**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 priorities 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.

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

**Problem Definition: -**

As we all know that churn prediction is a common use case in machine learning domain and here, we will analysis the churn prediction in telecommunication sector. Churn simply means “leaving the company”. In any business or company, it is far less expensive to keep their existing customer to get new customer. We will use different methods to analysis and observe the churn prediction using machine learning model.

**Source of dataset: -**

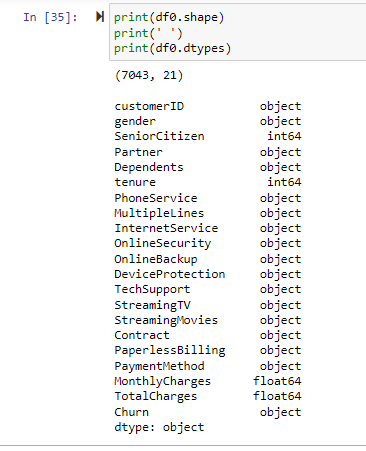
The dataset is given by “Datatrained”.

**Link: - https://github.com/dsrscientist/DSData/blob/master/Telecom\_customer\_churn.csv**

**Data Analysis: -**

First, we need to import the important libraries and required dataset from the above given link.

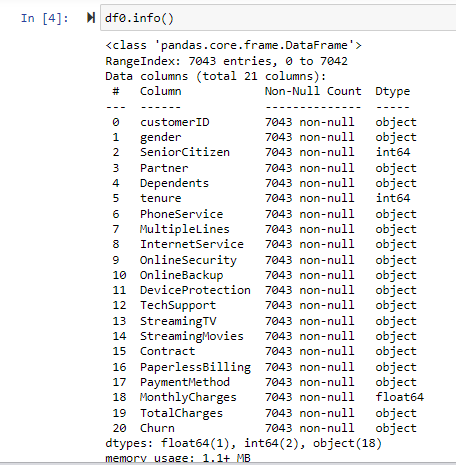




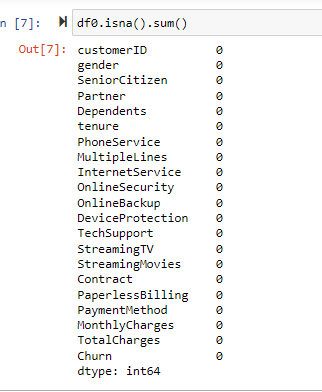
The given dataset contains 21 columns and 7043 rows. In which 20 features are independent variables and 1 is our target (dependent variable). Target variable indicates if a customer has left the company (i.e., churn=yes) or not (i.e., churn==no).

There are different types of variables in the given dataset. customerID, gender, Partner, Dependents, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, TotalCharges and Churn contains categorical datatypes and remaining i.e., SeniorCitizen, tenure, MonthlyCharges are of continuous datatype.

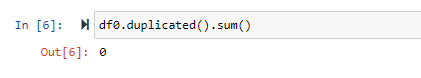
At first glance, only customerID seems irrelevant to customer churn. Other variables may or may not have an effect on customer churn. We will figure out.



**Let’s start with checking the missing values in our given dataset and there is no missing value.**

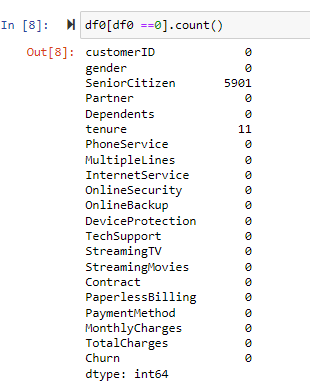


**Now, let’s check for the duplicate values**

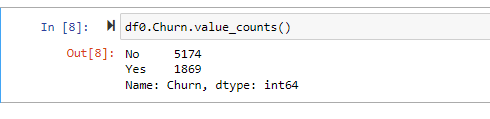


There is no duplicate value.

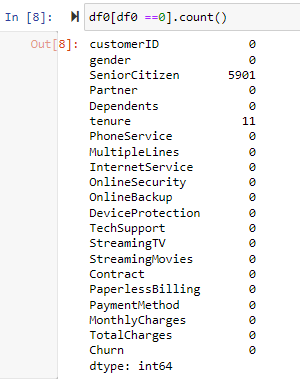
**Checking the variables that contain zero**



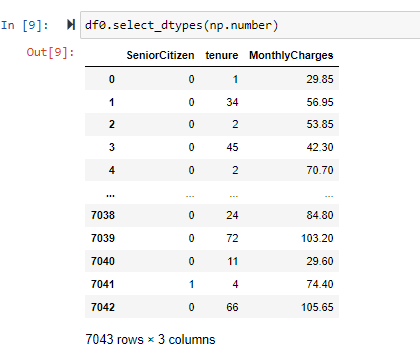
Target variable i.e., “Churn” has two value “yes” and “no”. in the given dataset the count of yes and no are as below



Yes has 1869 counts and No has 5174

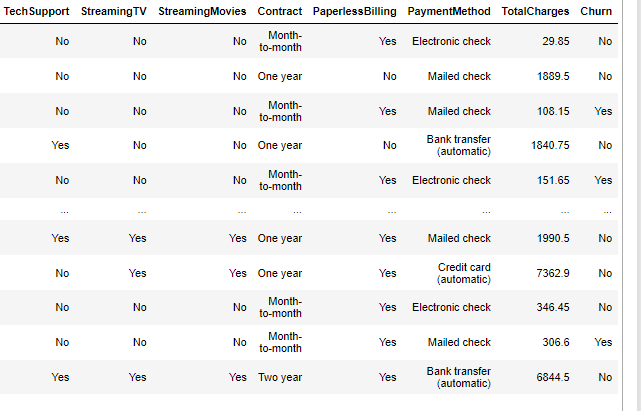
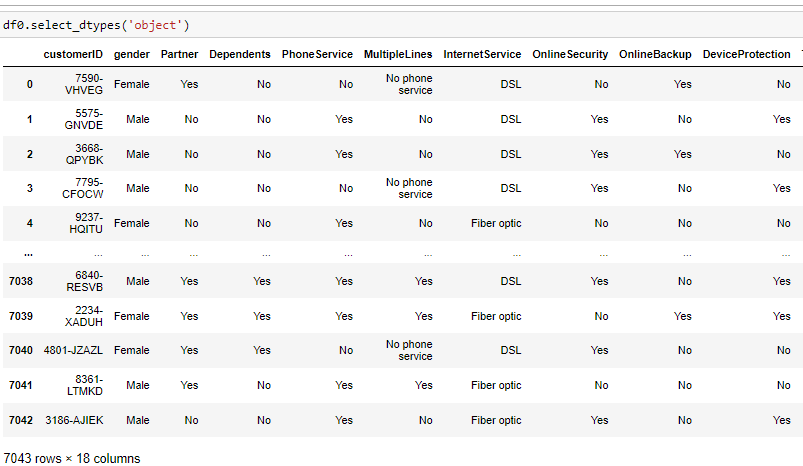
**Check the variables that contain “zero” as its value.**

As we can clearly see only Seniorcitizen and tenure have “0” as its value. Seniorcitizen column contains two values ‘0’ and ‘1’. ‘0’ indicates customer who are not senior citizen or not aged and ‘1’ indicates customer who are senior citizen or aged. Tenure columns also contains numerical values and shows the time period of customers associated with this company.

**Check the columns that contains only categorical data**.

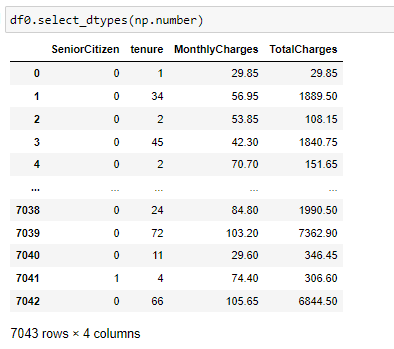
There are only three columns(variables) of numerical datatype.

**Now, let’s check the variables that contains categorical data i.e., datatpes= ‘object’.**

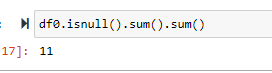
****

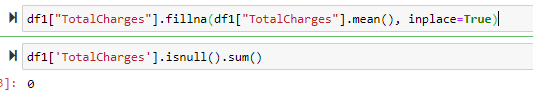
TotalCharges variable is containing continuous datatype but here it is shown as a categorical datatype. we will change it’s datatype.

It has been changed now. Let’s check it once again.

Now ,it’s showing as numerical data.

**Checking null values once again**

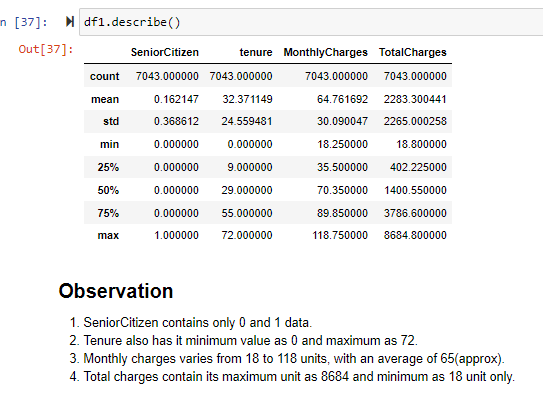
****It’s showing null values, we have fix it.

**Filling the null values**

we have fill the null values with mean(). Now, there is no null values present in our dataset.

**Dropping the irrelevent column i.e., CustomerID**

****

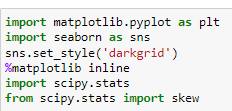
**After this, we done statistical analysis and wrote some observation.**

.describe() method is used to get descreptive analysis of numerical data with indexes like count, mean, standard deviation, min and max.

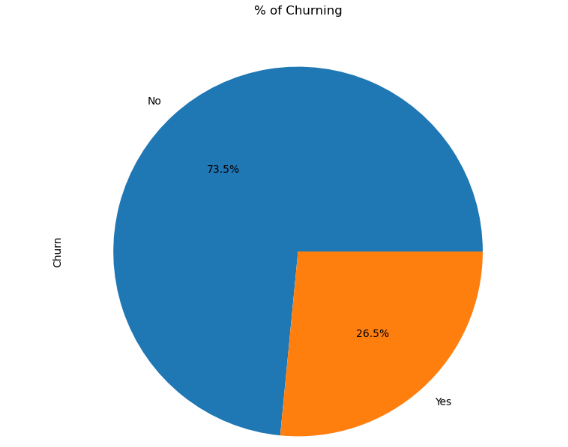
The observation from this table is written below.

**Now data visualization**

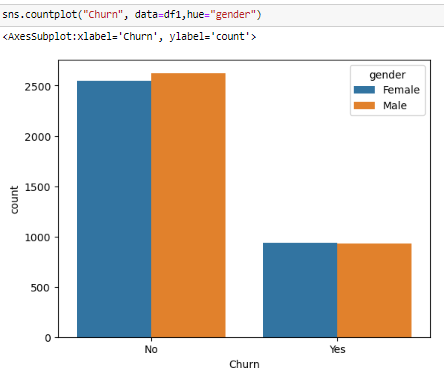
First, import all the necessary libraries to do datavisualization.

Other libraries will be imported when it is required.

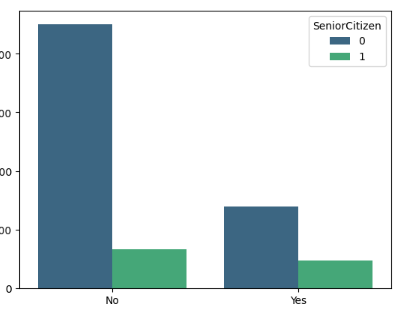
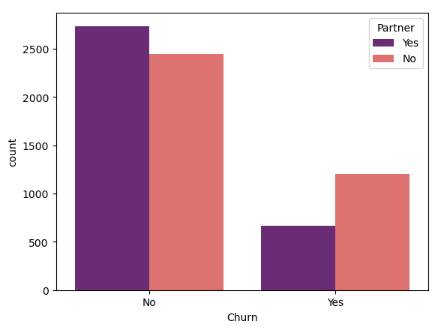
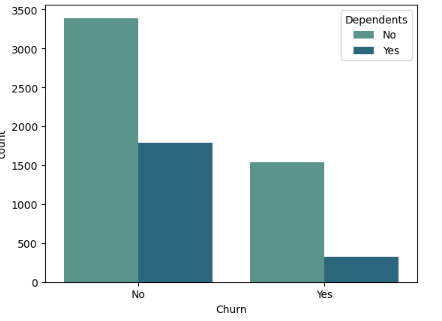
**Our target variable “Churn” has two values. So , we have created a pie plot of it.**

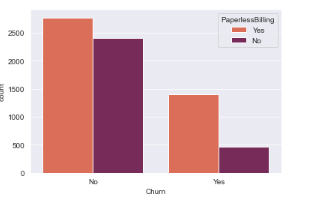
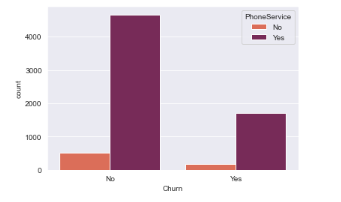
As we already know that in this given data set number of customer who has churned the company is relatively very low and this pie plot is showing it perfectly.

**Let’s see the churning percentage in different gender.**

**** In both female and male have nearly equal percentage of churning.

**Now, let’s check churn percentage different binary features**

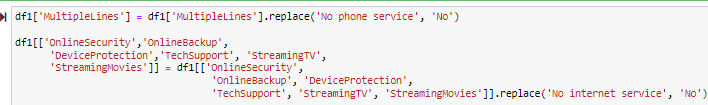
****



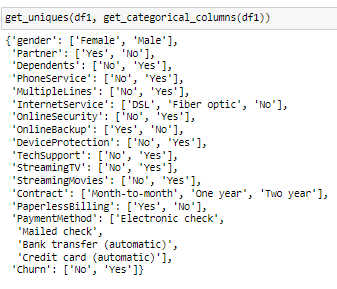
In above all the features churning number of customer is different in each feature.

**Done some changes in features value.**

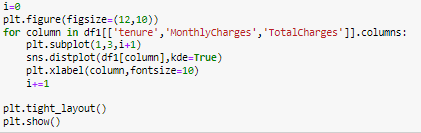
* MultipleLines column have 3 value :'No phone service','No' and 'Yes', in which 'no phone service' and 'no' indicates the same thing, so we try to replace 'no phone service' to 'no'.
* OnlineSecurity,OnlineBackup,DeviceProtection,TechSupport,StreamingTV,StreamingMovies, in these columns 'No internet service' and 'no' refers to the same point so we will replace one of these to other.

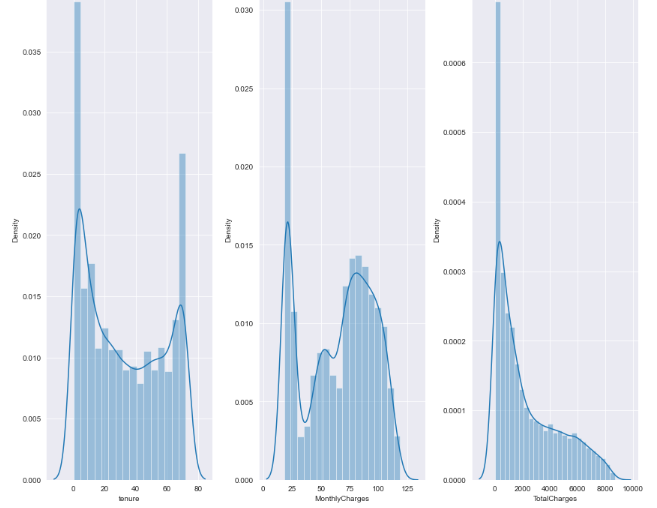


Using .replace() method.Now, let’s check it out.

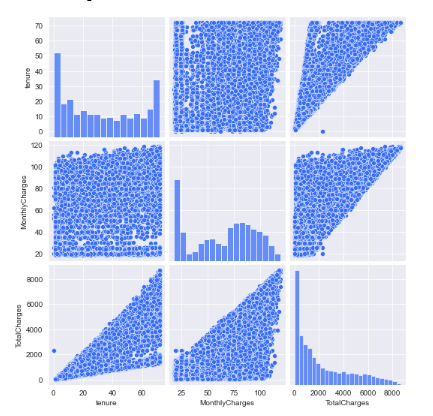
 All the above features values has been changed to correct manner.

**Now, let’s check the distribution of numerical features using distplot.**

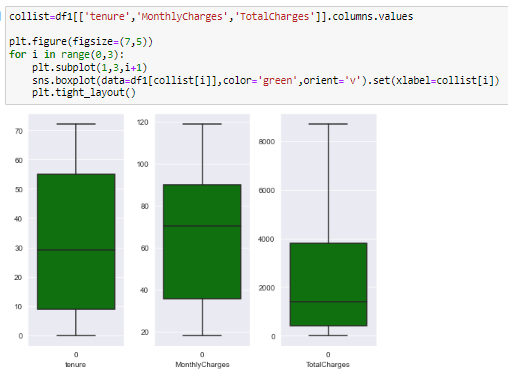
****

****The following features is not distributed well. We will check it later and will try to correct it .

**Pairplot:**

****

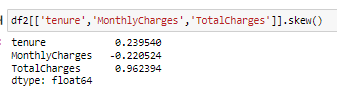
**Checking the outlier:**

****

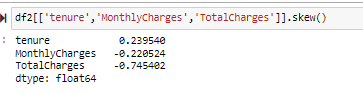
There is no outlier detected here. Now, let’s move and check the skewness of the following features.

1. First, we have to import required libraries to check the skewness of features.
2. Now, we will check it using .skew
3. As we know skewness values should lies between -0.5 to +0.5, if any feature’s skewness is not in between this range will transform the particular feature.
4. Let’s do it now

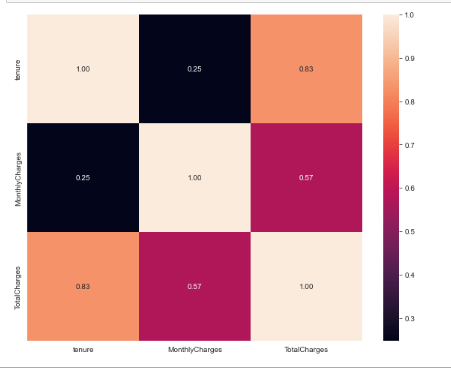
Libraries imported.

Only totalcharges’s skew value is not between above range. Let’s try to transform it with np.log1p method.



It skew value is decreased little bit. We will leave it as this and move ahead.

**Heatmap**:



There is not much collinearity. So, we will move to preprocessing now.

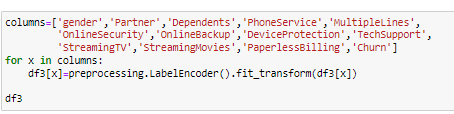
In data analysis, we have checked the all null values , duplicate, variables containg zero as it’s value. Also changed the datatype of totalcharges variable and filled it null values.Dropped the uneccessary column. For visualization we have different visualization graph like- distplot,countplot,scatterplot,pairplot, pieplot and etc. we have also checked and corrected the distribution of all numerical features. Also used the outlier detection technique. Skew values is reduced by using log transformation. For multicollinearity, we have ploted heatmap. Now, moving to next step i.e. data encoding…

**Data encoding:**

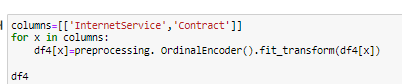
Import the libraries.



Now, using label encoder, we will encode the following features:



Using ordinal encoder, we will encode internetservices and contract features.



The remaining features will be encoded using one hot encoder.

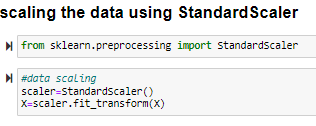


**Building the machine learning model:**

For machine learning model, first we have separate the target and feature variables.



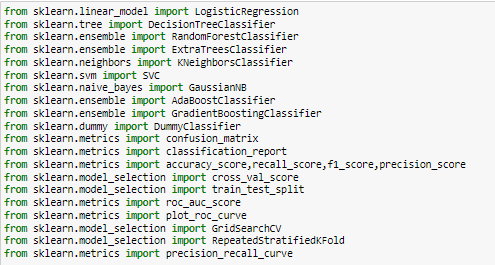
Now, we will scale the data using “**standardscaler**”



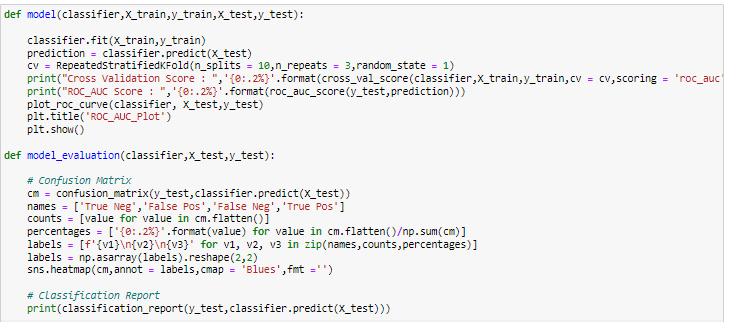
Now, we will split the train and test data using train\_test\_split method of from sklearn.model\_selection.

in the practise project, we were told to use only 25% for testing.

Important libraries for building the different machine learning model:

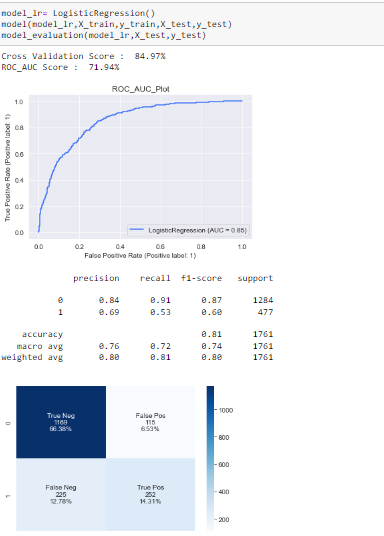


Now, we will use the required algorithm to get cross validation score, roc auc score , confusion metrix and classification report.



**cross validation score, roc auc score, roc auc curve, classification report and confusion metrix of different classification machine learning model.**

1. **Logistic regression.**

****

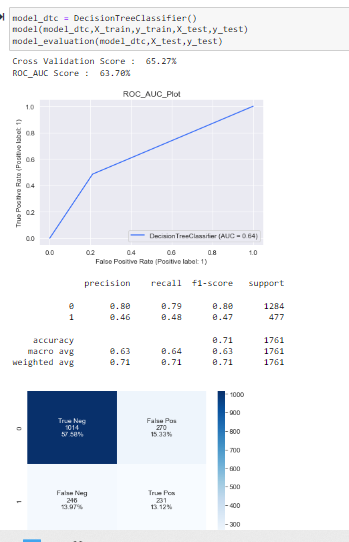
Cross validation score is 84.97%

Roc auc score is 71.94%

Accuracy is 81%.

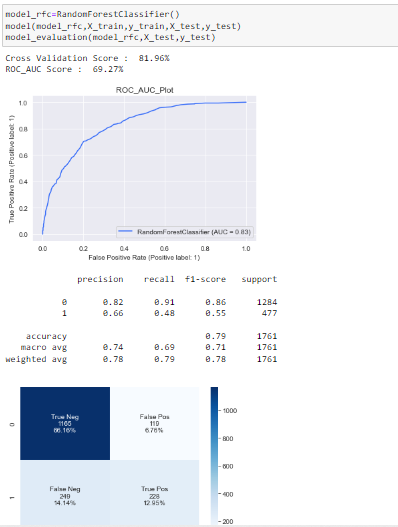
This model has a good accuracy of 81% but we will try to built some other model and check their performance.

1. **Decision tree classifier**

****

It’s accuracy is even less than logistic regression model. We will try someother model.

1. **Random forest classifier.**

****

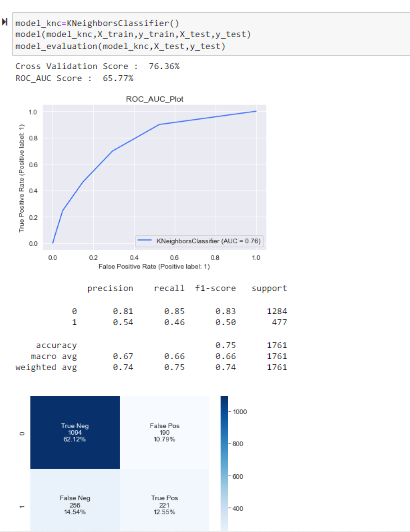
It’s accuracy is better than decision tree classifier but lower than logistic regression. Will try someother model and check their performance.

1. **Extra tree classifier.**



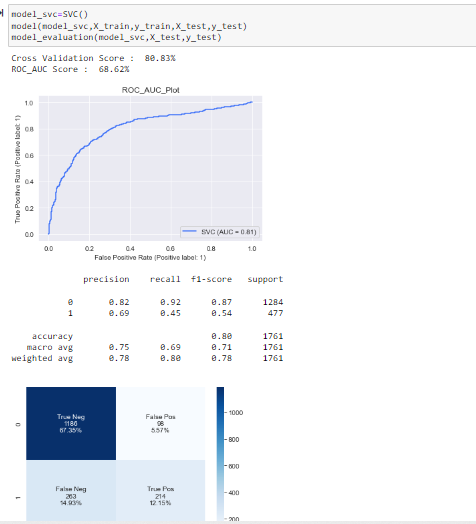
It accuracy is in between logistic regression and random forest classifier, we will see some more model and will compare it with others.

1. **KNeighbors classifier.**



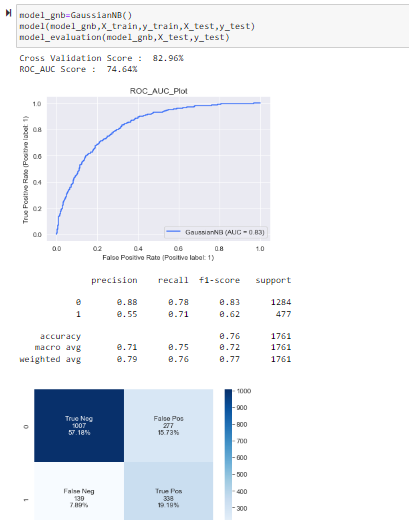
It’s accuracy is also not that good. So, we will try other model.

1. **SVC**.



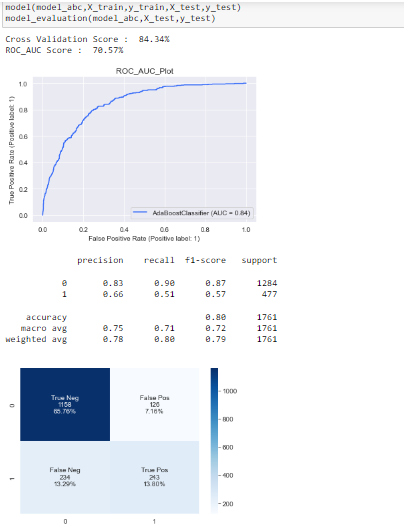
It’s accuracy is better than the above classifier but lesser than logistic regression model.

1. **GausianNB**.



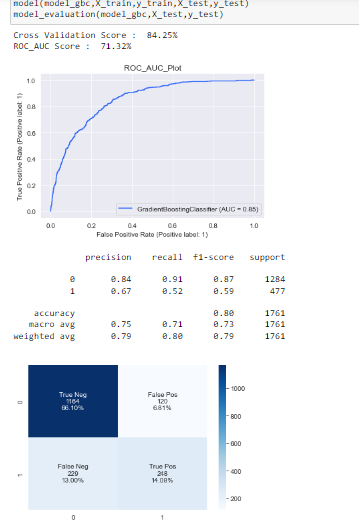
It’s accuracy is also not that good.

1. **AdaBoost classifier.**



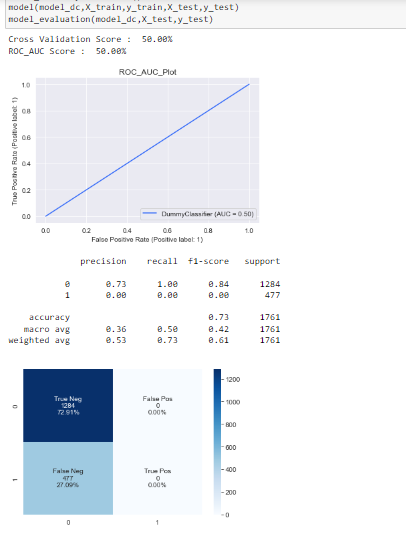
It has better accuracy than someothers model and also lesser than some.

1. **Gradient Boosting classifier.**



It has pretty much similar accuracy than above model.

1. **Dummy classifier.**



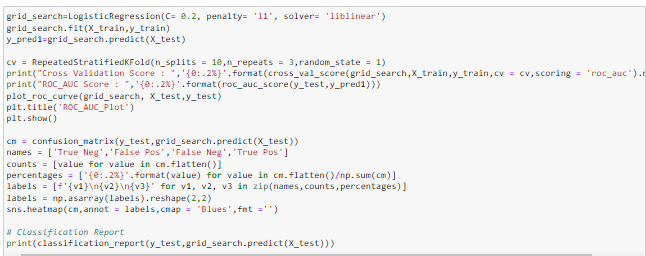
It has worst accuracy than others.

Out of all above different models, Logistic regression has highest accuracy and better cross validation score with roc auc score. So, we will tune this using gridsearchcv library with required parameters and check, if the accuracy of the model is increased or decreased or have same as it was before the tuning.

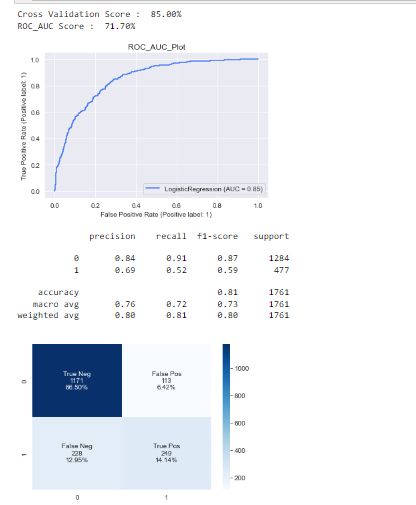
**Library: from sklearn.model\_selection import GridSearchCV**

****

In above algorithm we have used the hyperparameter of logistic regression and than fit it with X\_train and y\_train data. After this, we will get our best parameter for this model, using this “best\_params\_” we will again fit the X\_train and y\_train, and will predict the X\_test. After all this , we will again built the logistic regression model with these parameter and these data.



This is the required algorithm.Now, will check the performance of this model and will compare it with the model before tuning.

 If we see both model (before tuning and after tuning) closely, we can clearly observe that the accuracy and other performance parameter have increased by very small margin. It’s cross validation score and roc auc score has also increased.

Now, we will save the model using pickle library.



**Conclusion:-**

From the above modelling technique, we can reduce the customer churn percentage by analysis all features individualy and can get the exact reasons of why the existing customer is churning or likely to churn. This model will also help the company to retain the existing customer by providing slight changes in the different services(mentioned as features in this model). Companies very well know the the expenses of keeping an existing customer is far less than to get new customer and they can also make the required changes to keep their existing customer as if they know the exact reasons of churning. These machine learning model is vastly used to get the exact reasons of churning and also help in lowering the churning rate of any company.