**Predicting Customer Churn Analysis**

**1.Problem Definition: -**

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.

What exactly is Churn Prediction?

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.

## Why does Churn occur?

## Churn is occurring due to any other new company giving the good offering to move the one company to another company.

## Hence, there is no correct answer as to why exactly the customer wants to churn

## A data scientist’s job is to find such patterns in the data given and see what facts are churned out during data analysis.

**2.** **Data Analysis: -**

I have used the customer data from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models.

**Note: You can find the dataset in the link below.**

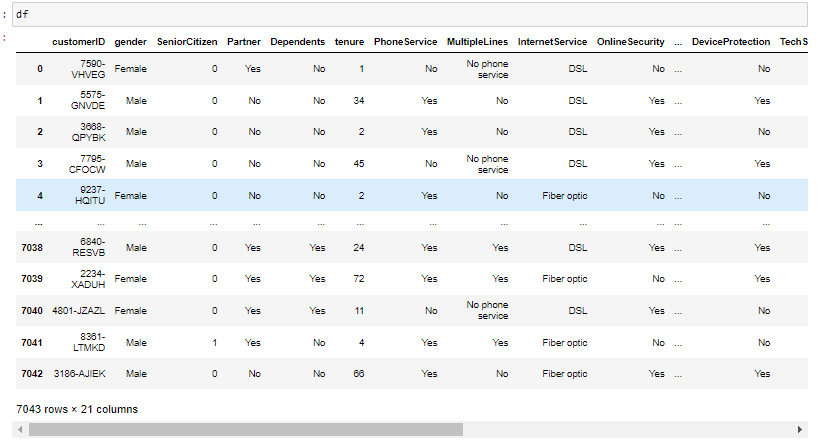
* <https://github.com/dsrscientist/DSData/blob/master/Telecom_customer_churn.csv>

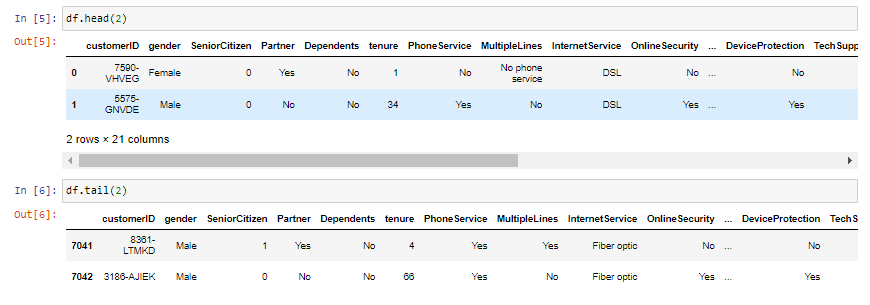
# Importing the Libraries



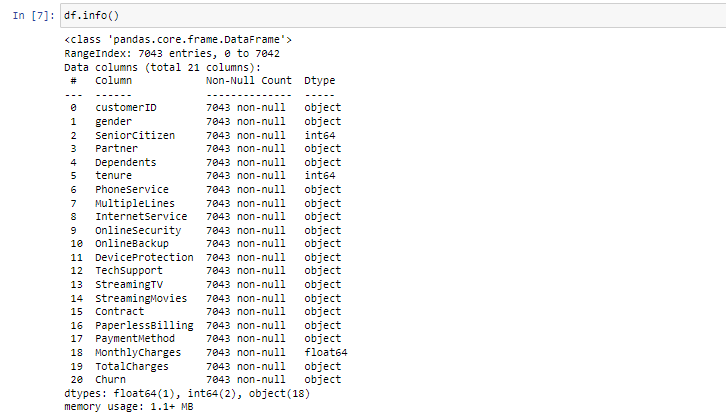
### **Loading the Dataset**

df=pd.read\_csv('D:/I/02.Flip Robo Assignments/00.DataTrained Assignment/01.Evaluation Phase/00.Projects/3rd Phase/Telecom\_customer\_churn.csv')



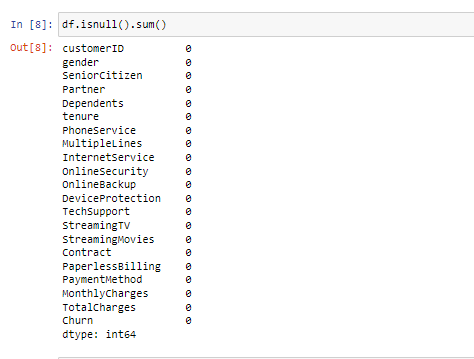


**Information about Dataset: -**



As mentioned above dataset we have 20 columns. Out of these, only 3 columns are numerical data.

### **Checking Null Values: -**



As above checking no null values are present in the dataset, so need to fill the null values.

Missing value imputation is one of the biggest challenges encountered by the data scientist.

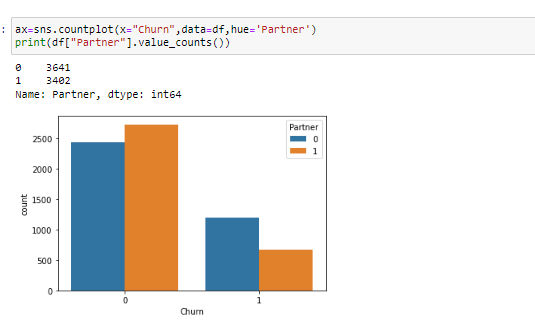
There are two methods of handling missing data, namely ignoring missing data and imputation of missing data. Ignoring missing data is a simple technique which deletes the cases that contain missing data.

**3. EDA Concluding Remark: -**

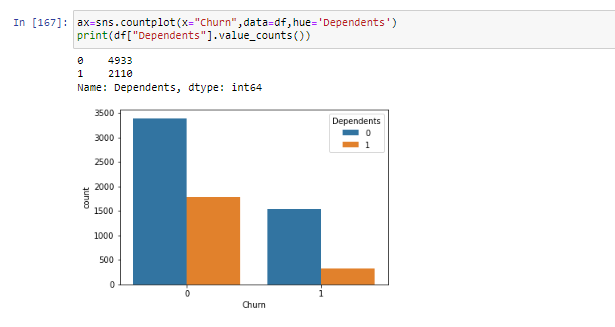
For Categorical data using the count plot to plot a graph against the Churn Column.



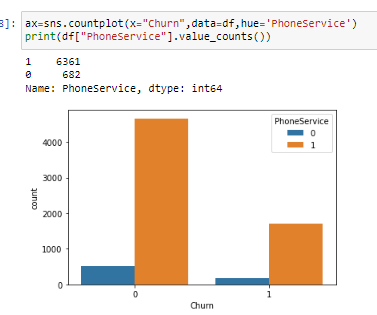
As checked above graph gender is not contributing the Churn.Both the gender values are same.



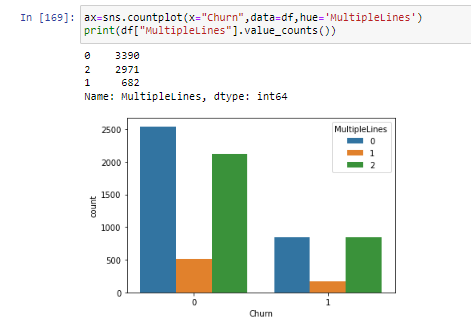
As checked above graph partner is little contributing the Churn.



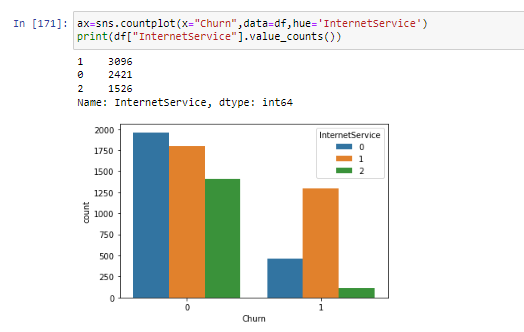
As checked above graph Dependents is 0 value is more contributing the Churn.



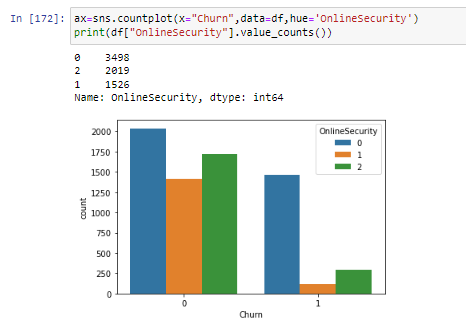
As checked above graph PhoneService is 1 value is more contributing the Churn.



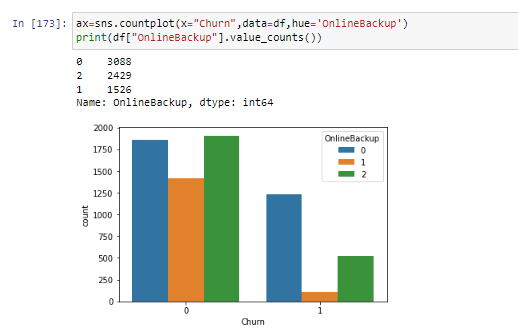
As checked above graph MultipleLines is 0 & 2 value is more contributing the Churn.



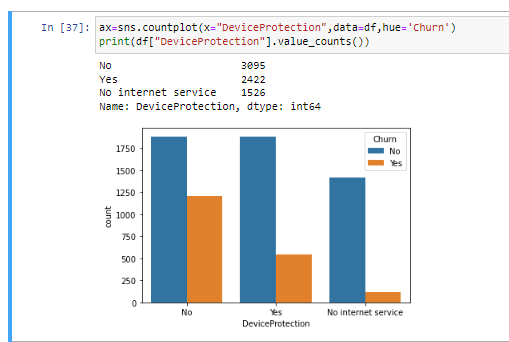
As checked above graph MultipleLines is 1 value is more contributing the Churn.



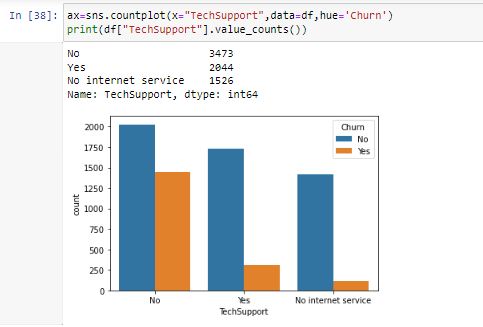
As checked above graph OnlineSecurity is 0 value is more contributing the Churn.



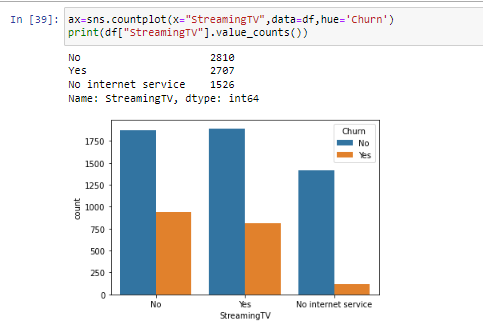
As checked above graph OnlineBackup is 0 value is more contributing the Churn.



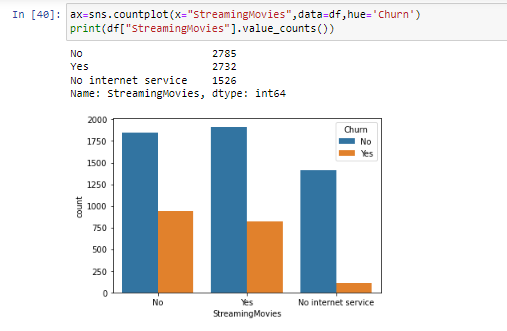
As checked above graph DeviceProtection is No value is more contributing the Churn.



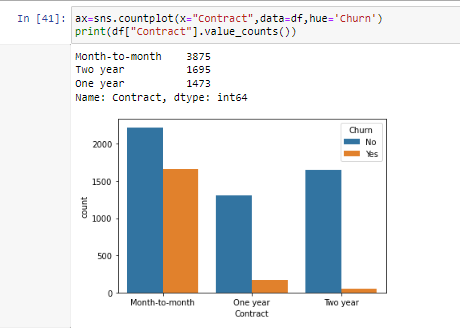
As checked above graph TechSupport is No value is more contributing the Churn.



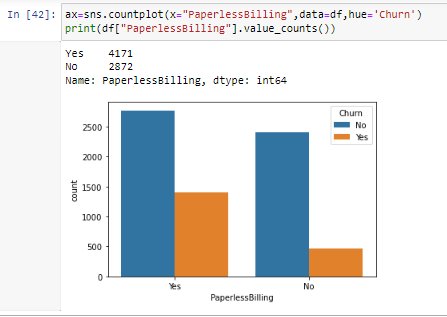
As checked above graph StreamingTV is No value is more contributing the Churn.



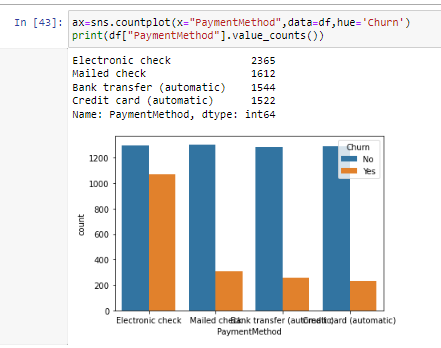
As checked above graph StreamingMovies is No value is more contributing the Churn.



As checked above graph Contract is No value is more contributing the Churn.

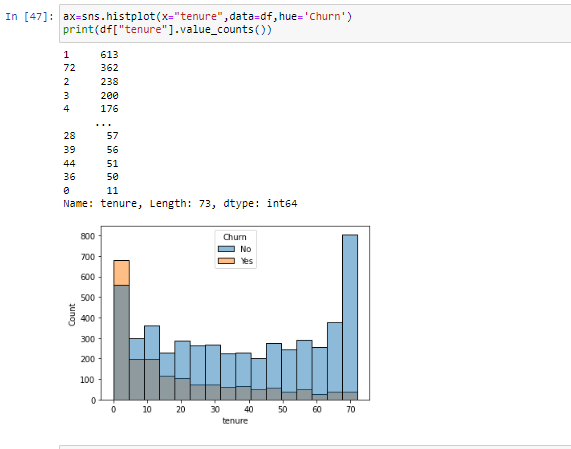


As checked above graph PaperlessBilling is No value is more contributing the Churn.

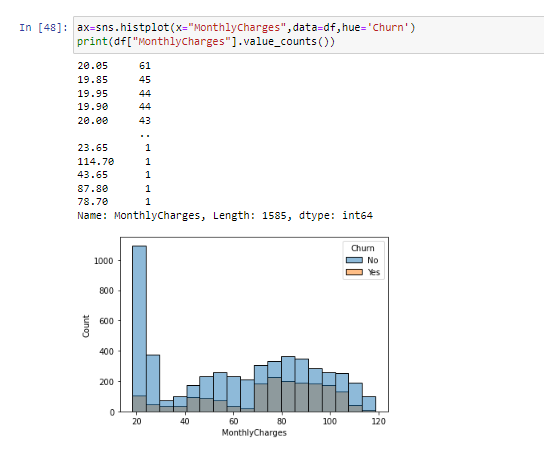


As checked above graph PaymentMethod is No value is more contributing the Churn.

For Numerical data using the hist plot to plot a graph against the Churn Column.



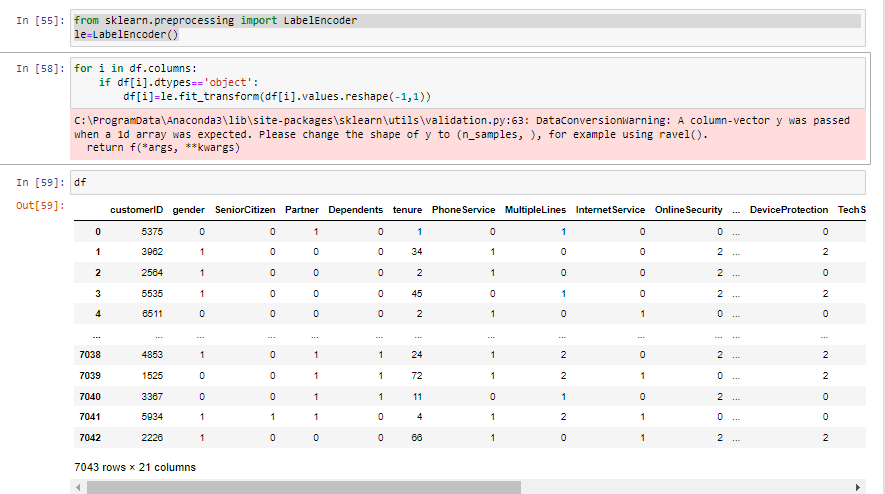
As checked above graph tenure customers who don’t have tech support have churned more.



As checked above graph The churn amount is higher in the initial 5 months.

**4. Pre-Processing Pipeline: -**

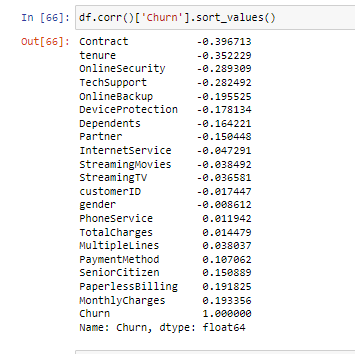
**Data Preparation:-**



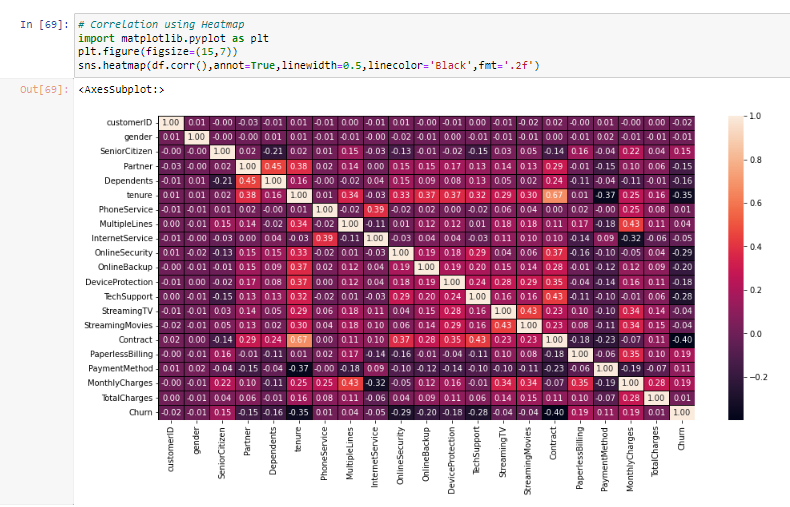
Machine Learning models do not work well with categorical inputs. So, we convert the categorical variables in our data set to numerical values by using one-hot encoding(LabelEncoder).

Checking Correlation: -

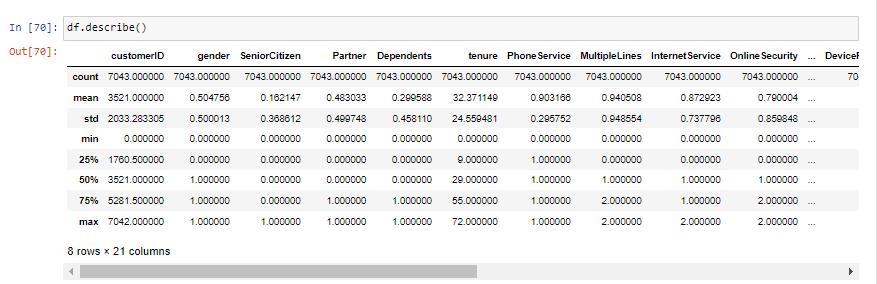
Correlation analysis defines the strength of a relationship between two variables, which can be between two independent variables or one independent and one dependent variable.

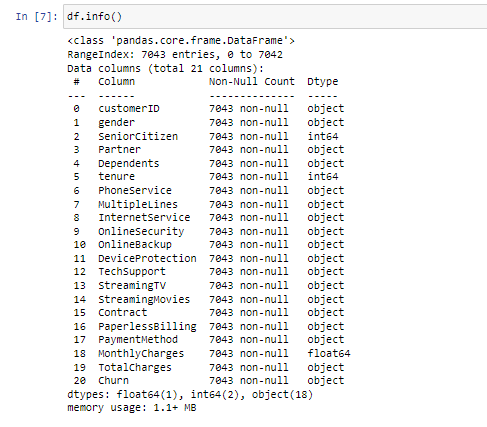


Checking Correlation using Heatmap: -



Describing Values: -

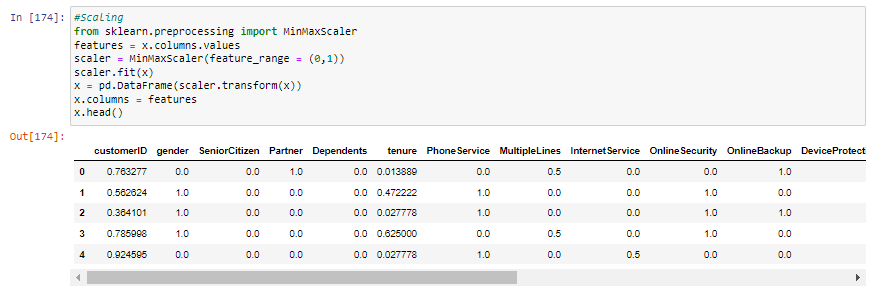




As checked above datatypes most of the datatype is categorical, so no need to check Outliers & z-score method.

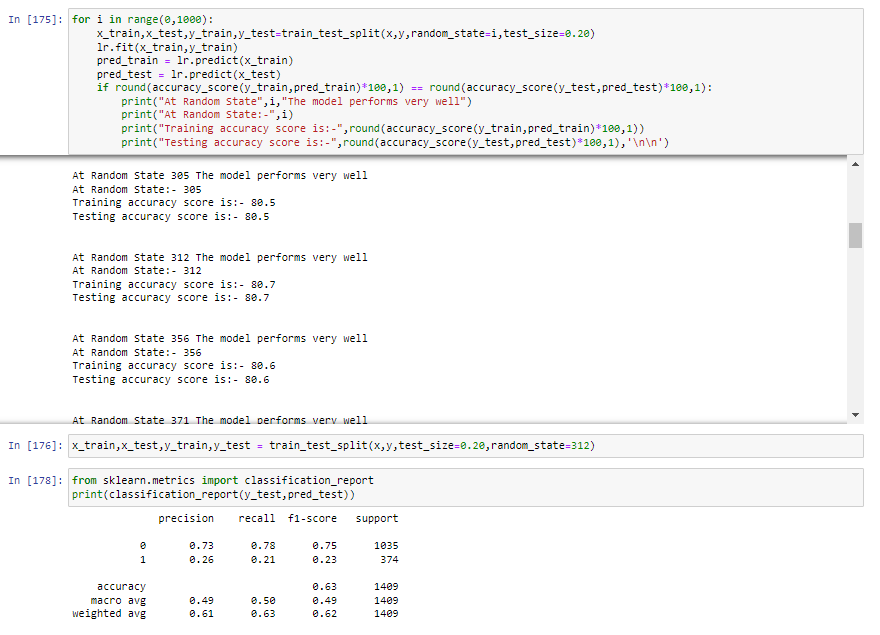
**5. Building Machine Learning Models: -**

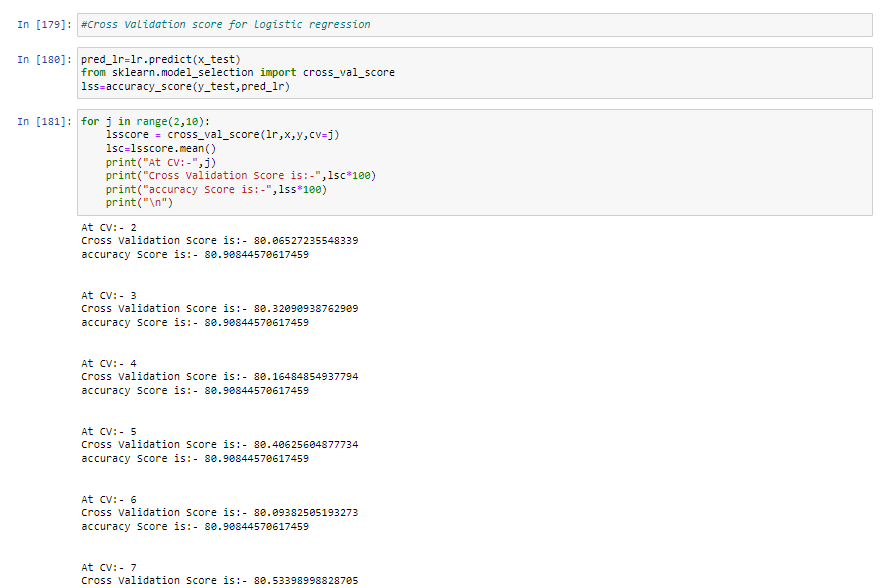
**Scaling the dataset using MinMaxScaler: -**

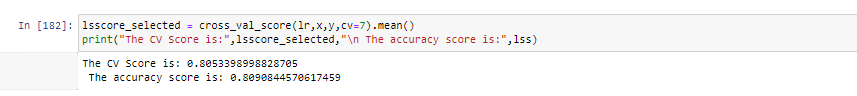


Prediction: -

Splitting the data into 2 datasets training & testing







Below we are checking prediction using other algorithms like

KNeighborsClassifier –

The K in the name of this classifier represents the k nearest neighbors, where k is an integer value specified by the user. Hence as the name suggests, this classifier implements learning based on the k nearest neighbors. The choice of the value of k is dependent on data.

DecisionTreeClassifier-

Decision tree classifiers are **supervised machine learning models**. This means that they use prelabelled data in order to train an algorithm that can be used to make a prediction. Decision trees can also be used for regression problems.

RandomForestClassifier-

Random forest is a supervised learning method, meaning there are labels for and mappings between our input and outputs. It can be used for classification tasks like determining the species of a flower based on measurements. When performing a classification task, each decision tree in the random forest votes for one of the classes to which the input belongs.

AdaBoostClassifier-

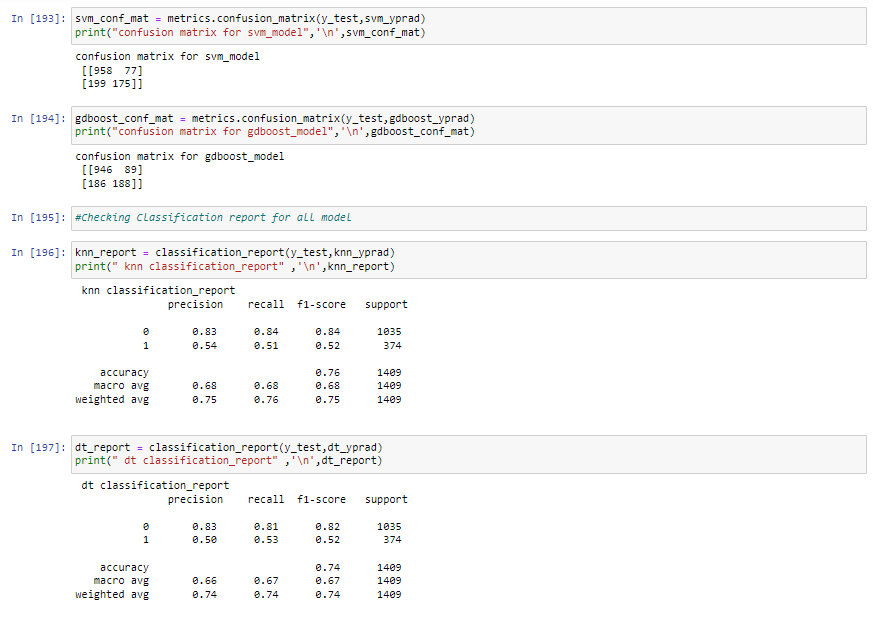
Ada-boost or Adaptive Boosting is one of ensemble boosting classifier. It combines multiple classifiers to increase the accuracy of classifiers. AdaBoost is an iterative ensemble method. AdaBoost classifier builds a strong classifier by combining multiple poorly performing classifiers so that you will get high accuracy strong classifier.

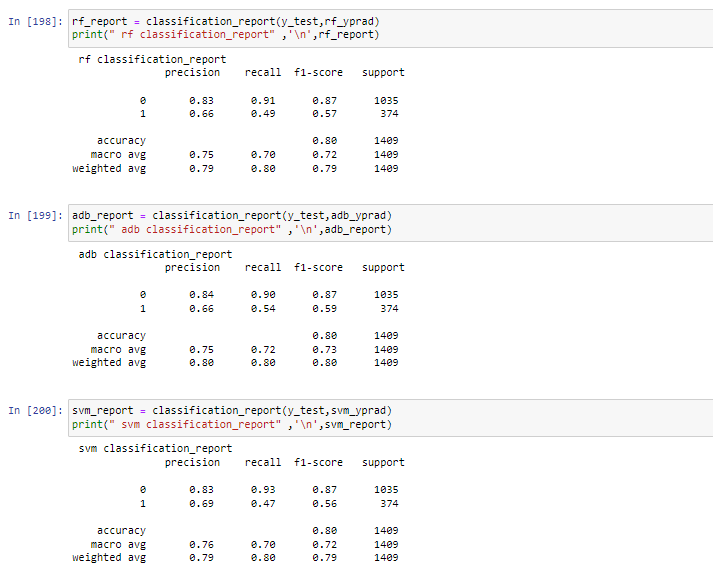
GradientBoostingClassifier-

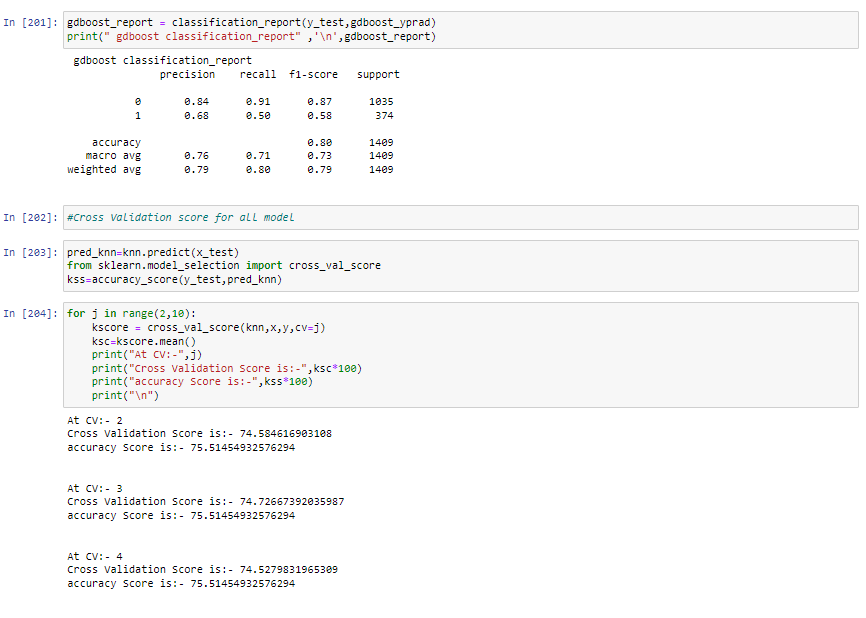
Gradient Boosting for classification. The Gradient Boosting Classifier is an additive ensemble of a base model whose error is corrected in successive iterations (or stages) by the addition of Regression.

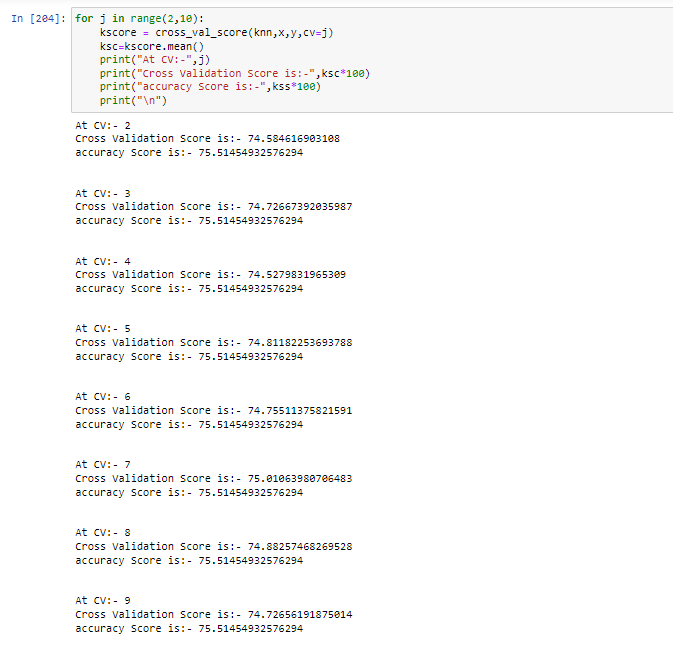


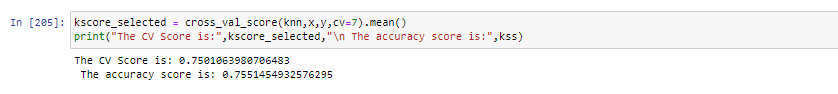


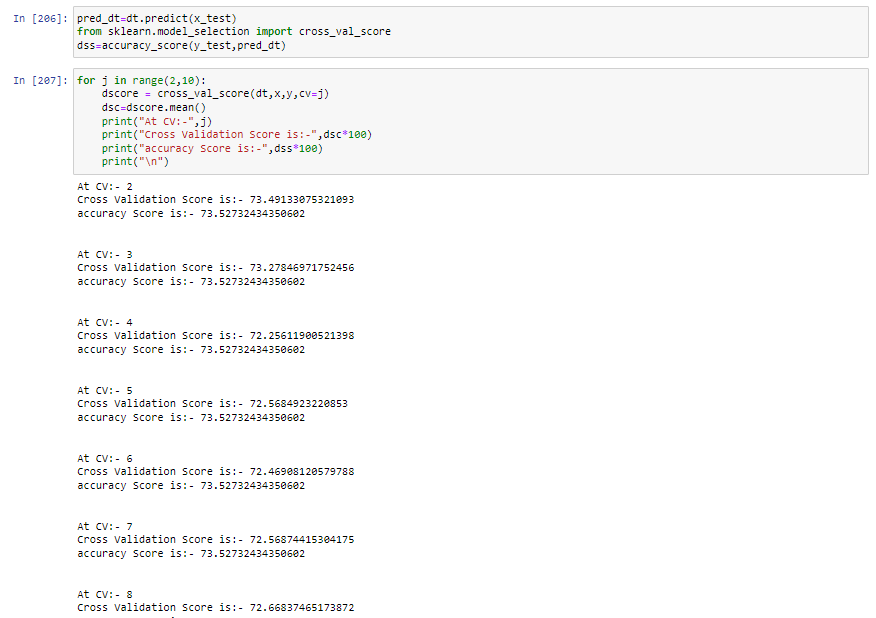


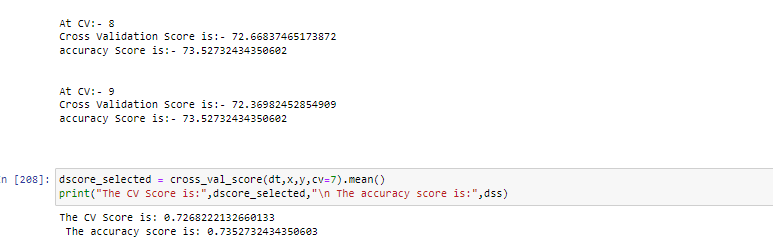


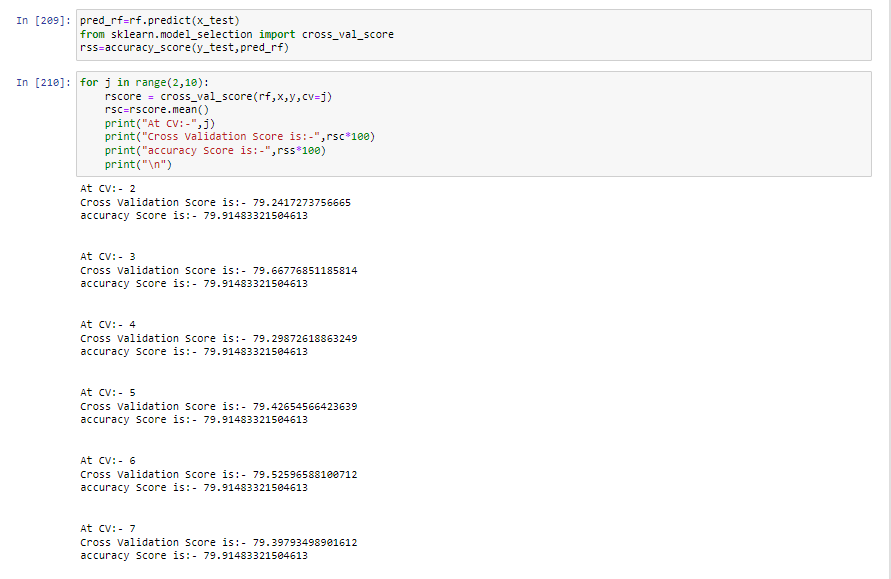


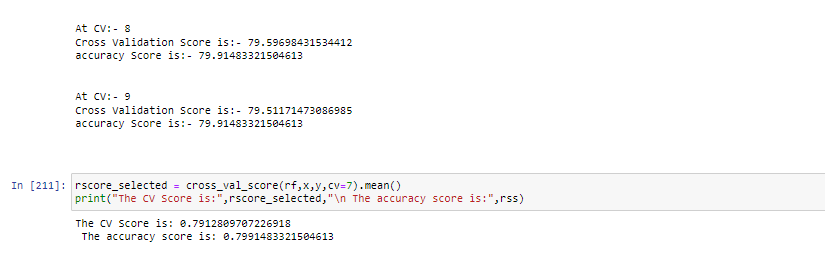


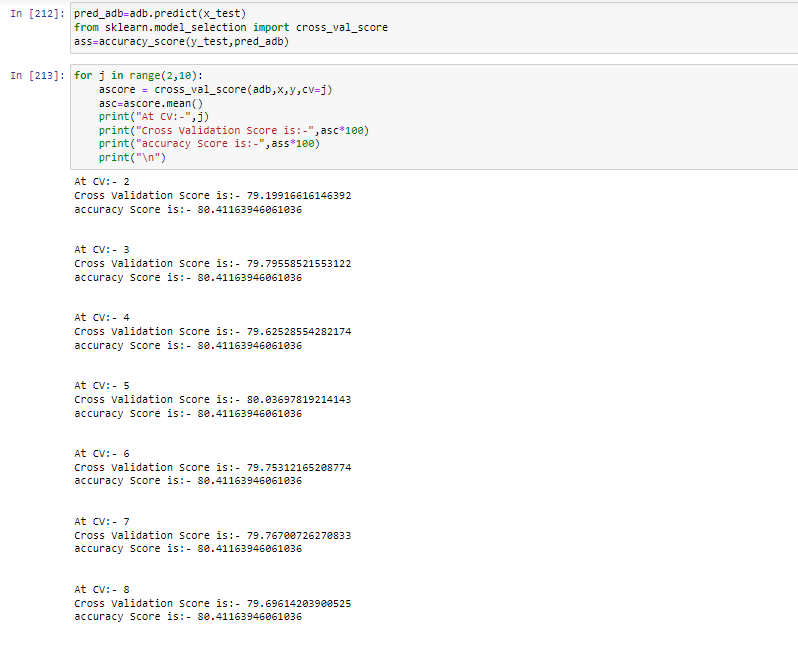


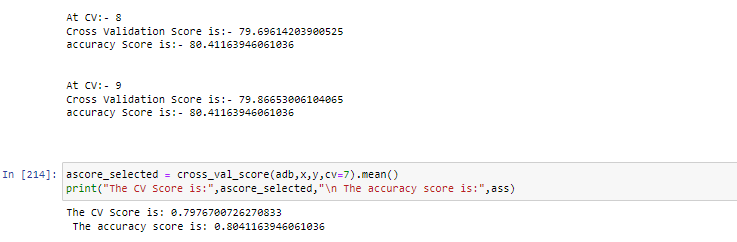


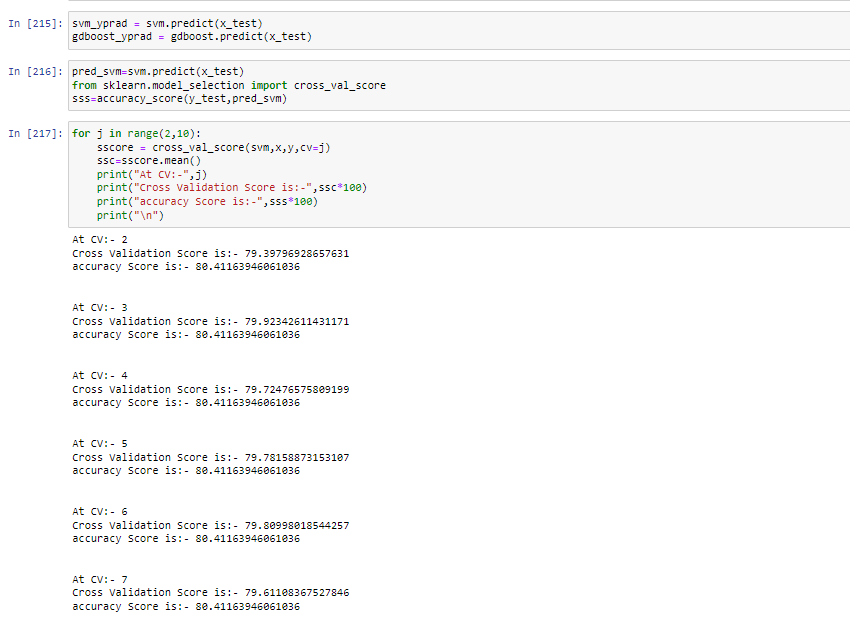


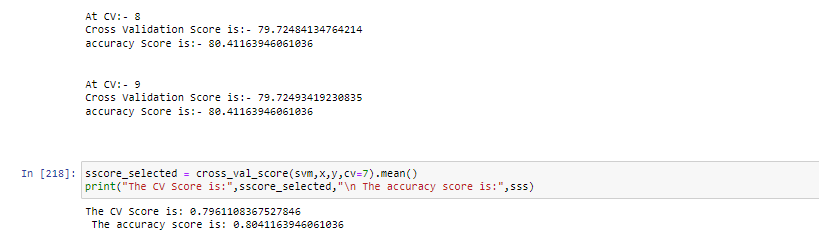


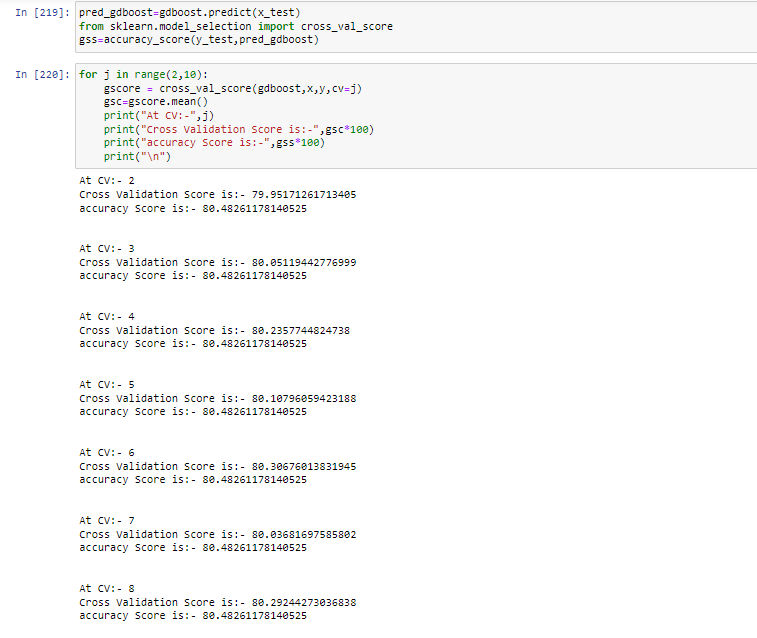


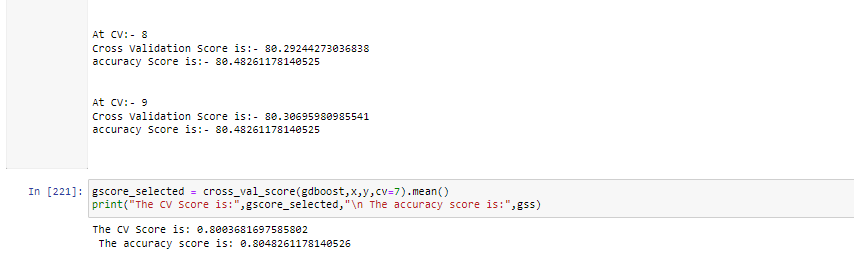












As above Accuracy Score for each prediction model is: -

Logistic Regression – 0.80

KNeighborsClassifier() – 0.75

DecisionTreeClassifier() – 0.73

RandomForestClassifier() – 0.79

AdaBoostClassifier() – 0.80

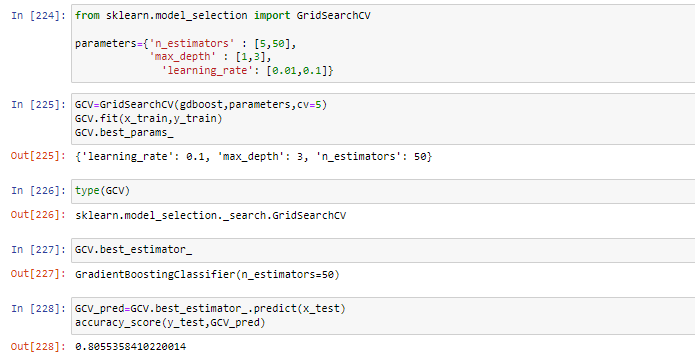
SVC() - 0.80

GradientBoostingClassifier() – 0.80

As compare above model GradientBoostingClassifier is highest accuracy.

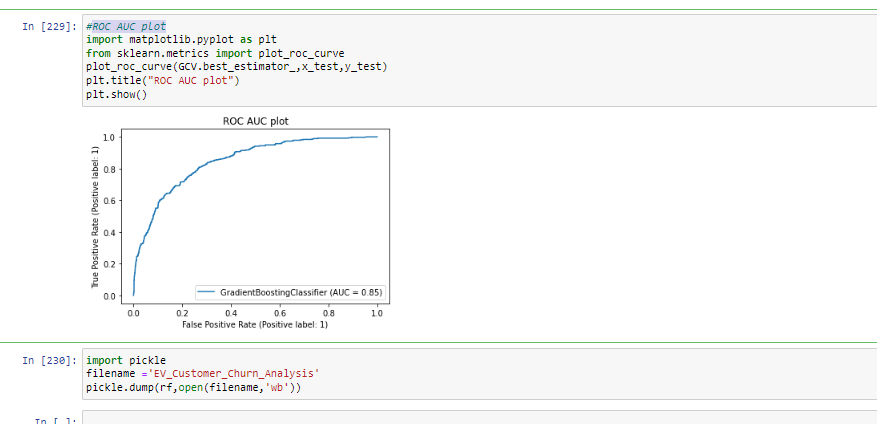
**Using Hyper Parameter Tuning : -**

In machine learning, hyperparameter optimizationor tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process.



**5. Concluding Remarks: -**

**Using ROC AUC plot : -**



As checked above ROC AUC plot using GradientBoostingClassifier() Accuracy is 0.85%.

It is very difficult to find Churn prediction but as soon as possible tried to find churn prediction using above data.