**Customer Churn Analysis**

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 prioritize 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.

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

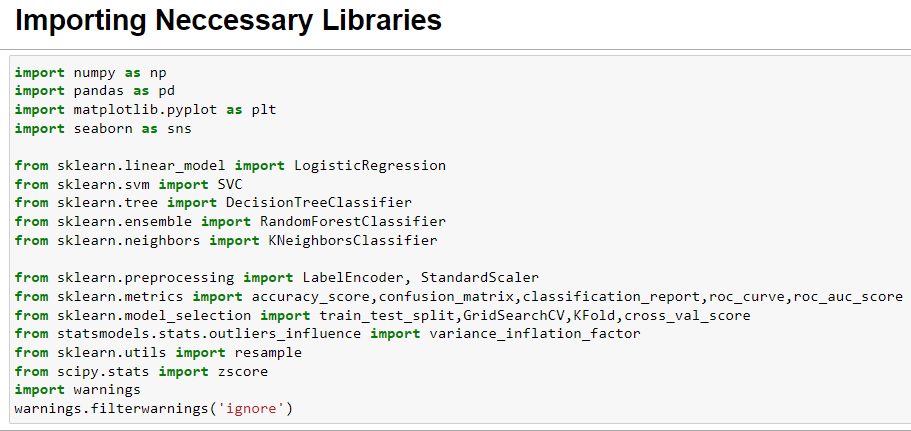
**OBJECTIVE**

**“The Goal is to predict whether Customer will loose or Churn or not, churn is the target col with binary classification.”**

So let’s begin,

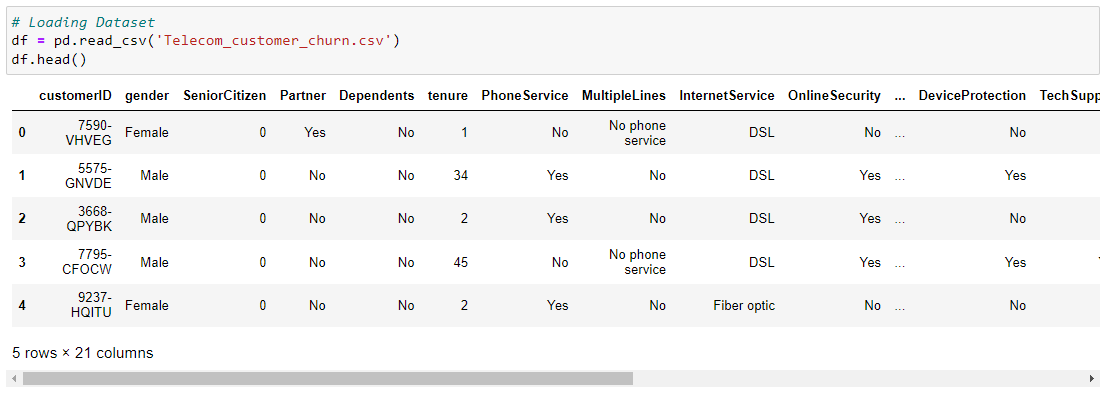
**Data Analysis**

1. Importing the necessary Libraries which will help us in doing the analysing and modelling dataset.



* We will be using Numpy, Pandas, Matplotlib & Seaborn (for plots) for basic analysis.
* Five Machine Learning algorithm used from sklearn library that is, Logistic Regression ,SVC , Decision tree classifier, Random Forest Classifier, and KNN Classifier.
* We have imported some metrics module from Sklearn metrics, as the metrics is common for all classification problems so imported common metrics all at once (Confusion Metrics, Accuracy Score, Classification Report).
* We have also imported some Libraries i.e, Label Encoder for encoding data to simplified format so that our machine can learn easily and Standard Scaler from Sklearn Pre-Processing.
* We have also imported resample from Sklearn Utils, so that we can balance or resample the target variable (if it is not balanced).
* From Sklearn model selection we have imported train test split, Grid Search CV, KFold, and cross val score.

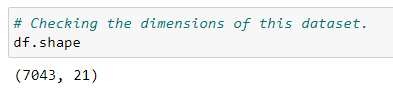
1. Loading Dataset and Creating data frame.



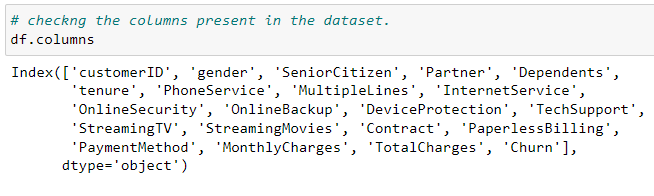
* We have successfully imported libraries and also imported the dataset.
* From the dataset we can see that there is various (21) columns including the target variable (Churn).
* Features information extracted from Dataset Target Column or Target Variable:
* Churn: This column has binary labels which says whether the customer will churn or not.
* Independent Variables:
* Customer ID: Customer Unique ID.
* Gender: Gender of a customer male or a female.
* Senior Citizen: Does the customer comes under senior citizen category.
* Partner: Any partner of the customer.
* Dependents: Any Dependent of the customer.
* Phone Service: Does the customer avail phone service or not.
* Multiple Lines: Does the customer has multiple lines.
* Internet Service: The type of customer internet service provider.
* Online Security: Does the customer has online security feature with internet provider.
* Online Backup: Does the customer has online backup benefits features with the provider.
* Device Protection: Does the customer avail the service for device protection.
* Tech Support: Does the customer avail the service for tech support.
* Streaming TV: Does the customer avail the service for streaming TV.
* Streaming Movies: Does the customer avail the service for streaming movies.
* Tenure: Duration of time the customer linked with service Provider Company.
* Contract: Terms & Condition of the contract of the customer.
* Paperless Billing: Does the customer preferred paperless billing.
* Payment Method: Preferred payment method used by the customer’s.
* Monthly Charges: This feature indicates monthly charge of the customer.
* Total Charges: This feature indicates the total amount charged to the customer.

**Exploratory Data Analysis (EDA)**

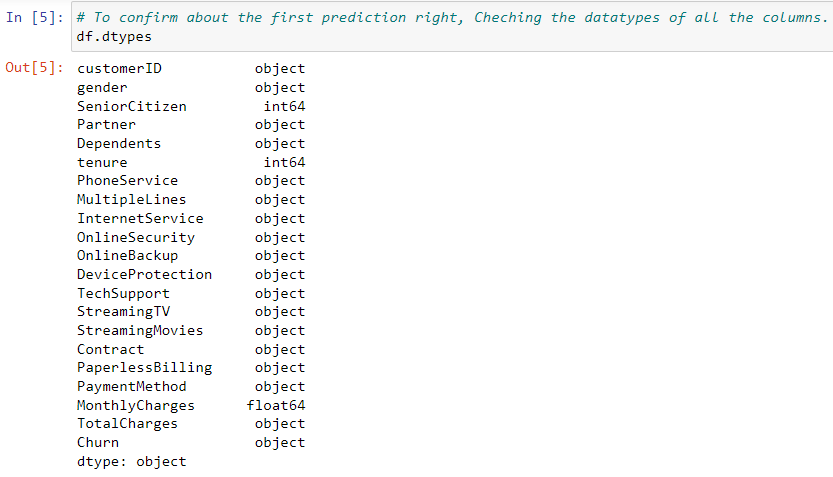
Firstly to understand about the dataset, we should know how much data is there in the dataset.



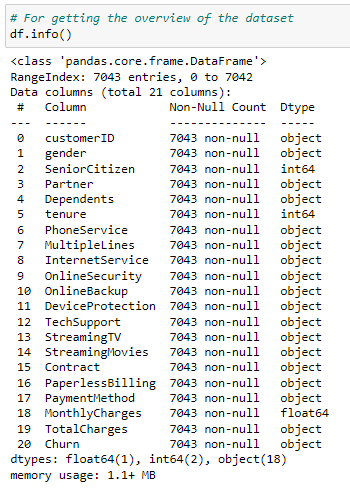
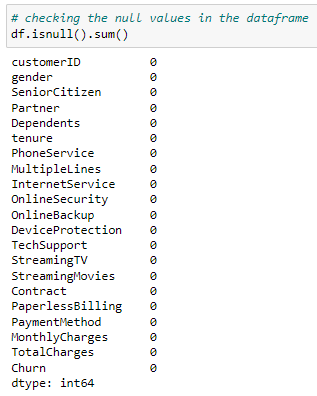
* Dataset consists of 7043 rows & 21 columns including the target variable.



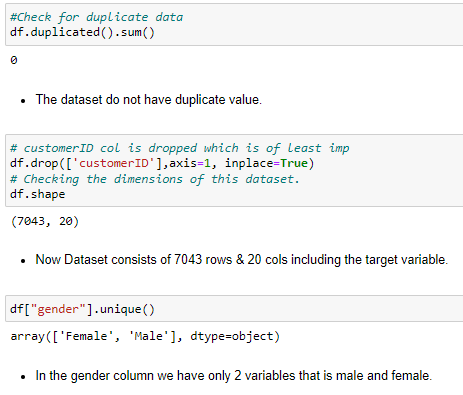
* No space or special characters seen in the title of the columns.



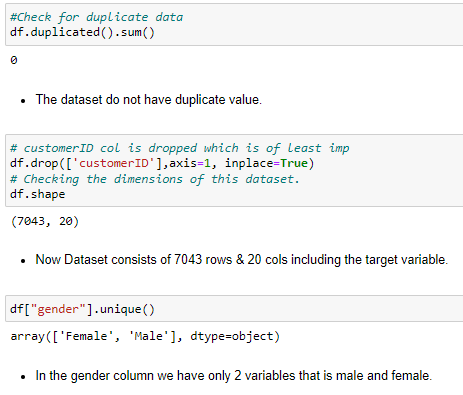
* 1 float, 2 int and 18 objects variables.
* int64 = 2 (SeniorCitizen, tenure)
* float64 = 1 (MonthlyCharges)
* object = 18
* dataset will be separated with numeric & objects cols

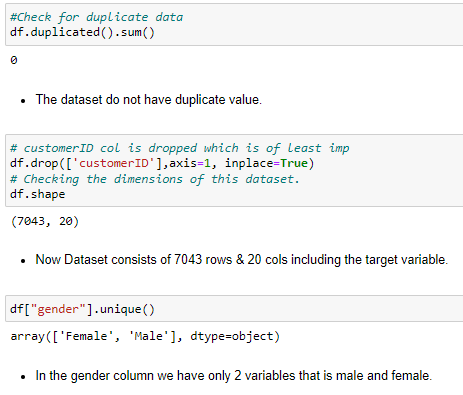
* The dataset do not have missing value still lets reconfirm.
* We can see that there are no null values in our dataset but still we have to check it closely for any missing, nan,? values to clean our dataset but before that lets check any duplicate values are present.



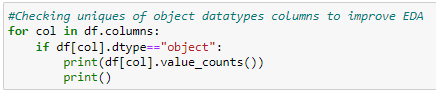
* The dataset do not have duplicate value.



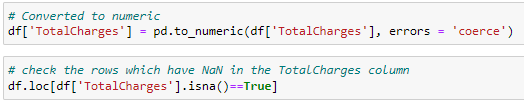
* Now Dataset consists of 7043 rows & 20 cols including the target variable.



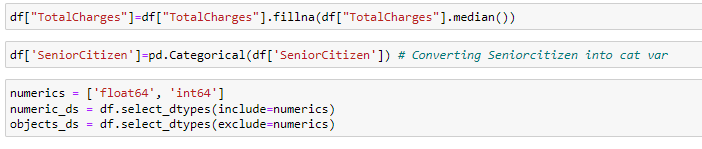
* In the gender column we have only 2 variables that is male and female.



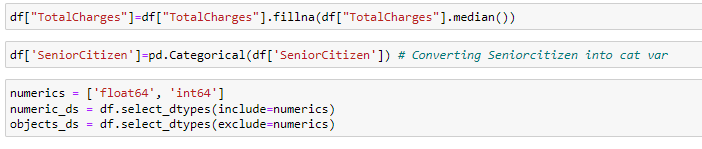
* Out of 7043 customers, 6361 have Phone Service whereas 682 do not have Phone Service.
* Out of 6361 customers, 3390 have Multiple Lines and 2971 do not have Multiple Lines.
* Out of 7043 customers, 3096 have Fiber optic , 2421 DSL that is total of 5517 have Internet Service.
* Out of 7043 customers, 1526 No Internet Service.
* Out of 5517 customers, 2019 have Online Security and 3498 dont.
* Out of 5517 customers, 2429 have Online Backup, and 3088 dont.
* Out of 5517 customers, 2422 have Device Protection, and 3095 dont.
* Out of 5517 customers, 2044 have Tech Support, and 3473 dont.
* Out of 5517 customers, 2707 have Streaming TV, and 2810 dont.
* Out of 5517 customers, 2732 have Streaming Movies, and 2785 dont.
* Out of 7043 customers, 3875 have Month-to-month Contract, 1695 have Two year Contract & 1473 have One year Contract.



* Some of the columns in Total Charges have NAN values.
* Let’s replace it with median.



* Converting the Senior Citizen into categorical variable for easy analysis.



**EDA Concluding Remarks:** In the given data, ‘churn’ feature is the Target feature or variable. The unique values of this feature are only 2 i.e Yes and No, which means it has only two classes. So, as there are only two unique values this is a ‘Classification Problem.’

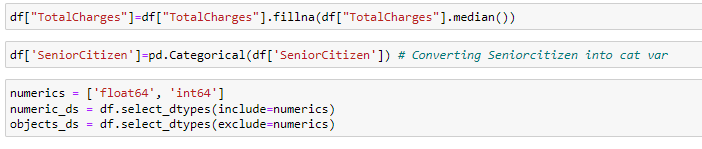
**Pre-processing Pipeline:**

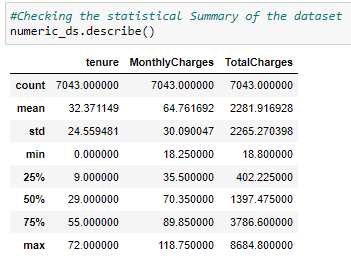
The data set has variables in both object type and numerical type (int and float)

Therefore we have to pre-process the data to move forward.

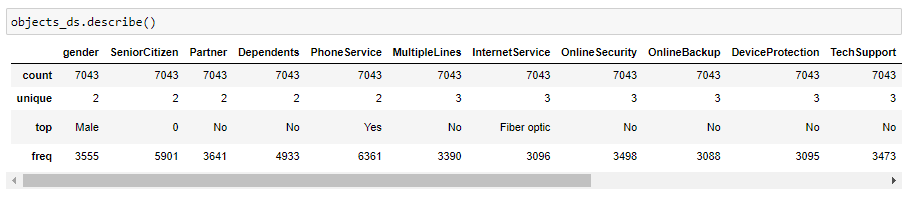
All the float type or int type variables should be converted into the same scale since the range of values of raw data varies widely, in some machine learning algorithms, objective functions do not work correctly without normalization.

* Therefore normalization is to be performed only on the numerical type (int and float type) variables.
* Diving the dataset into object and numeric to simplify the EDA.



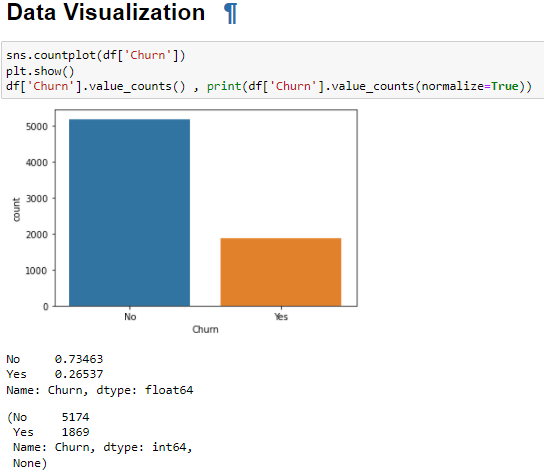


* Monthly charges shows big diff between max & 75th Percentile.
* Total charges also shows big diff bewteen max & 75th Percentile.
* Std is high for Total Charges & Mean is also > then 50th percentile
* As per above observations it seems data is skewed, spreaded.

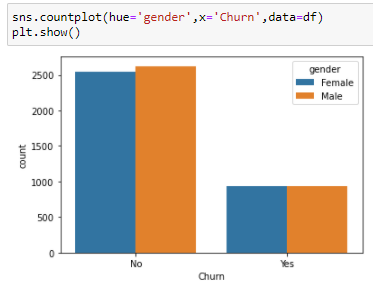


* From above it is clear that the top in gender is male with frequency 3555.
* 6361 customers have Phone Service and 3390 do not have Multiple Lines.
* Fiber optic is the most used Internet Service.
* Most of the customers that is 3875 have preferred Month-to-month Contract.
* Most of the customers that is 4171 have preferred Paperless Billing.

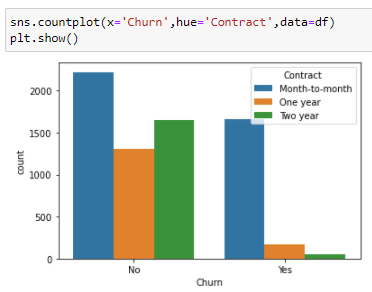
**Data Visualization**



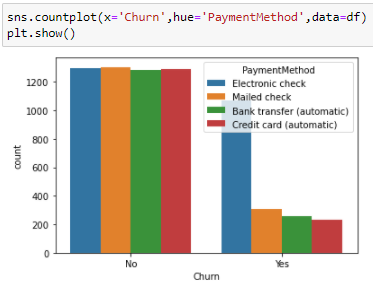
* We can see that 73.4% customers who are likely to churn compared 26.5% customers who are likely to churn.



* We can see that Female customers are not likely to churn as compared to Female customers who are likely to churn.
* We can see that Male customers are not likely to churn as compared to Male customers who are likely to churn.



* We can see that customers who are on Monthly contract are more and more likely to churn then who are on 1 year and 2 year contract

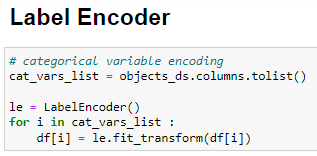


* We can see that customers whose payment method is Electronic Cheque are more likely to churn as compared to other mode of Payment

After the Normalising is done all the numerical data fall under the same range or scale.

We can see that majority of variables are object-type. These variables contain string data that cannot be passed into the machine learning model as it won’t be able to recognize string data type. It only recognizes numerical data.

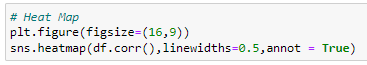
Therefore we need to convert the string data into numerical data. This can be done by manually encoding or by using an encoder such Label Encoder, one-hot encoder etc. For example: The target variable churn consists of only two unique values, Yes & No. after encoding this will get converted to 0 and 1. Similarly, if there are three unique values then it will be converted to 0,1, and 2.



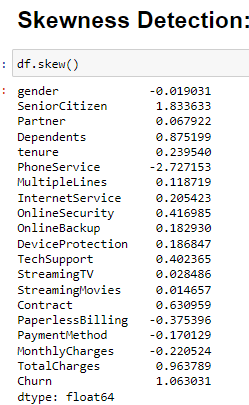
* Converting the values into simple codes for simpler learning.

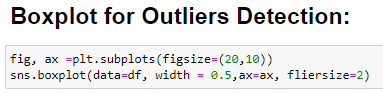
**Correlation between ‘Churn' and 'Independent features'**

* Now we can check the correlation between all the variables. (Note: correlation of all independent variables can be only done after encoding as correlation does not consider string values)



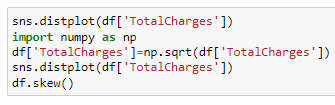
* Churn doesnt show corr with almost any variables.
* Tenure has corr with TotalCharges & contract





**Removing Skewness** :

The only numerical column totalcharges shows right skewed and positive skewed data which is removed by sqrt method after which skewness of the data is removed and comes under within the normal range +0.5 till -0.5.



**Class Imbalance:**

Imbalance Dataset: Basically classes or Labels of Target variables is not in proper ratio due to which model will get trained to predict the larger class as output so we have to balance ratio of class, Because ratio is not balanced model may predict low class cases as high class, So we need to balance the ratio of class so that model will predict low value class properly.

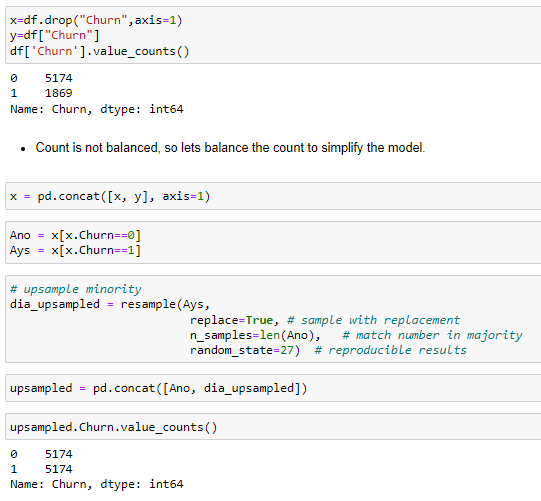
Imbalance Dataset is related with Classification problem having binary class, linear regression there is no issue with imbalance Dataset:

There are two methods to balance the ration of Target Variables:

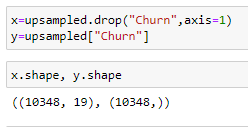
1. Oversampling/Upsampling : In upsampling Minor class is balanced with major class to get ratio balanced with Major class.

2. Undersampling/Downsampling : In Downsampling Major Class is Balanced with Minor class to get the ratio balanced with Minor Class.

In this example target variable churn has class value 1 as low value class and class value 0 as high value class so the up sampling method is used to balance the low value class to high value as we can see after up sampling both the class has same value count.



Dividing Dataset into Independent Variables(x) & Target Variable(y)



**Building Machine Learning Models:**

We have to now split the data into independent and target variables.

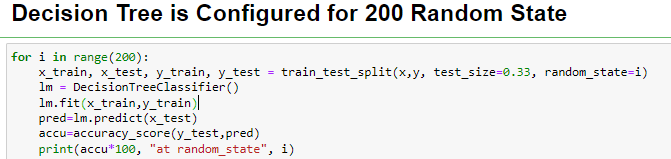


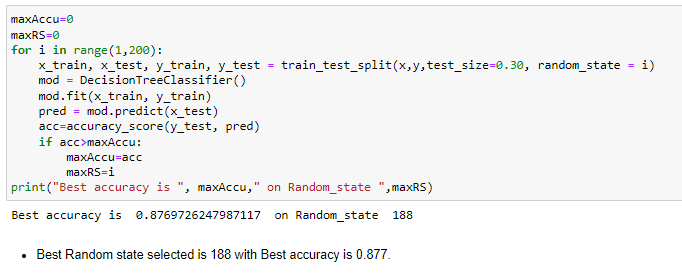
Here the target variable is ‘Churn’ and the rest of them are independent variables.

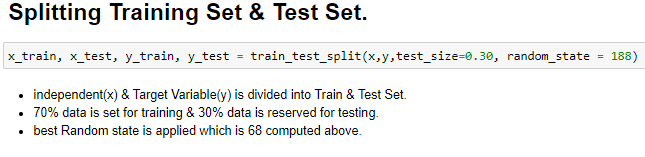
We have to now split the independent and target variables into training and testing datasets as shown below.

Applying Standard scaler to independent variables





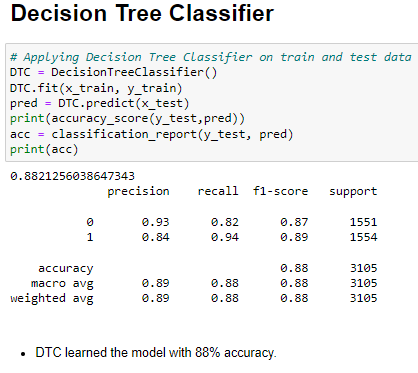
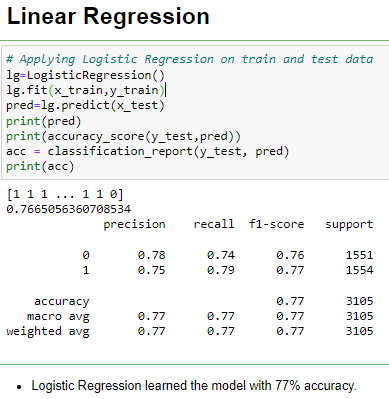


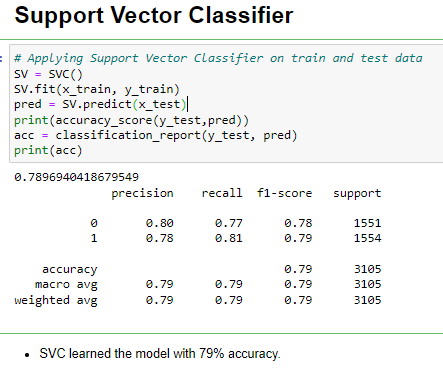
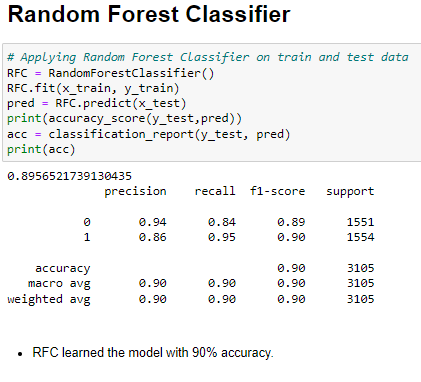


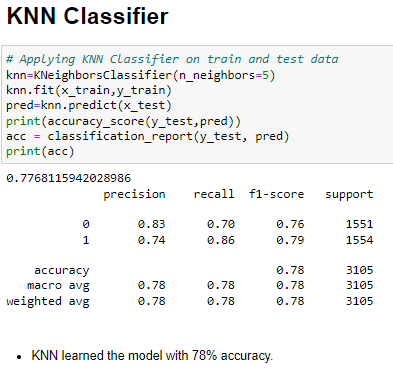
We will use a machine-learning algorithm to learn from the training set and use the model to predict the testing set and compare it with the predicted data with the target testing set to know how close the values. If the error between the predicted and target testing data is less that means the accuracy of the model is high and we can use this model to predict the result of similar datasets.

In this, we have used 5 Machine learning Algorithms

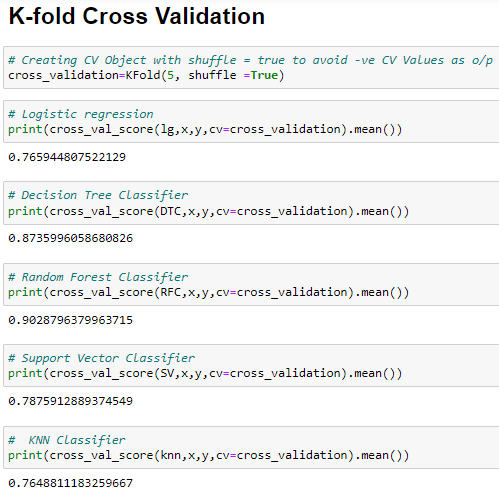
* Logistic Regression
* Support vector classifier
* Random Forest Classifier
* K Neighbors Classifier
* Decision Tree Classifier



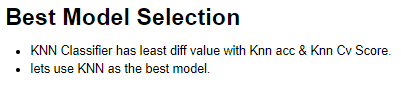




Further we can see Random forest classifier gives highest accuracy score among rest other machine learning algorithm, but the best model will be selected the one which has least difference value between Model accuracy and K-fold CV Score.



Cross Validation is a technique to prevent the model to get overfitted. We say model is overfitting when the model learned the training set well where the model gets high accuracy on the training set but when the same model is applied on the new set of data it is most likely to give bad accuracy, because it has never seen the data before and thus it fails to generalize the model well.



**KNN Hyperparameters & GridSearchCV :**

Hyper parameter tuning is basically we are fine tuning our model to get more accuracy, with Hyperparametr tuning we are helping model to learn better relationship of underlying pattern in train & test split by passing exclusive Parameters through GridsearchCV or Randomized SearchCV.

GridsearchCV is a method used to tune our hyper parameters we can pass different values of hyper parameters as parameters for grid search, it does a exhaustive generation of combination of different parameters passed. Using cross validation score. Grid Search returns the combination of hyper parameters for which the model is performing the best**.**

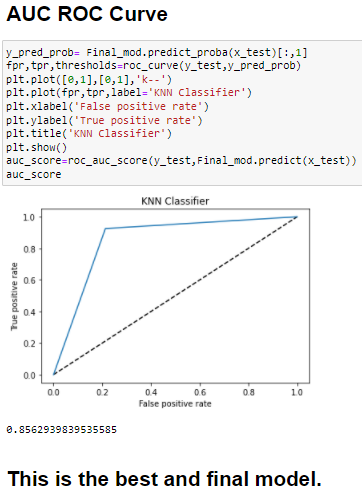
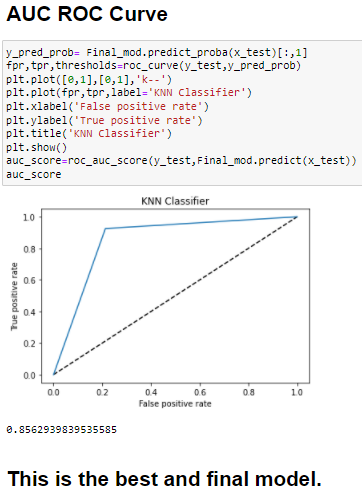


The following Knn Hyperparameters is used in Gridsearch CV : n\_neighbors, weights, algorithm, leaf size and metric and the best param value is applied on final KNN model where we got 85.63% as accuracy of the model.

**AUC ROC Curve:**

AUC ROC Curve is the Graphical Representation of Confusion Matrix for different threshold values where ROC Curve tells us about how good the model can distinguish between two things and the model will performed better if the value is close to the value 1.

AUC ROC curve basically used for binary classification problems it can also be applied for multiclass problem where there will be plot for each class of multiclass.



In this case we got the AUC Score of 0.8562 and can see the ROC Curve Closes to the value 1 indicate the model will perform well.

**Concluding Remarks:**

Churn prediction is really very important data science research and applications in all most all the companies as it helps company to know the overall profits earned by the business and also helps companies to deal with attrition of employees or churn of a customers, further it helps companies to make more profits because retention of customer is generating recurring income to the business of the companies.



