# **Machine Learning Case Study Customer Churn Prediction.**

**Introduction :**

According to telecom companies their aim is to attract the new customers and also retain the existing customers.There are many ways to attract the customers like giving offers,value added services etc.Similarly there are few reasons for the customer to terminate the service for example better price offers,service quality and customers personal situations.

In customer churn prediction analysis we are going to see the customers capabilities and also the various parameters that drive the customers perception.

Telecom companies approach the Data Analysis team for analysing and predicting the various offers to be given to the customers for retention and to improve the revenue.we are going to predict the analysis by the given dataset and analyse whether the customer will stay back or not.

Customer Churn Analysis comes under Supervised Learning and it's a Classification Algorithm Problem.The prediction and model we select to solve the problem depends on the target variable.Since we are going to calculate the customers opinion and its binary value we are going to solve using the classification model. By learning new facts and new skills we are going to solve the problems more efficiently and accurately.

In this article we are going to see the following steps :

Step 1: Problem Definition

Step 2: Data Analysis

Step 3: EDA Concluding Remarks

Step 4: Pre-processing Pipeline

Step 5: Building Machine Learning Models

Step 6:Concluding Remarks

Step 1: Problem Definition:

Our task is to analyse key parameters that drive the company to increase their revenue and get hold of the existing customers. We are going to check the features and provide the prediction according to the dataset given.

Comparing the models and selecting the best model for this task accuracy is measured.Based on the characteristics of dataset we are going to calculate the "Customer Churn" by considering the metrics used in the model.

Step 2: Data Analysis

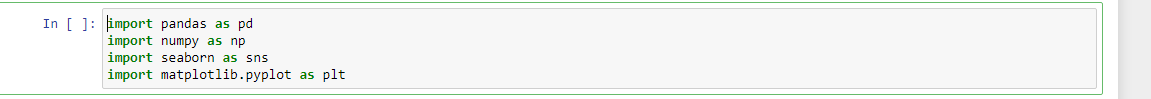
The dataset for this Regression Problem is taken from the IBM Samples Data sets with the aim of building and comparing several customer churn prediction models.

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

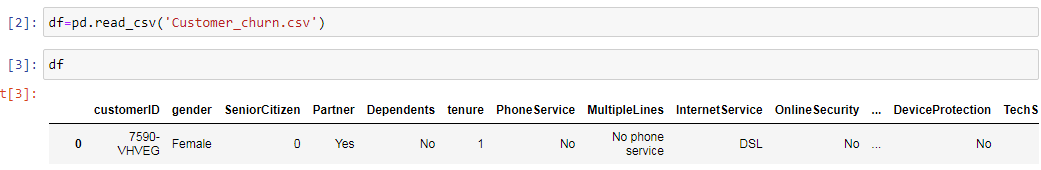
**Download Link:**

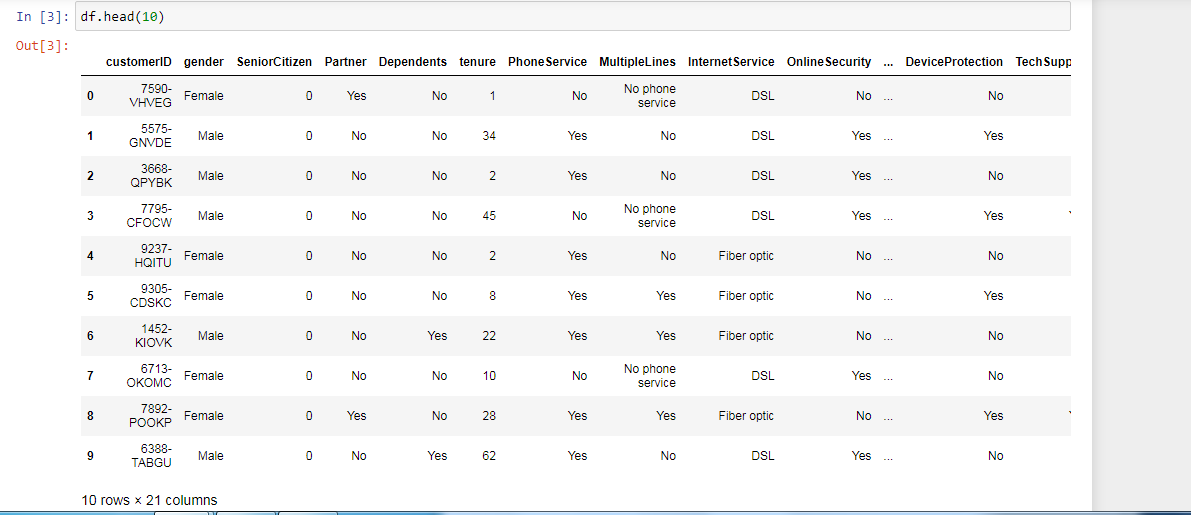
* [**https://github.com/dsrscientist/DSData/blob/master/Telecom\_customer\_churn.csv**](https://github.com/dsrscientist/DSData/blob/master/Telecom_customer_churn.csv)

The Data Analysis is started up by using the basic libraries.We are going to use Numpy,Pandas,Seaborn Visualization and Matplotlib,Encoder,Standard Scaler Method etc.

****

For this exercise, the data set (.csv format) is downloaded and read into the Jupyter notebook and stored in a Pandas DataFrame.

****

****

Step 3: Exploratory Data Analysis Concluding Remarks:

Once data collection is done we perform several steps to explore the data.In this step we are going to an understanding of the data structure ,doing initial preprocessing ,clean the data,identify the patterns and features of the data to finally build and validate hypothesis.

Exploring given data:

In this part of EDA the data frame is evaluated for number of columns,data types and null values in the dataset.The aim of this step are to get a general understanding of the data set,check domain knowledge and first get the dataset idea on topics for further investigatation.Here we use some basic pandas library.

Defining Each Features in the Dataset.

By Analyzing the columns and their unique values we are grouping them into category accordingly:

Customer's Information on the Dataset:

Tenure--->How many months does the customer having the service connection.

Total Charges---> It is the total amount charged from the customer according to the services they are using the charges are provided.

Services provided to the Customer:

Phone Service--->Whether the customer is having the service connection (Yes/No)

MultipleLines --- >Whether the customer is having multiple lines of services (Yes, No, No phone service)

InternetService--->what type of internet connection does the customer hold (DSL, Fiber optic, No)

OnlineSecurity --- >Does the customer is having online security protection (Yes, No, No internet service)

OnlineBackup--->Does the customer is having online backup with hand(Yes, No, No internet service)

DeviceProtection--- >Is the customer having device protection (Yes, No, No internet service)

TechSupport --->Is the customer has tech support (Yes, No, No internet service)

StreamingTV --->Does the customer is having streaming TV service(Yes, No, No internet service)

StreamingMovies--->Does the customer is having streaming movies package along with the service(Yes, No,No Internet service).

Contract--->Term billing of the customer(month-month etc.)

Paperless Billing--->If the customer is paying online or through direct billing method.

Payment Method--->The customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)).

MonthlyCharges---> Monthly charges by the customer.

Total Charges--->Total charges paid by the customer for the period of time.

Customer's Info:

customerID---> Customer ID

Gender---> Either the Customer is a male or a female.

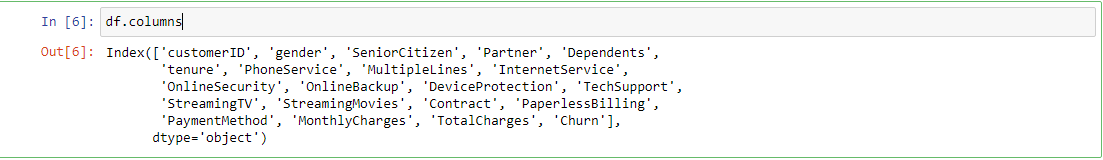
SeniorCitizen---> Either the Customer is a senior citizen or not (1,0).

Partner --->Whether the customer has a partner or not (Yes, No).

Dependents ---> Whether the customer has dependents or not (Yes, No).

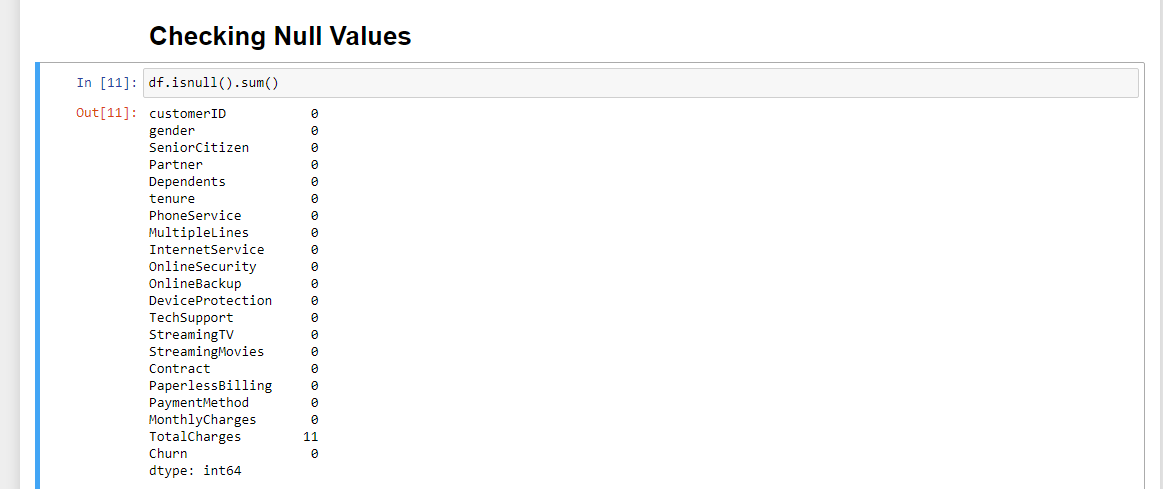
Step 3: EDA Concluding Remarks

We are going to get printed all the column names,data types and shape of the dataset.

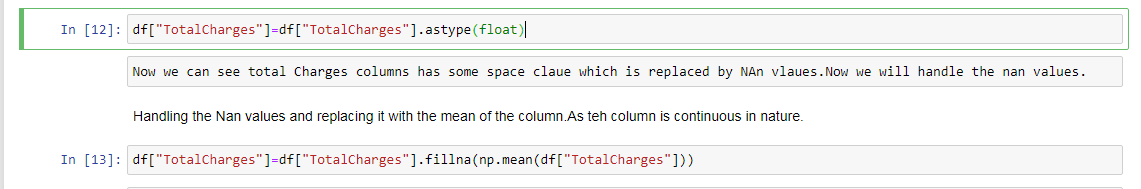


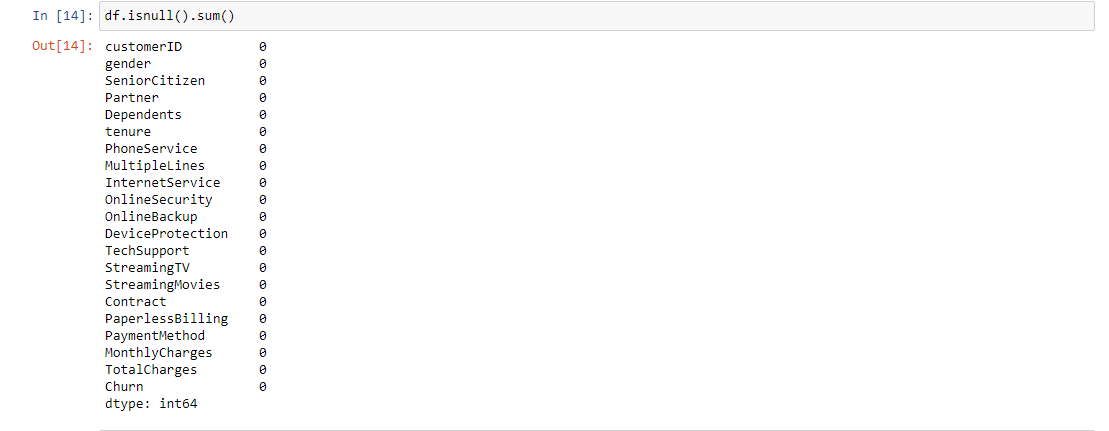


Checking For Null Values in the dataset,

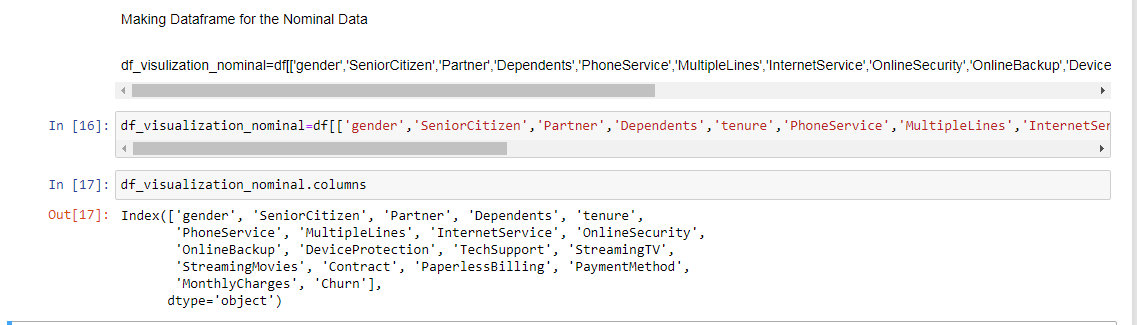


We have null values in the total charges column and the datatype is in object type,we shall handle this.

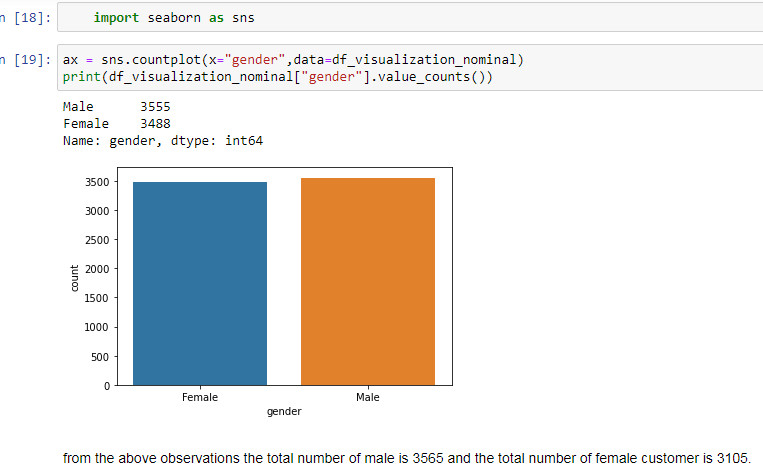
****

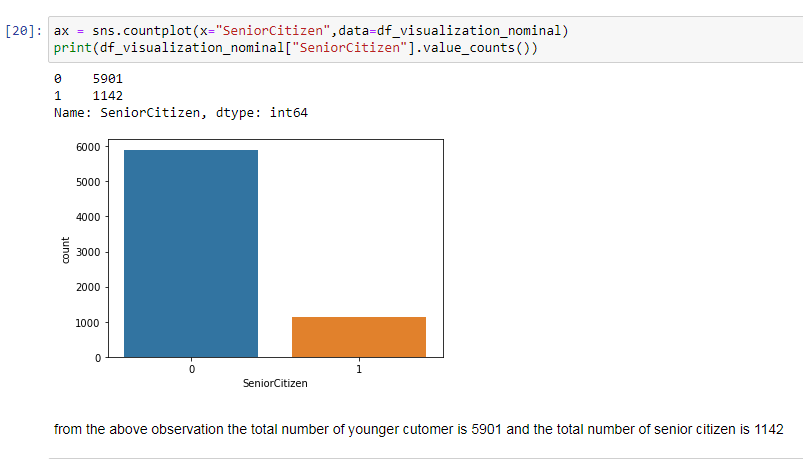
****

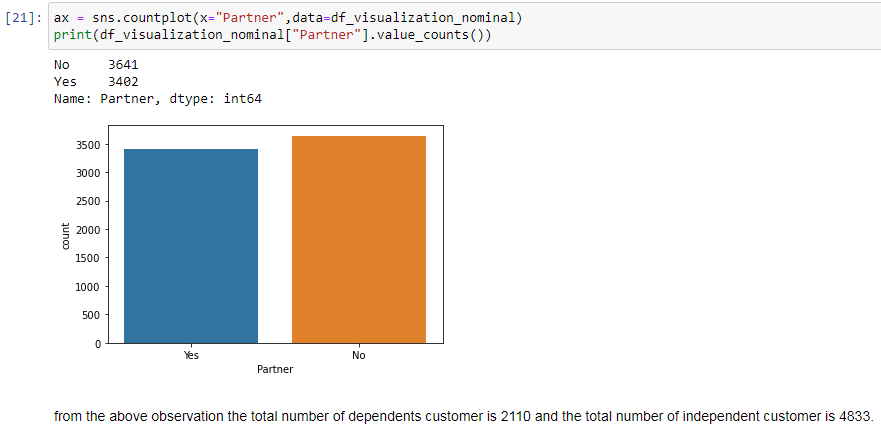
# Visualization of the Data:

****

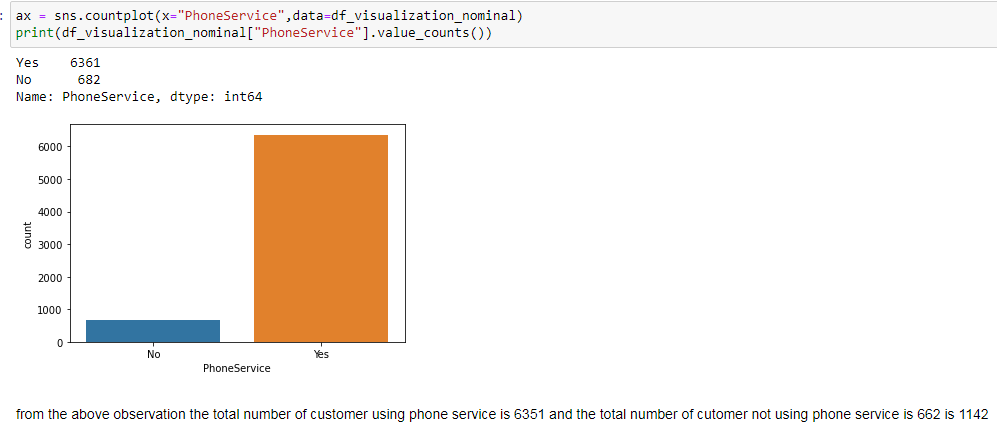
For the nominal categorical data we will use countplot as it will give the frequency of the classes of the columns.

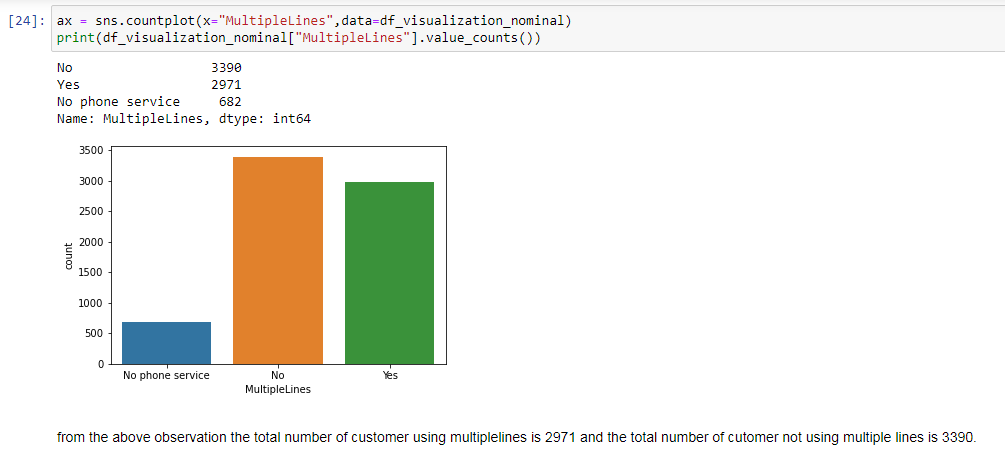
****

****

****

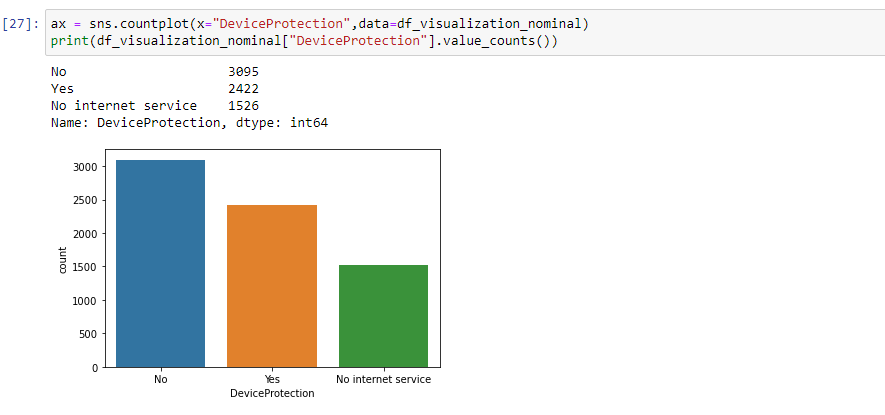
****

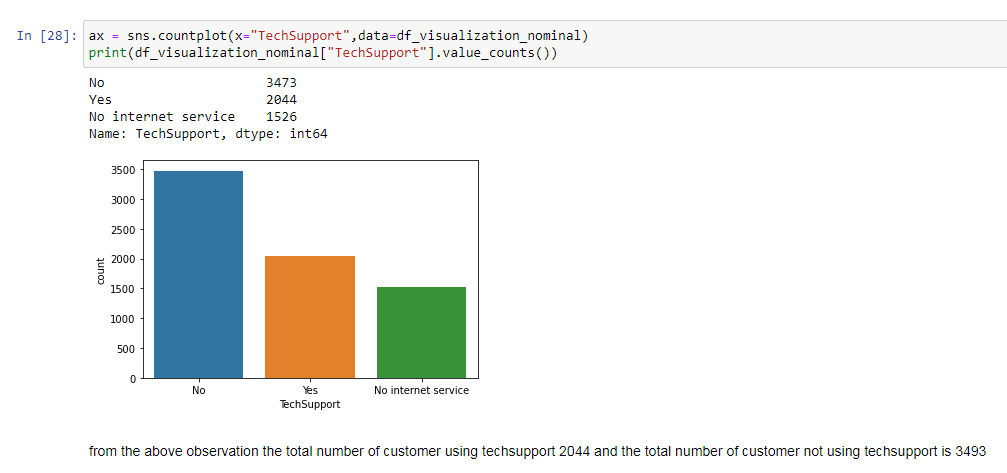
****

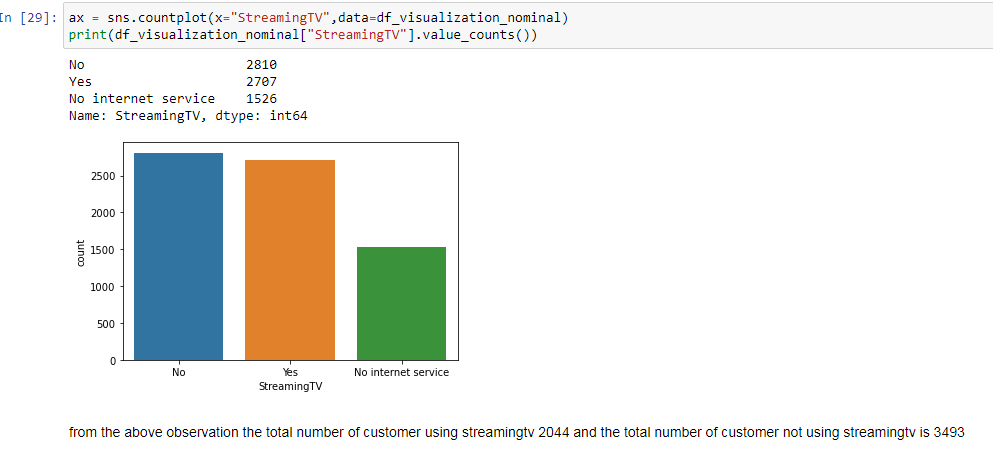
****

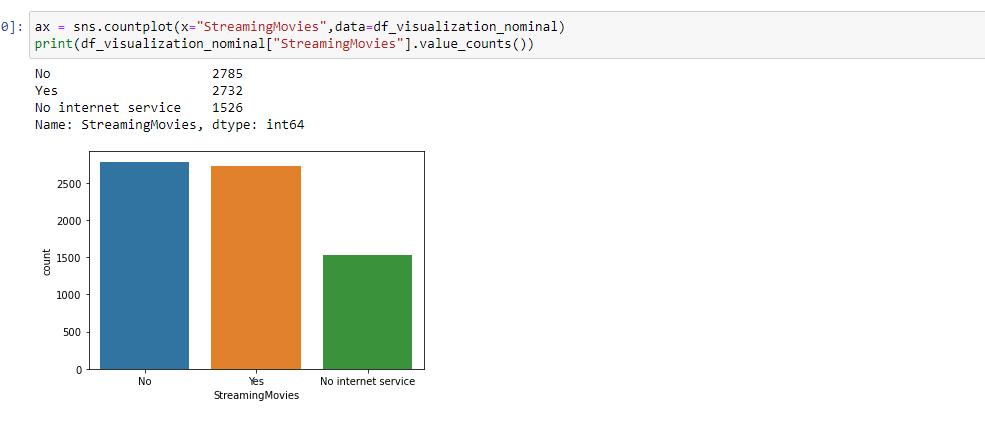
****

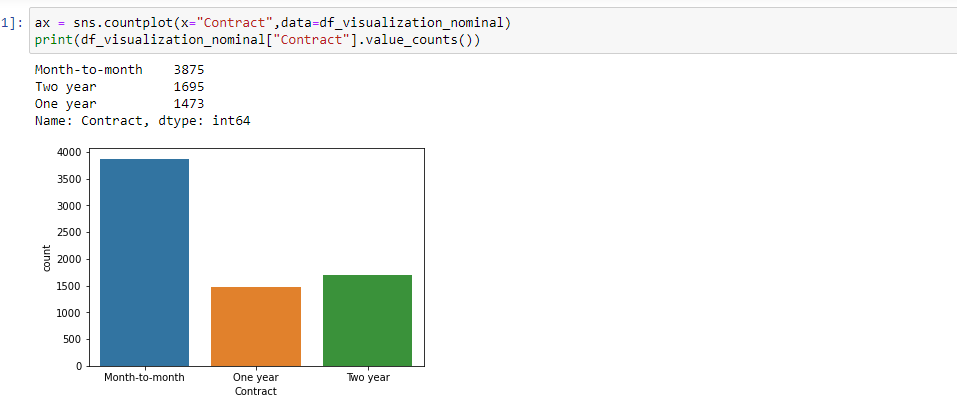
****

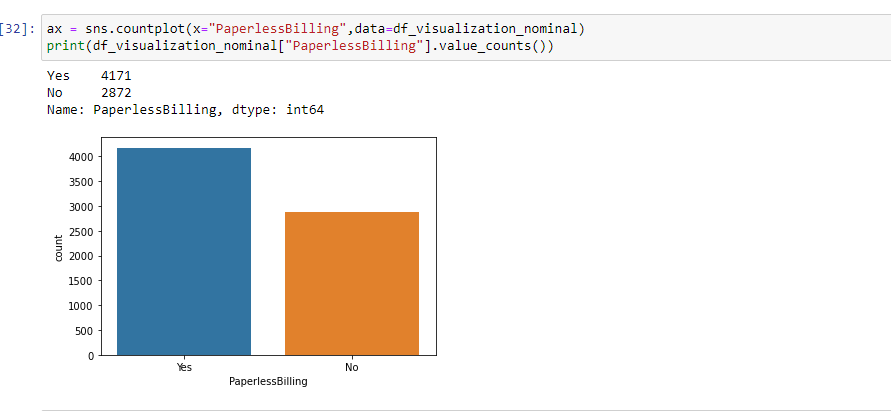
****

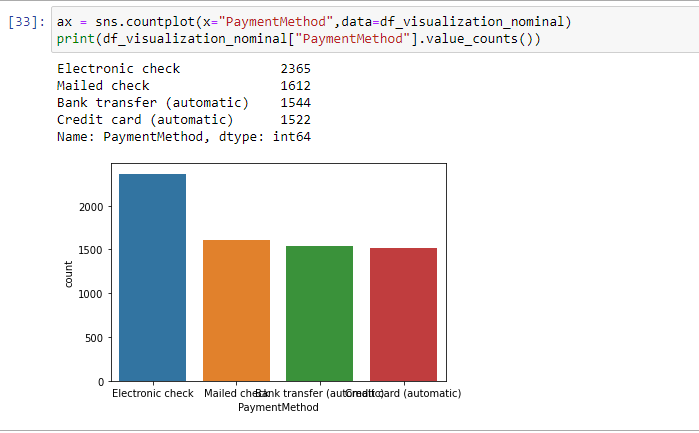
****

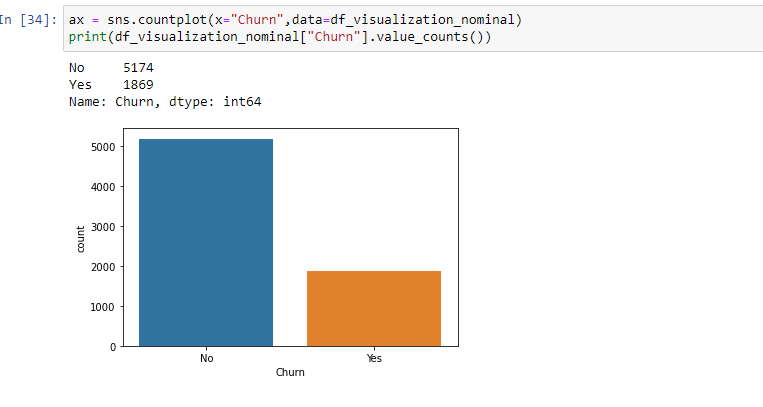
****

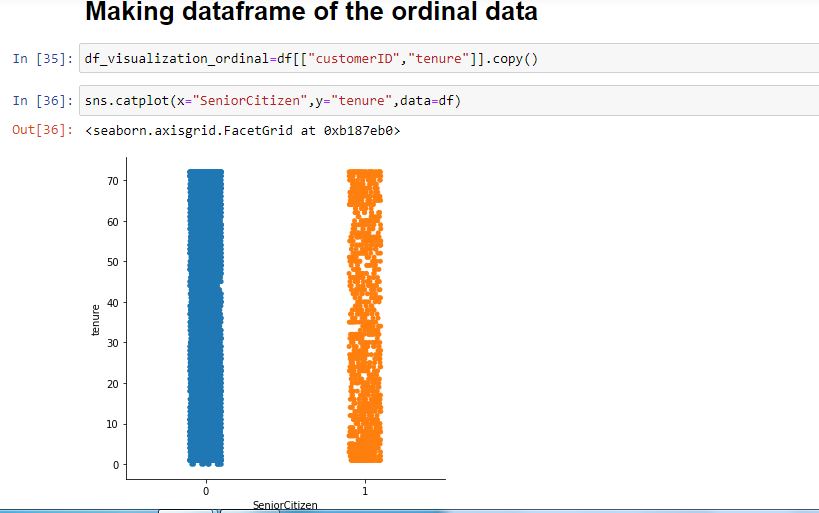
****

****

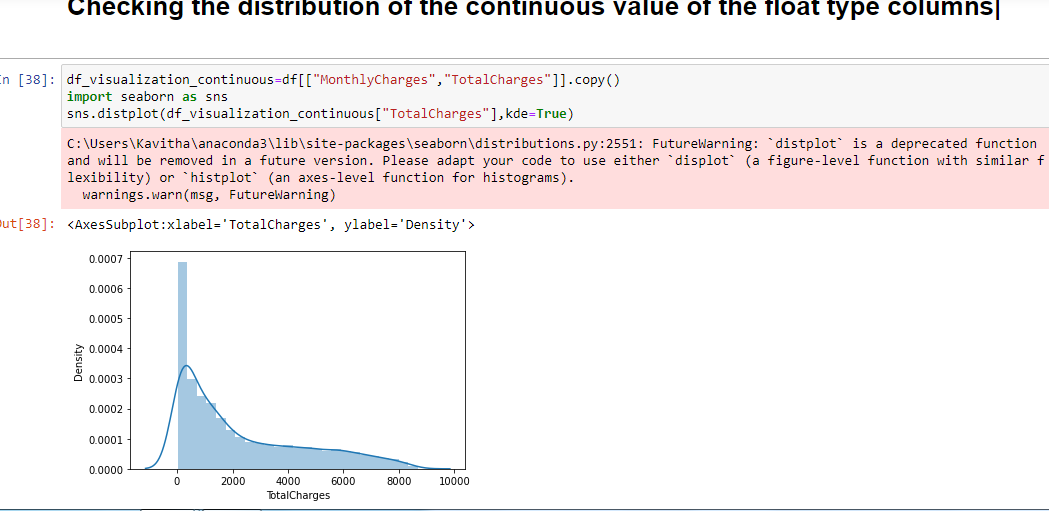
****

****

****

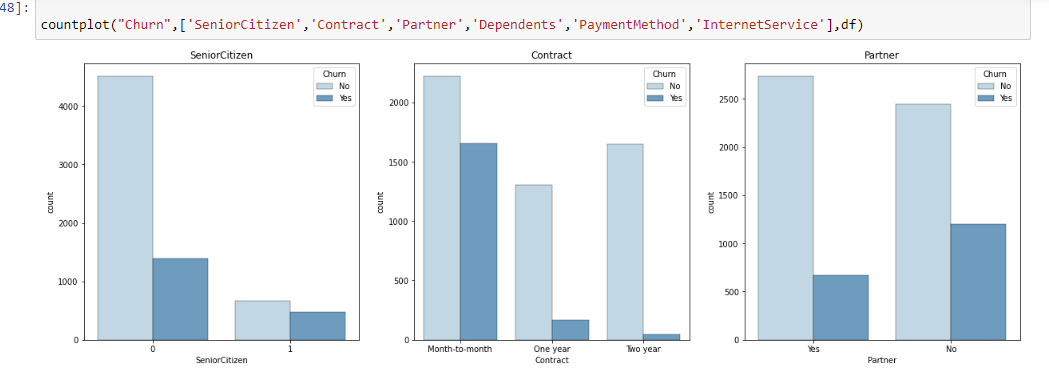
****

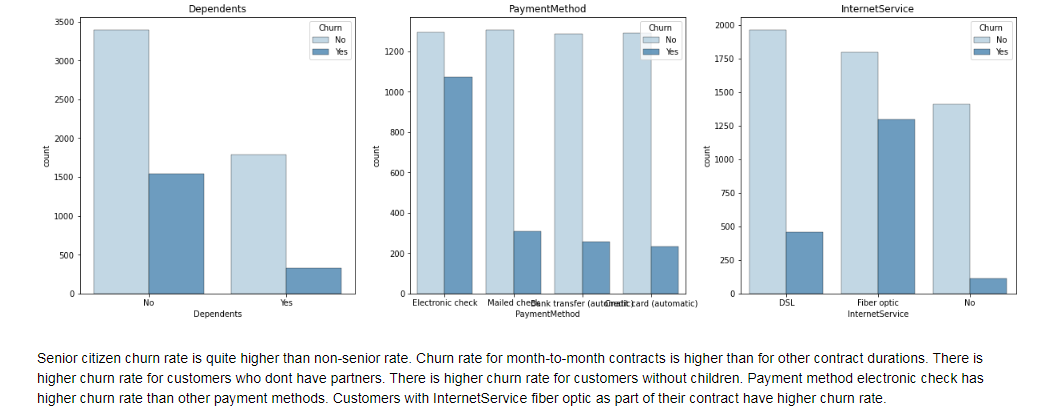
Now we are going to check the distribution plot of the target variable plotted against to its related columns.

****

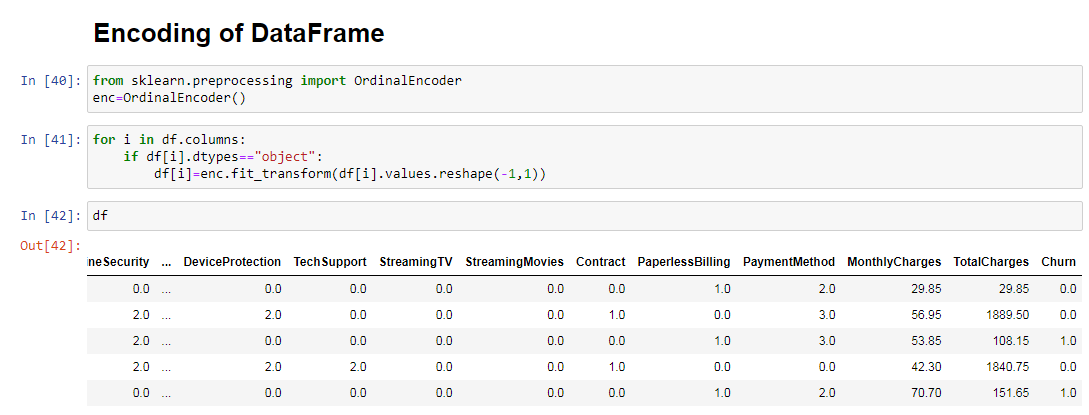
****

****

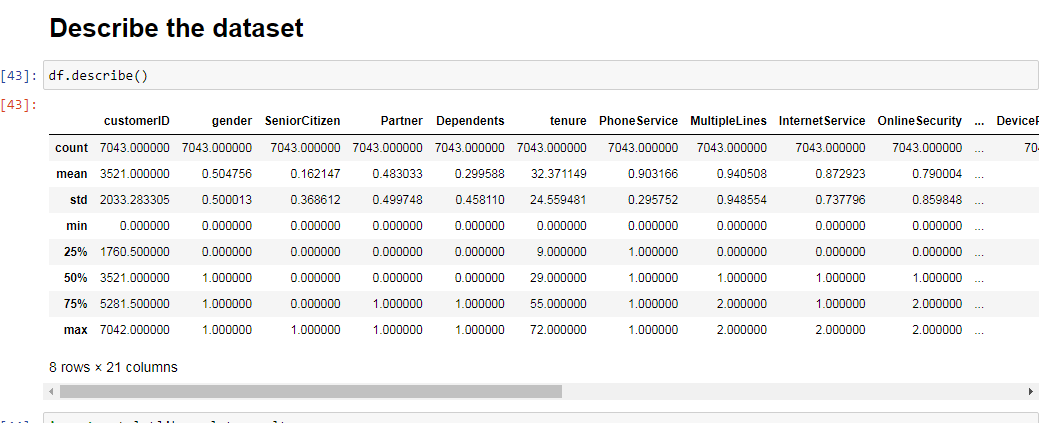
****

****

**Step 4: Pre-processing Pipeline:**

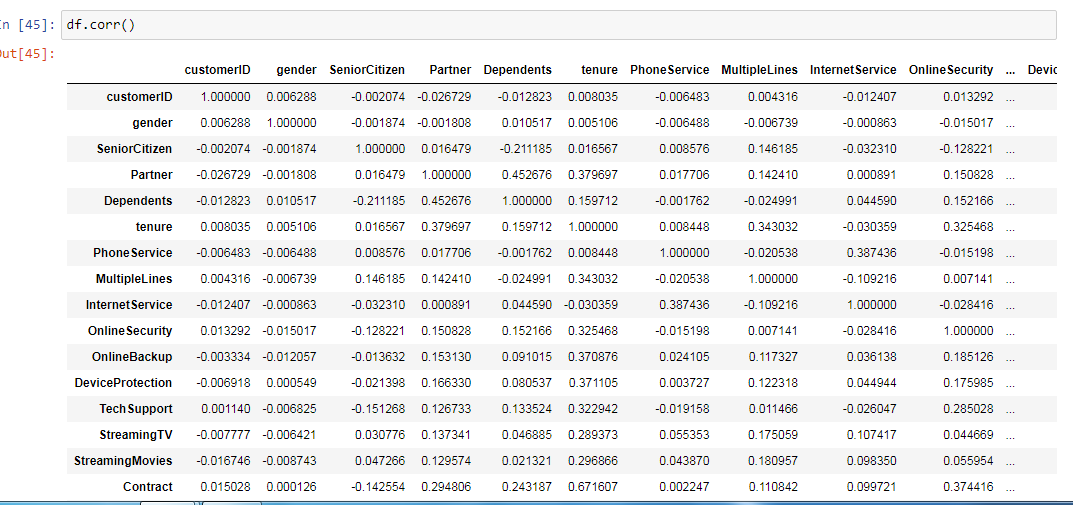
****

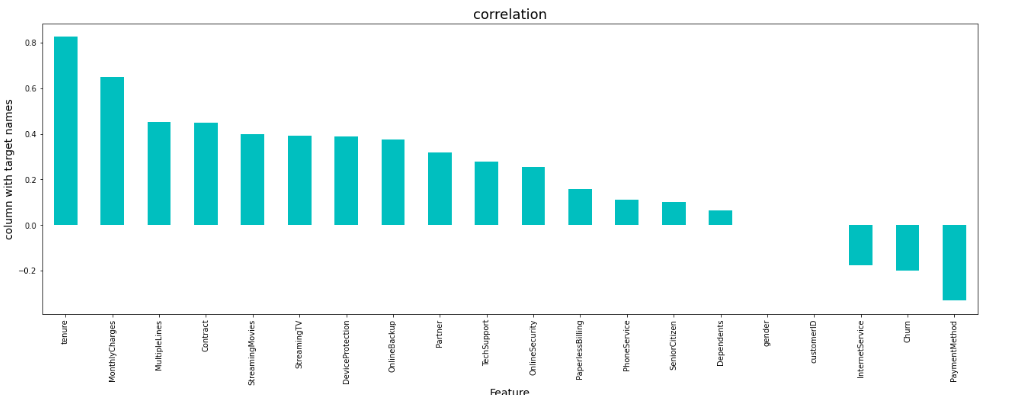
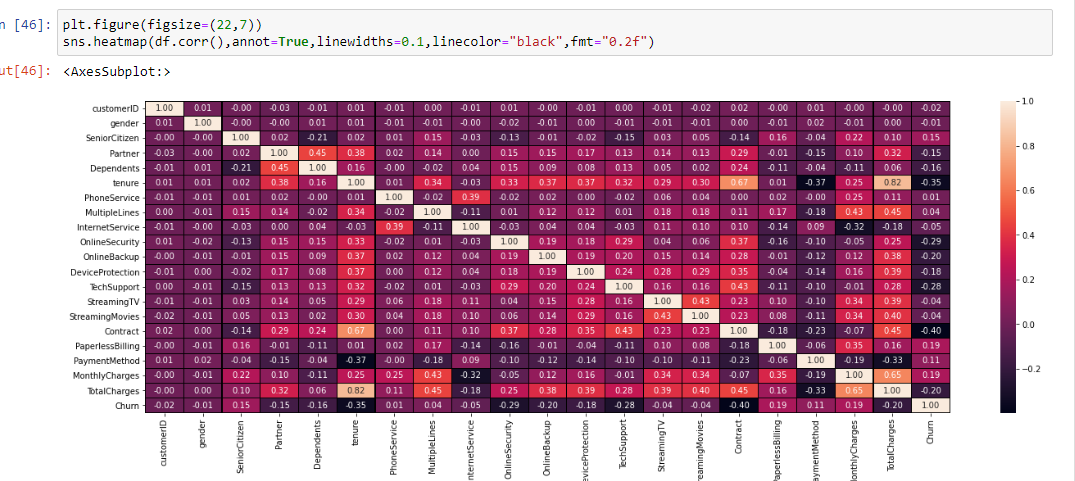
**Statistical Analysis of the Dataset:**

****

****

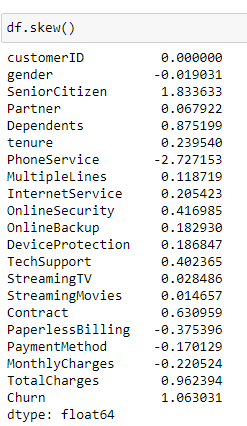
**Correlation of the Dataset:**

****

****

Checking the columns which are positively and negative correleated with the target columns:

**Skewness and Outliers :**

****

Keeping +/- 0.5 as the range for skewness here are the columns which does not lie within this range.

Senior Citizen -categorical

Dependents - categorical

Phone Services - categorical

Contract-categorical

Total charges-Numerical

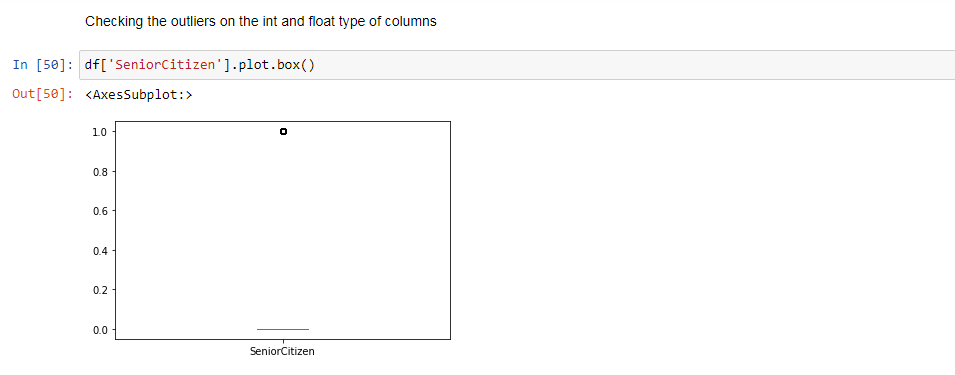
Churn-categorical

We are calculating the skewness in the dataset.we shall remove these outliers by the z score method.

Z score method:

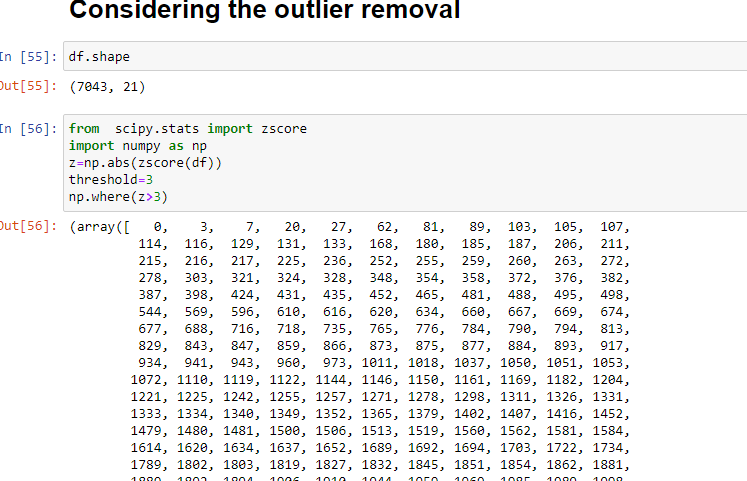
Z score is also called standard score. This score helps to understand if a data value is greater or smaller than mean and how far away it is from the mean. More specifically, Z score tells how many standard deviations away a data point is from the mean.

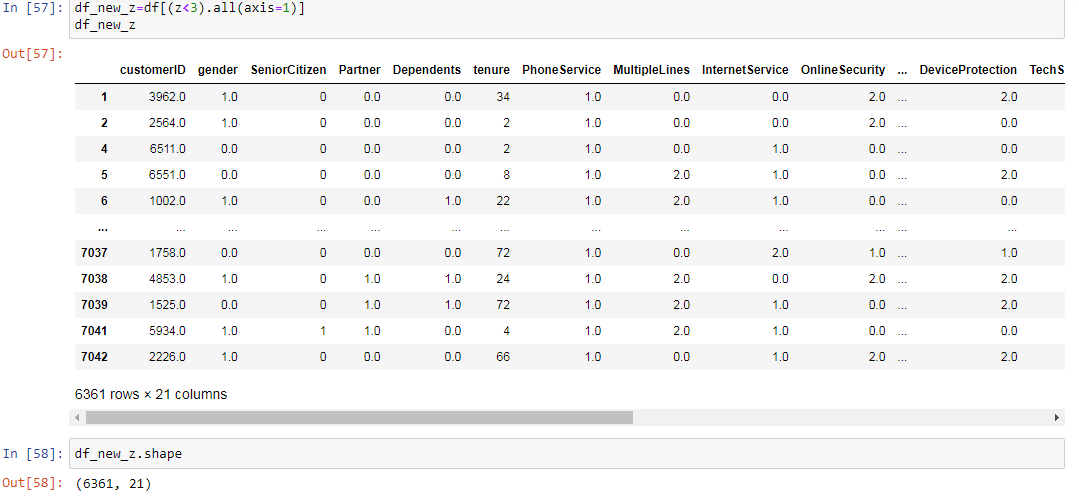
If the z score of a data point is more than 3, it indicates that the data point is quite different from the other data points. Such a data point can be an outlier.

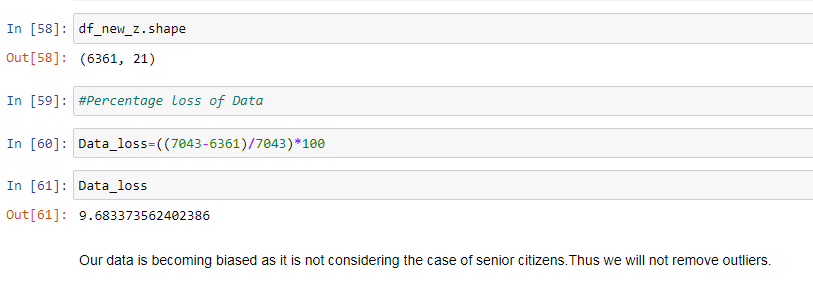
****

****

TotalCharges are the result of tenure and MonthlyCharges, which are more insightful on an individual basis.

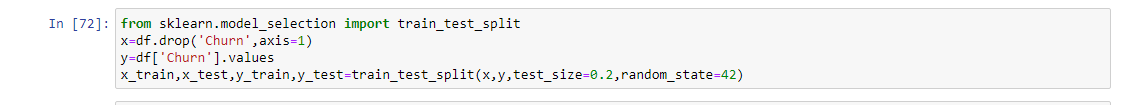
****

****

****

Train-Test-Split:

In train-test-split model training the data is split into 80% training data and 20% testing data.Here our target variable column "Churn" is defined as the class "y" and the remaining columns as the feature "x"**.**

****

Here now we are going todo the Model Buliding after splitting the train and test data.We can see the evaluation metrics and its parameter which we are going to use in our model building.

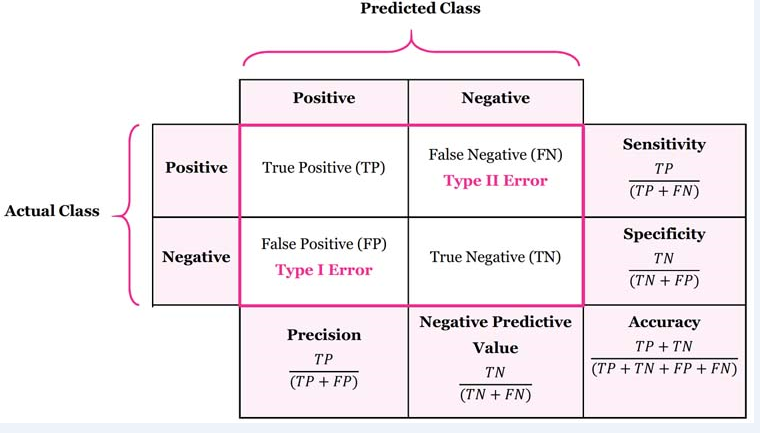
In Machine Learning, performance measurement is an important task.

This data comes under Supervised Machine Learning and typical Classification Problem.So the metrics involved here are:

1.Confusion Matrix:

It is an N\*N matrix used for evaluating the performance of a classification model,where N is the number of target classes.The matrix compares the actual target values against the predicted by the machine learning model.The rows represent the predicted values of the target variable.

It can be plotted as a grid of True and False predictions compared against the Actual values in the dataset.Confusion matrix is also used to visualize important predictive analytics like recall,precision and accuracy.It is very useful because they give direct comparsion of values like True Positives,False Positives,True Negatives and False Negatives.



2.Accuracy Score :

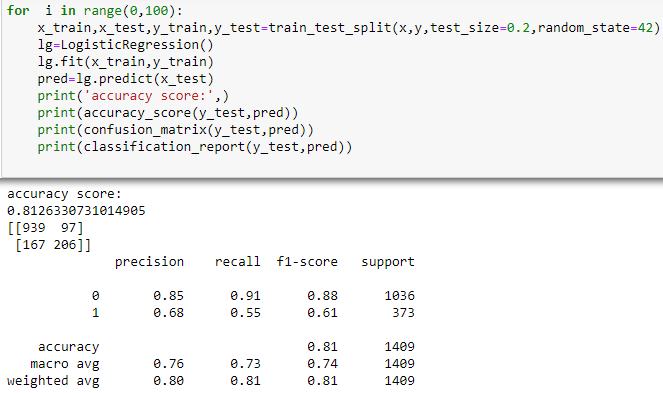
Classification accuracy is what we calculate while using the term accuracy.It is the ration of number of correct predictions to the total number of input samples.

Actually it gives fine score when there are equal number of samples in each class.

We can simple measure accuracy by number of correct decisions our classifer makes,divide by the total number of test examples and th result is the accuracy of our classifier.



3.Classification Report:The classification report visualizer displays the precision,recall,F1 and support scores for the model.



We are going to see the classification report calculation.

Precision:

Precision is the ability of a classifier not to label any instances positive which is actually negative. In each class, it is defined as the ratio of true positives to the sum of a true positive and false positive.

Precision:- Accuracy of positive predictions.

Precision = TP/(TP + FP)

Recall:

Recall is the ability of a classifier to find all positive instances. In each class it is defined as the ratio of true positives to the sum of true positives and false negatives.

Recall:- Fraction of positives that were correctly identified.

Recall = TP/(TP+FN)

F1 score:

The F1 score is a harmonic mean of precision , and recall such that the best score is 1.0 and the worst is 0.0. F1 scores are lower than accuracy measures as they include precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.

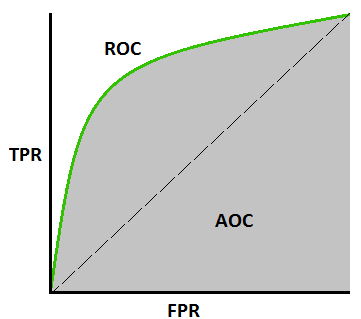
F1 Score = 2\*(Recall \* Precision) / (Recall + Precision).

4.AUC ROC Curve:

In measuring the performance evaluation of classification problem, we can count on an AUC - ROC Curve. When we need to check or visualize the performance of the multi-class classification problem, we use the AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve. It is one of the most important evaluation metrics for checking any classification model’s performance. It is also written as AUROC (Area Under the Receiver Operating Characteristics)

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.

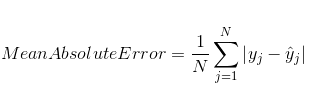
The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis.



An excellent model has AUC near to the 1 which means it has a good measure of separability. A poor model has an AUC near 0 which means it has the worst measure of separability. In fact, it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means the model has no class separation capacity whatsoever.

Mean Absolute Error:

Mean Absolute Error is the average of the difference between the Original Values and the Predicted Values. It gives us the measure of how far the predictions were from the actual output. However, they don’t gives us any idea of the direction of the error i.e. whether we are under predicting the data or over predicting the data. Mathematically, it is represented as :



Mean Squared Error:

The only difference being that MSE takes the average of the square of the difference between the original values and the predicted values. The advantage of MSE being that it is easier to compute the gradient, whereas Mean Absolute Error requires complicated linear programming tools to compute the gradient. As, we take square of the error, the effect of larger errors become more pronounced then smaller error, hence the model can now focus more on the larger errors.

Model Building Algorithms:

First we will test out several models and measure their performance by several metrics. Those models will be optimized in a later step by tuning their hyperparameters. The models used include:

1.Logistic Regression:

Logistic regression is a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for given set of features(or inputs), X.

Logistic regression models the data using the sigmoid function.

2.Ensemble Machine Learning Algorithm:(Random Forest Classifier):

Random Forest is another ensemble machine learning algorithm that follows the bagging technique. The base estimators in random forest are decision trees. Unlike bagging meta estimator, random forest randomly selects a set of features which are used to decide the best split at each node of the decision tree.

Random subsets are created from the original dataset (bootstrapping).

At each node in the decision tree, only a random set of features are considered to decide the best split.

A decision tree model is fitted on each of the subsets.

The final prediction is calculated by averaging the predictions from all decision trees.

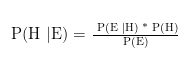
Note: The decision trees in random forest can be built on a subset of data and features. Particularly, the sklearn model of random forest uses all features for decision tree and a subset of features are randomly selected for splitting at each node.

3.Decision Tree Classifier:

It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

4.Gaussian NB: Using the Bayes theorem the naive Bayes classifier works. The naive Bayes classifier assumes all the features are independent to each other. Even if the features depend on each other or upon the existence of the other features. Naive Bayes classifier considers all of these properties to independently calculate the probability .Naive Bayes classifier is a straightforward and powerful algorithm for the classification task. Even if we are working on a data set with millions of records with some attributes, it is suggested to try Naive Bayes approach.Naive Bayes classifier gives great results when we use it for textual data analysis.

Below is the formula for calculating the conditional probability.



It works on conditional probability. Conditional probability is the probability that something will happen, given that something else has already occurred. Using the conditional probability, we can calculate the probability of an event using its prior knowledge.

5.AdaBoost Classifier:The general idea behind boosting methods is to train predictors sequentially each trying to correct its predecessor.The two most commonly used boosting algorithms are AdaBoost and Gradient Boosting.

AdaBoost is similar to Random Forest in that they both tally up the predictions made by each decision trees within the forest to decide on the final classification. There are however, some subtle differences. For instance, in AdaBoost, the decision trees have a depth of 1 (i.e. 2 leaves). In addition, the predictions made by each decision tree have varying impact on the final prediction made by the model.

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

6.K neighbors Classifier:

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

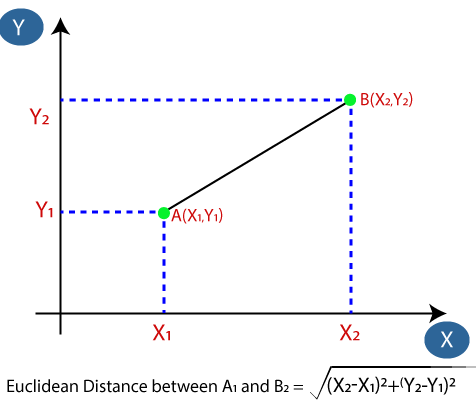
It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

This algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

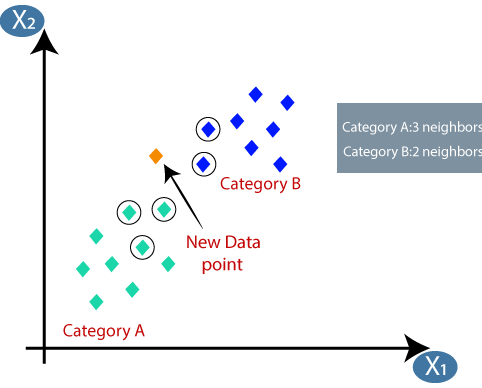
It is a non-parametric algorithm, which means it does not make any assumption on underlying data.

This algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:



By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:



As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

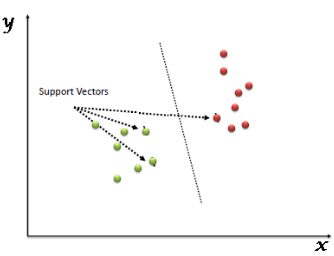
There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.

A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.

Large values for K are good, but it may find some difficulties.

7.Support Vector Classfier:

Support Vector Machine is a supervised machine learning algorithm which can be used for both classification or regression challenges.It is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well .Consider the below image:

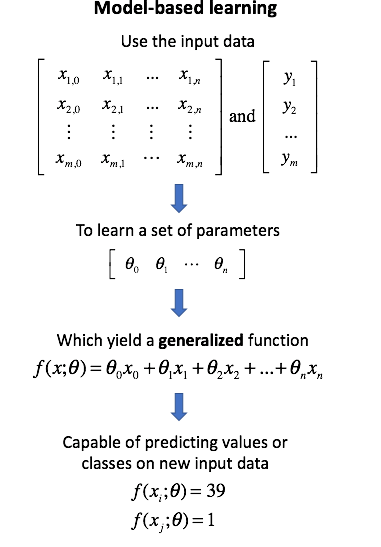


Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line).

8.Hyper Parameter Tuning:

Hyperparameter optimization or 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. By contrast, the values of other parameters (typically node weights) are learned.

Hyperparameters are not model parameters and they cannot be directly trained from the data. Model parameters are learned during training when we optimize a loss function using something like gradient descent.The process for learning parameter values is shown generally below.



Whereas the model parameters specify how to transform the input data into the desired output, the hyperparameters define how our model is actually structured.

9.Cross-Validation Score:

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data.This procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

The general procedure is as follows:

Shuffle the dataset randomly.Split the dataset into k groups

For each unique group:

Take the group as a hold out or test data set.

Take the remaining groups as a training data set.

Fit a model on the training set and evaluate it on the test set.

Retain the evaluation score and discard the model.

Summarize the skill of the model using the sample of model evaluation scores.

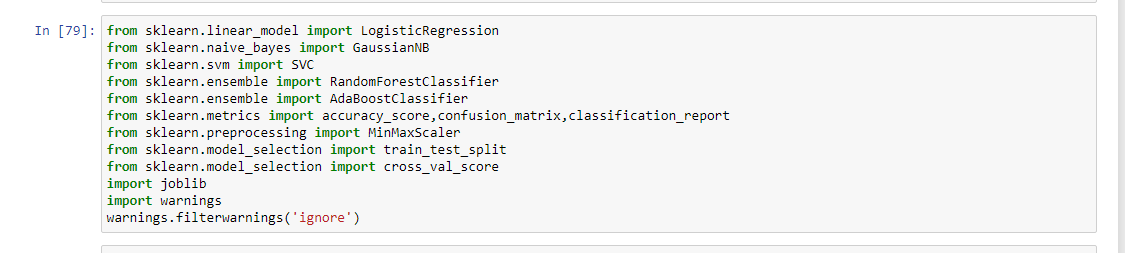
Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.

This approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining k − 1 folds.

We have seen all the metrics definition and Model Building Algorithms in short: Now we shall proceed with the model building for our dataset.

First we shall apply Min Max scaler to our dataset and then the model building approach.

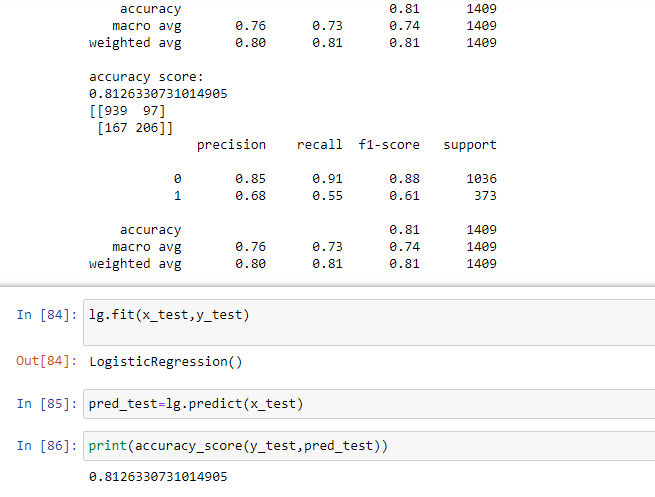
**Step 5: Building Machine Learning Models:**

****

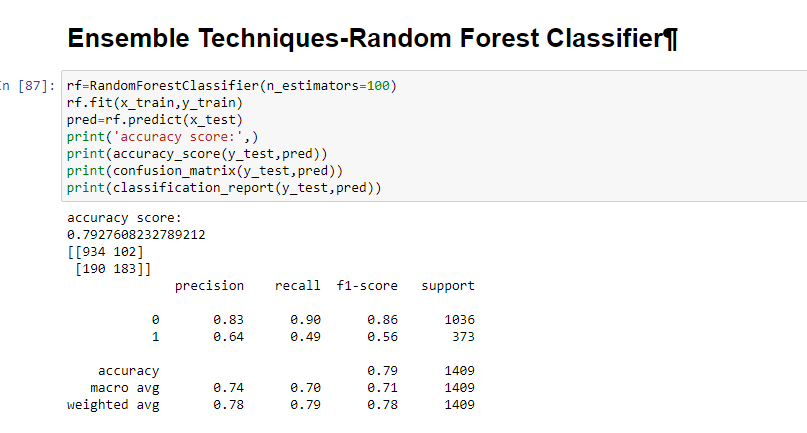
**Min-Max Scaling:**Values of numerical features are rescaled between a range of 0 and 1. Min-max scaler is the standard approach for scaling. For normally distributed features standard scaler could be used, which scales values around a mean of 0 and a standard deviation of 1. For simplicity we use a min-max scaler for all numerical features.

Tenure ,Total Charges,Monthly Charges.

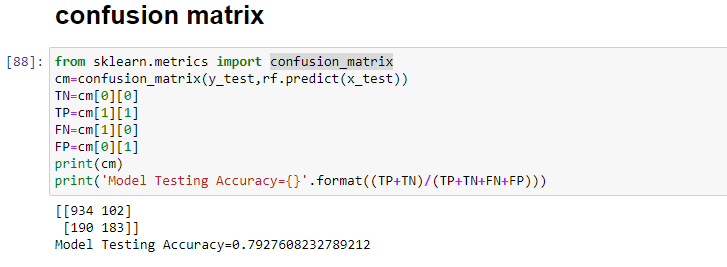
****

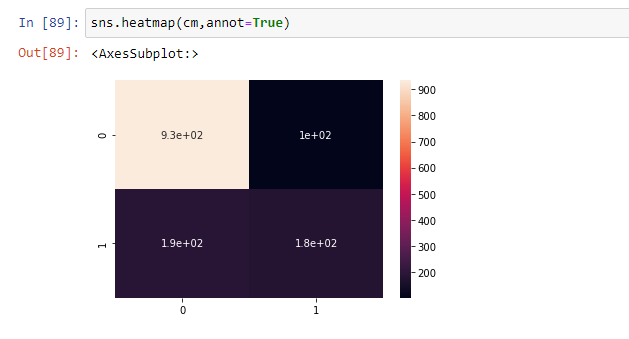
****

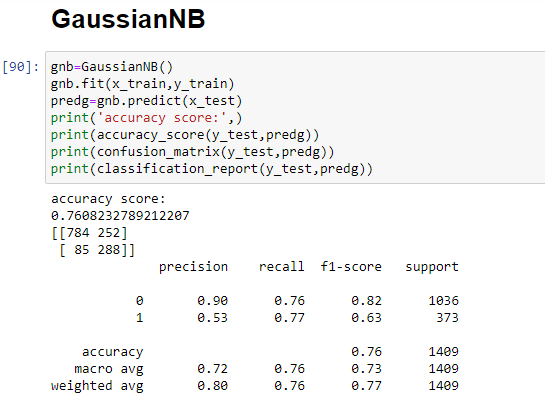
**The accuracy score of Logistic Regression is 0.81.**

****

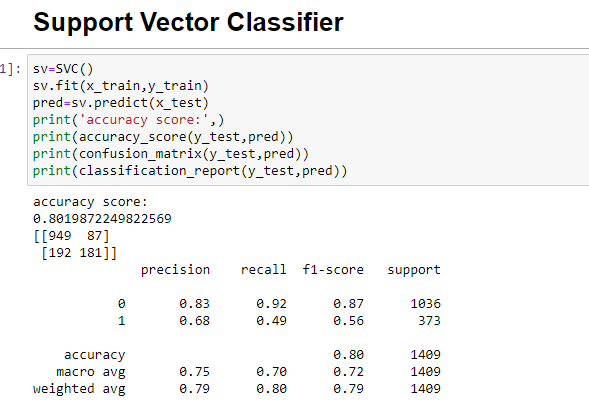
**The accuracy score of Random Forest Classifier is 0.79.**

****

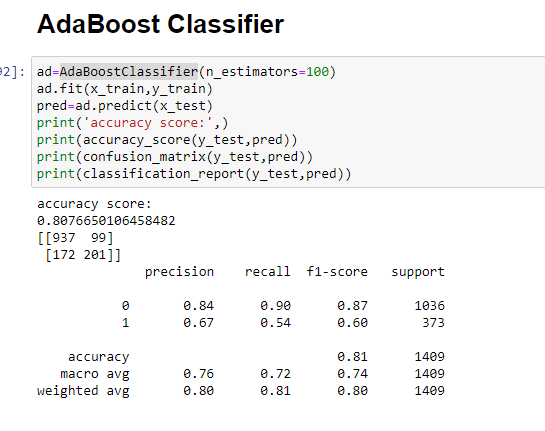
****

****

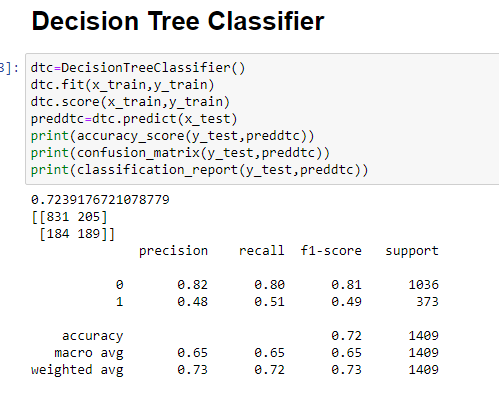
**The accuracy score of Gaussian NB is 0.76.**

****

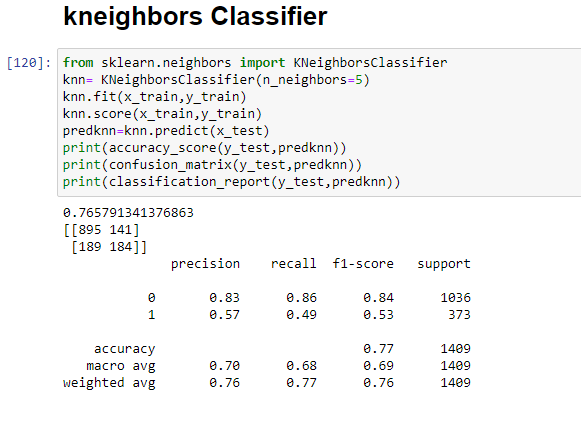
**The accuracy score of support vector classifier is 0.80.**

****

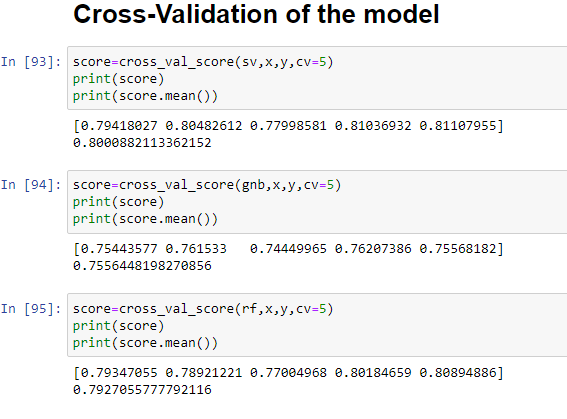
**The accuracy score of Ada Boost Classifer is 0.80.**

****

**The accuracy score of decision tree classifier is 0.72.**

****

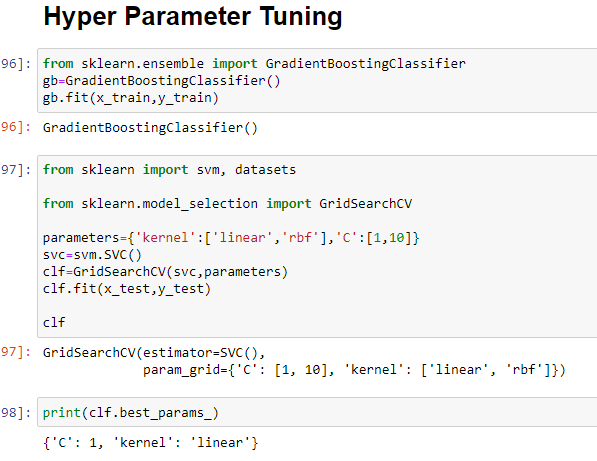
**The accuracy score of Kneighbors classifier is 0.76.**

****

**The cv score of support vector is 0.79,**

**The cv score of gaussianNB is 0.75,**

**The cv score of RandomForest Classifer is 0.79.**

****

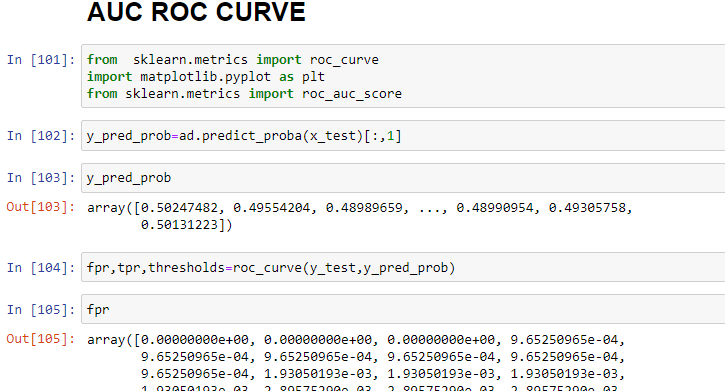
****

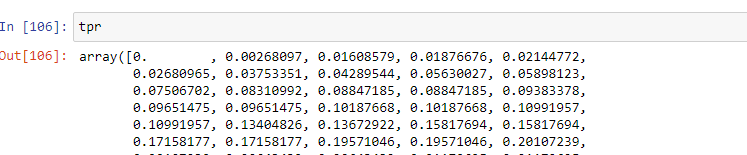
**The hyper parameter tuning has calculated**

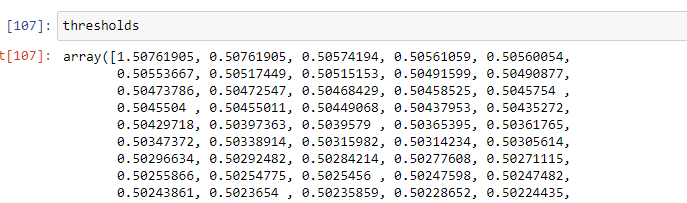
**0.83 in support vector machine.**

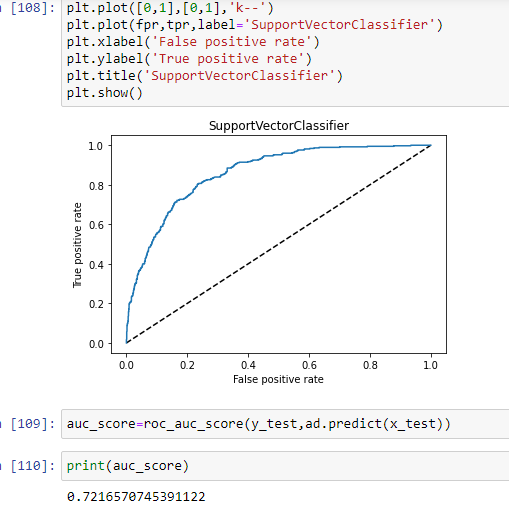
****

**The Grid search cv score for decision tree is 0.72.**

****

****

****

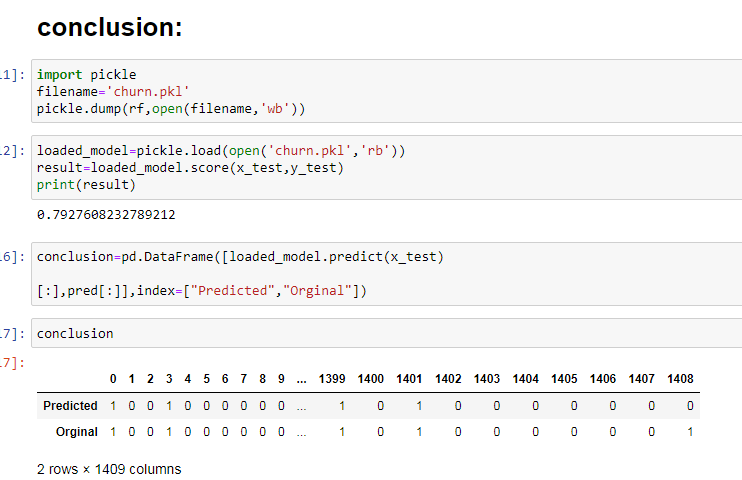
****

We have successfully drawn our auc roc curve.The auc roc score is 0.72.

Note:(The AUC value lies between 0.5 to 1 where 0.5 denotes a fair classifier and 1 denotes an excellent classifier.)

**7.Concluding Remarks:**

Finally we shall save our predictions in pickle file.We have learnt to build a complete machine learning project. In the process, we built custom transformers that can be used with sklearn’s Pipeline class. We also learned to fine-tune our model and save it for further use.

****

**Final Conclusion :**Our Best model is Support Vector Classifier which gives 0.83 score.

We have printed our prediction and saved the prediction in pickle file named as churn.pkl.

**This Article about Customer Churn Prediction Analysis is written by myself,**

**M.Kavitha,**

**Data Trained,**

**Batch number:1829.**

**Reference was taken on the Data Analysis model done by myself.**

**Please click the link below to see the model solved.**

<https://github.com/kavi4m/Data-Trained-Practice-Project/blob/main/Customer%20Churn%20analysis.ipynb>