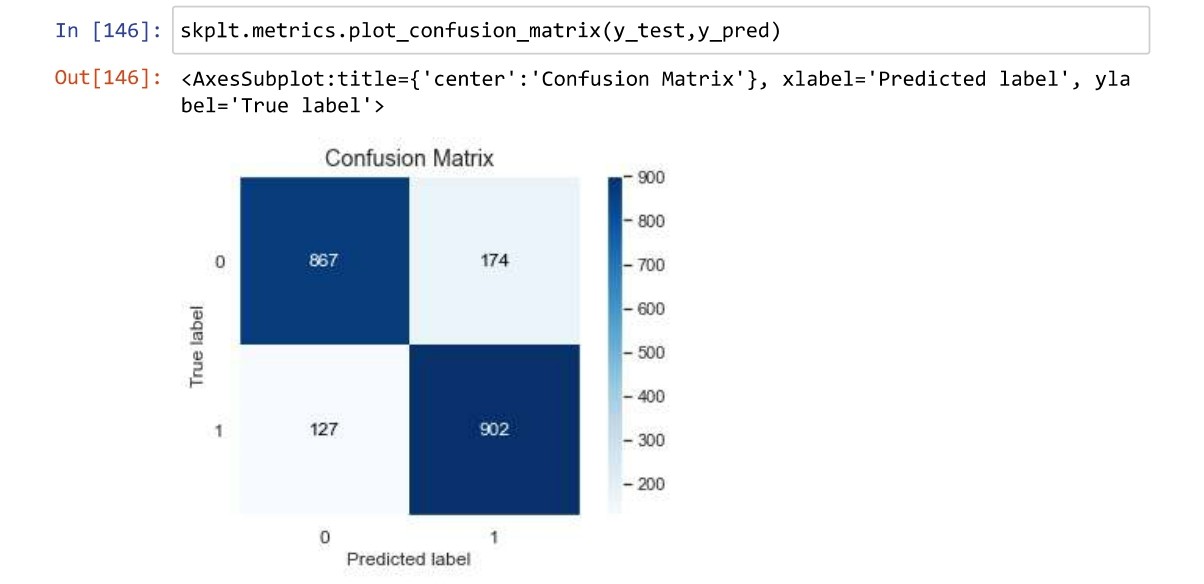
finally after difference of r2score accuracy and cross validation accuracy we got random forset classification is our best model and its 86% of good accuracy

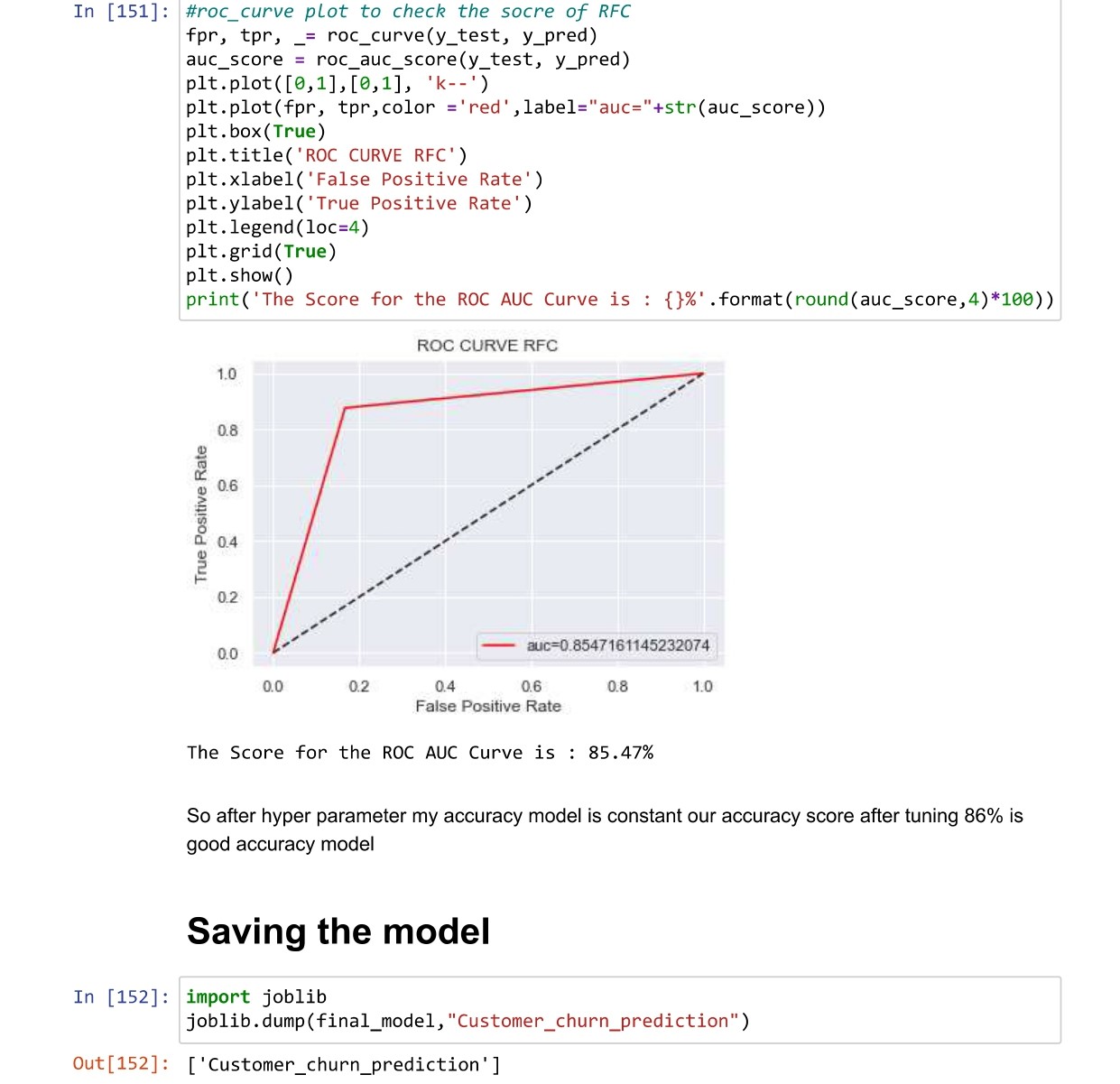
now we do increase our accuracy score with hyper parameter tuning

# Hyper Prameter Tuning









So after hyper parameter my accuracy model is constant our accuracy score after tuning 86% is good accuracy model

Final conclusion: people having very high tenure or very less tenure are leaving company

People don’t have the phone services are not enjoying otherservices.So,probably customer is leaving ,here company can work upon new schemes so that customer can get attract towars services but max customers can’t afford for two yers subscription,company should comeup with again new schems and offers

After model building , RandomForestClassifier algorithm looks best for the telecom customer dataframe, which will predict churn analysis