Customer Churn Prediction Analysis

Today I am going to write about a complete end-to-end project for Customer Churn Analysis to understand the basic architecture that’s required in the real world for creating a Data Science project.

**Customer churn** refers to the loss of customers or subscribers for any reason at all. Businesses measure and track churn as a percentage of lost customers compared to total number of customers over a given time period. This metric is usually tracked monthly and reported at the end of the month. It's important to note that churn rates vary by industry and knowing your market is key to reducing churn with more precision.

This article is containing the following sub-topics

1. Problem Definition

* How attrition impact the business
* How to achieve customer retention

1. Data Analysis

* Understanding the data

3. EDA

4. Pre-Processing Pipeline

5. Building Machine Learning Models

6. Concluding Remarks.

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.

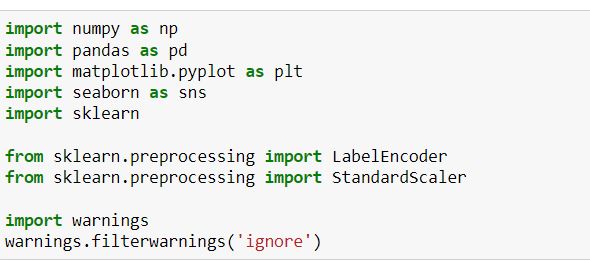
Customer retention can be achieved with good customer service and products. But the most effective way for a company to prevent attrition of customers is to truly know them. The vast volumes of data collected about customers can be used to build churn prediction models. Knowing who is most likely to 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.

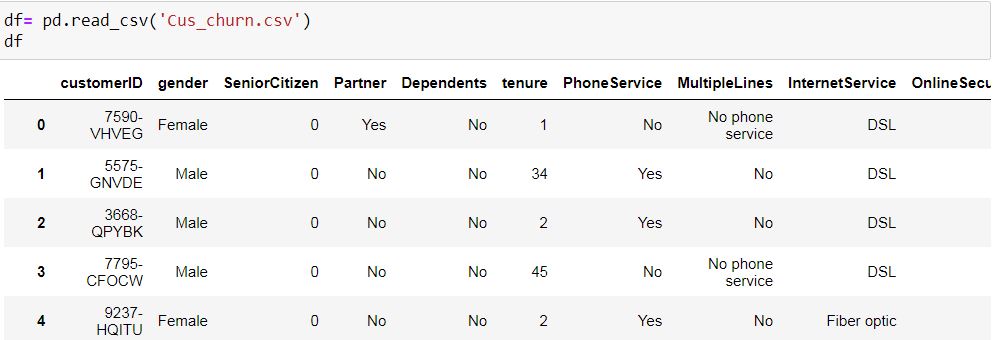
Customer attrition affecting companies is a major problem since high attrition is its cost to an organization. Therefore the major goal of this project is to identify the “Attrition” rate as a simple Yes or a No tag making this to be a classification problem!

**Importing libraries:-**

We need some libraries to be imported to work on a dataset.



We will import all the necessary libraries here that will be used in our project and obtain the rest as and when required. We must get the dataset before starting any further processing.



Dataset has been imported by using pandas read\_csv() function. We can see that it has mixed data types.

Capture.JPG

With this code we shall see the full column information on our Notebook directly.

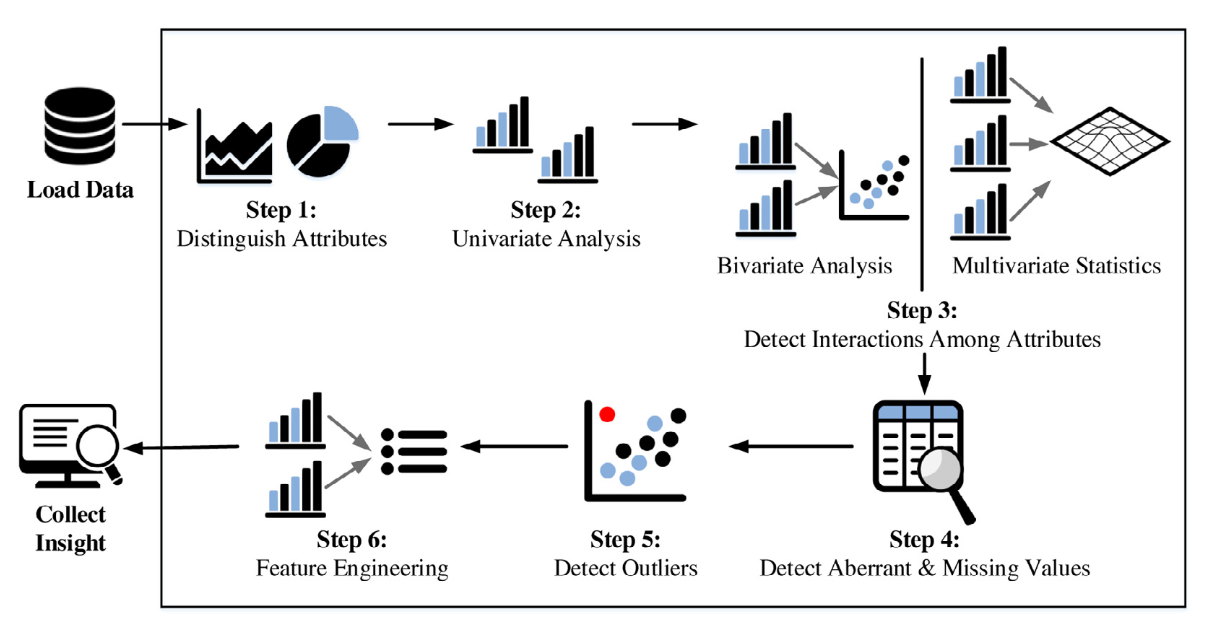
**EDA**

EDA refers to performing visualizations and identifying significant patterns, such as correlated features, missing data, and outliers. EDA’s are also essential for providing hypotheses for why these patterns occur.Exploratory data analysis is what data analysts do with large sets of data, looking for patterns and summarizing the dataset’s main characteristics beyond what they learn from modeling and hypothesis testing. EDA is a philosophy that allows data analysts to approach a database without assumptions. When a data analyst employs EDA, it’s like they’re asking the data to tell them what they don’t know.

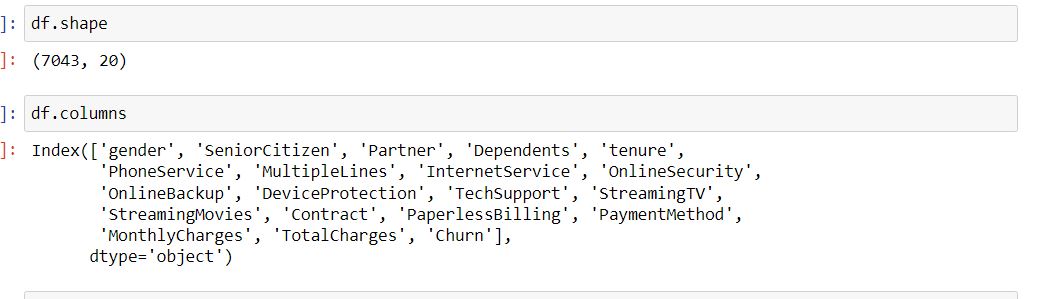
EDA is mainly used for these four goals:

* Exploring a single variable and looking at trends over time.
* Checking data for errors.
* Checking assumptions.
* Looking at relationships between variables.

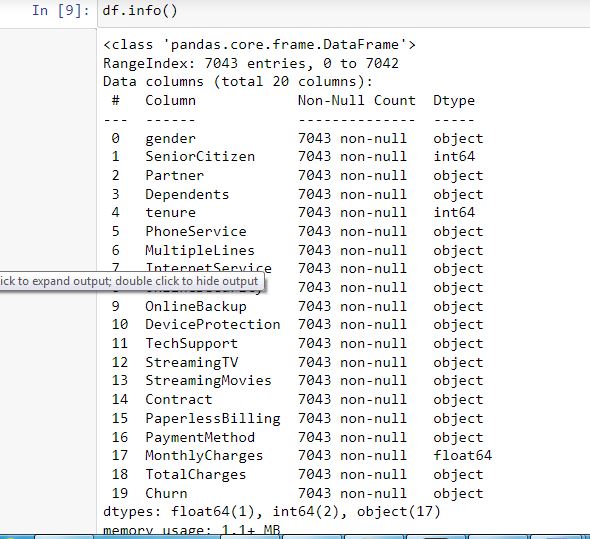
Steps Followed in EDA



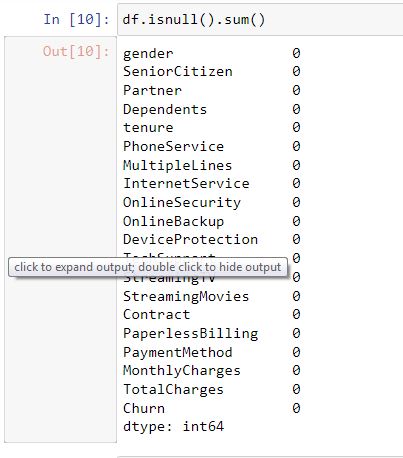
As a starting step, the shape and column names of the dataset are found using the below code.



Now we shall start the analysis with getting information about each columns as which data type it belongs to using the df.info() function.



Then, we shall check for missing values in the dataset using df.isnull() function.



The dataset has no null values in it, so its not necessary to do any filling or handling of the missing values. We shall proceed further with the analysis.

Now we shall separate the object datatype and numeric datatype values that allows for easier processing in further steps. The code to do that is a simple for loop usage.



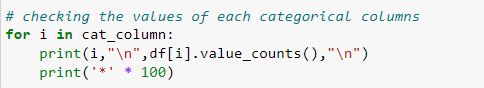
We did analyzing of the target variable with the following code.

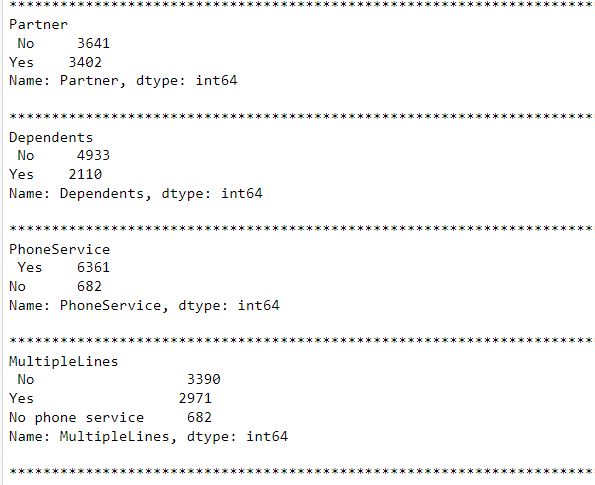


After storing the datatype column names in two separate lists, we will take a look at the overall unique values for all the columns and then the data numbers for only object datatype columns using the below codes.

This code provides us the output where we get the entire list of column names with unique data in the dataset. Also it provides the values for categorical object datatype columns.

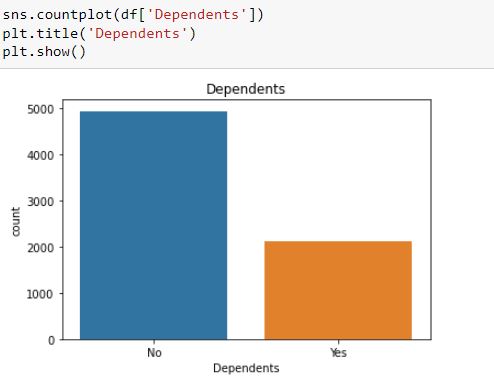
Capture.JPG

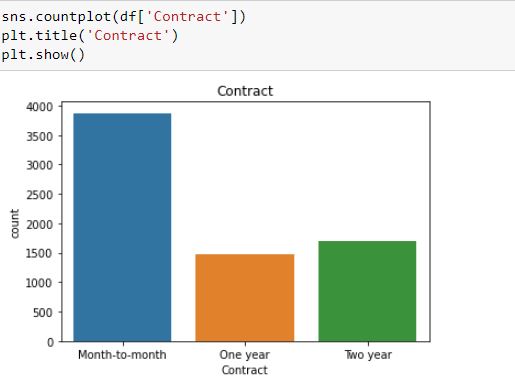




Now we shall take a visual look on how many rows or count of rows these values cover in our data set. Usage of various visualization techniques allows us to optimize and analyse the columns further. It gives an idea as to when and where data pre processing will be needed.

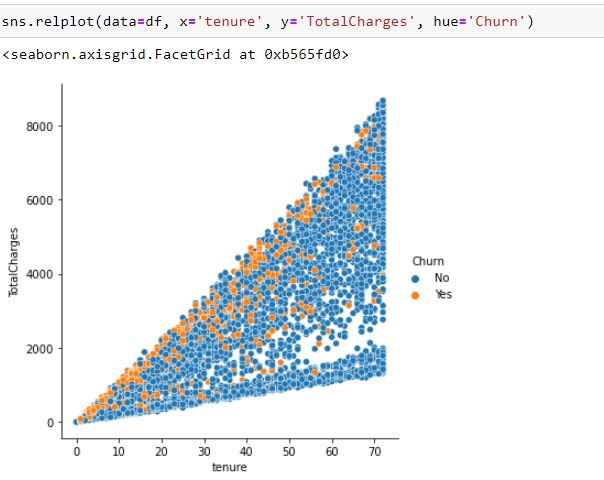
**Univariate analysis**





**Bivariate analysis**





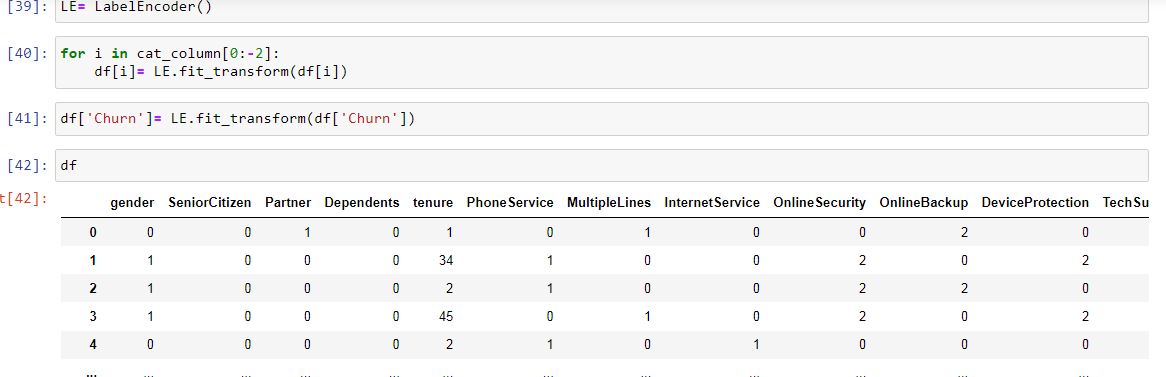
With the help of above codes and outputs we will be able to take a look at all the column values/counts, the univariate analysis gives us the insight to each column, where as bivariate analysis shows us the relationship between two or more columns.

**Pre-Processing Pipeline**

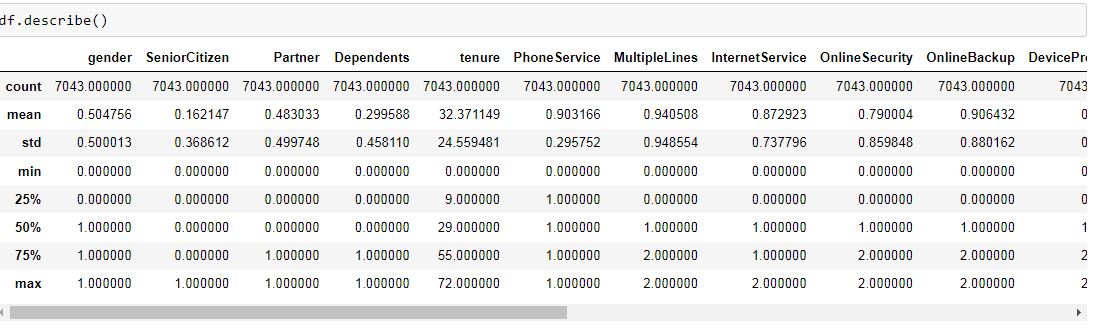
We have done some analysis and found interesting observations, missing value and relationship within columns and are treated accoringly.

Encode The Categorical Data

Machine learning algorithm can take only numerical input for learning, so it is important to covert categorical data into numerical data, it can be achieved by Label Encoding.

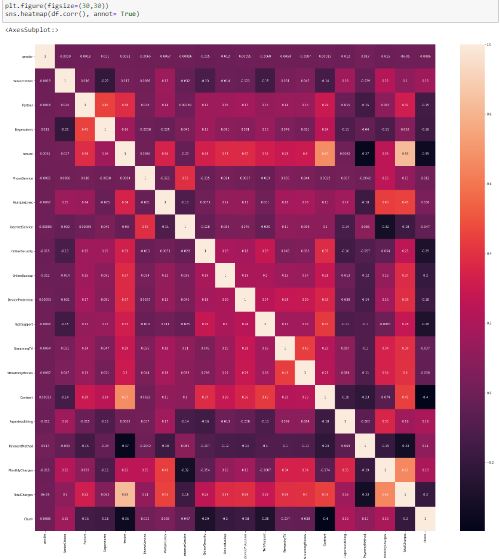


With the use of describe function, we shall get the statistical measurements between each columns.



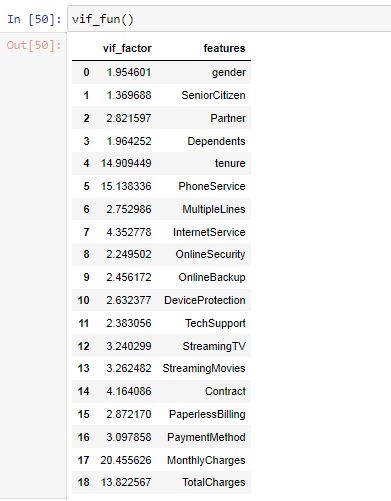
With the below code we shall get an idea about the correlation between each columns.

Capture.JPG

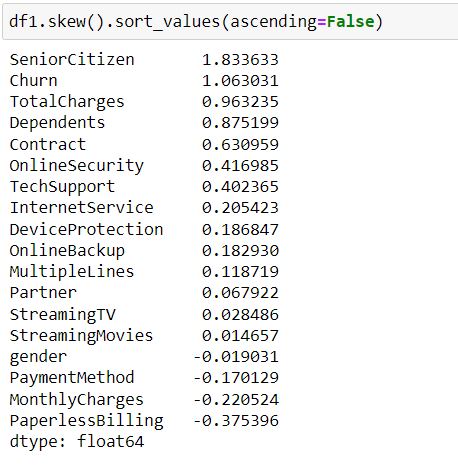


Now we should check for Multi Collinearity and remove columns with multicollinearity as it may affect the performance of the model.

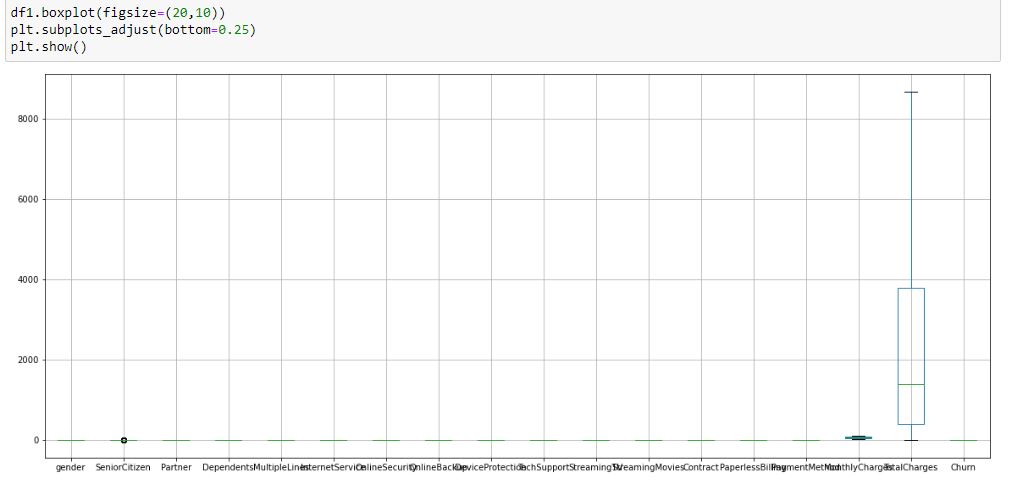




To avoid multi collinearitry we shall drop ‘tenure’, ‘phone service’. We are proceeding with skewness and outliers check.



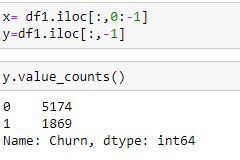
we see that there are columns that are above the acceptable range of +/-0.5 value of skewness. However most of those are categorical columns and we do not have to worry about outliers or skewness in catagorical data therefore we will ignore it.

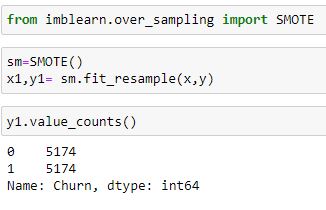


This dataset has no outliers present in it. So we shall proceed with Model Training.

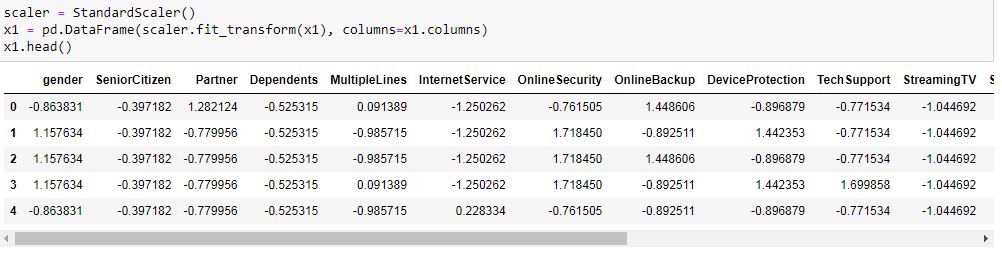
**Building Machine Learning Models**

The target variable is highly imbalanced which may affect lead to biasing and low performance. So we shall balance it with SMOTE.

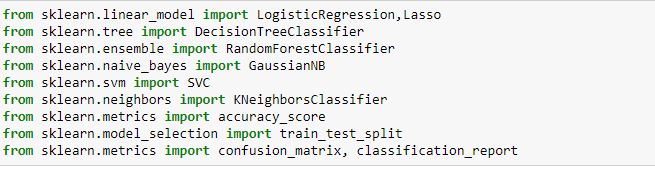




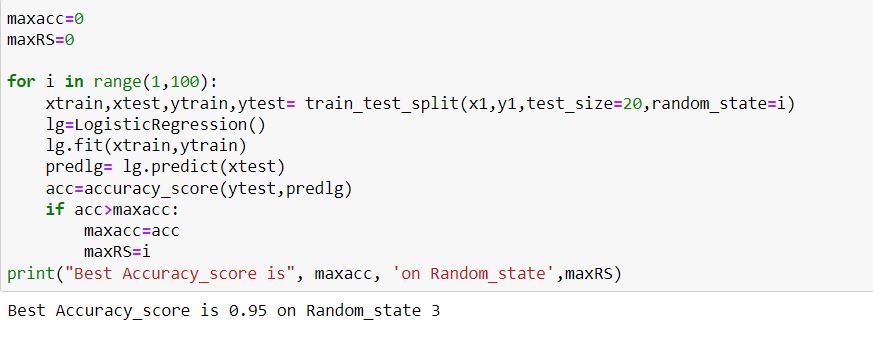
We shall scale the features to make the values to same range.



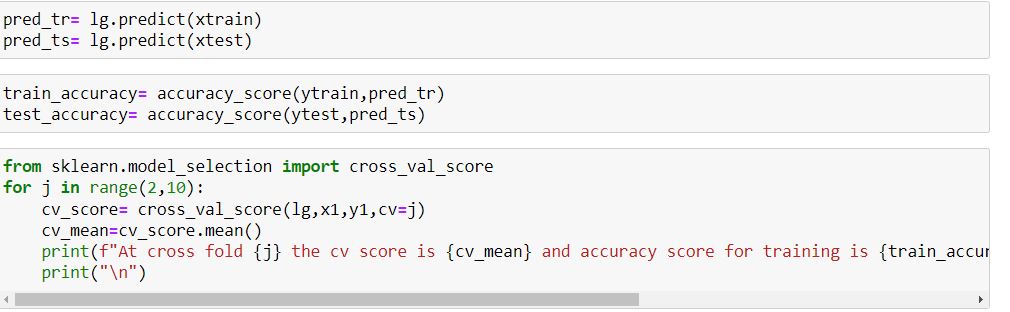
After balancing the dataset, we are moving ahead to train the model.



We shall use multiple machine learning algorithms to train model and pick the best model. Here we are taking Logistic regression as base model and running the loop for finding the best random state, let’s check the result of the model.



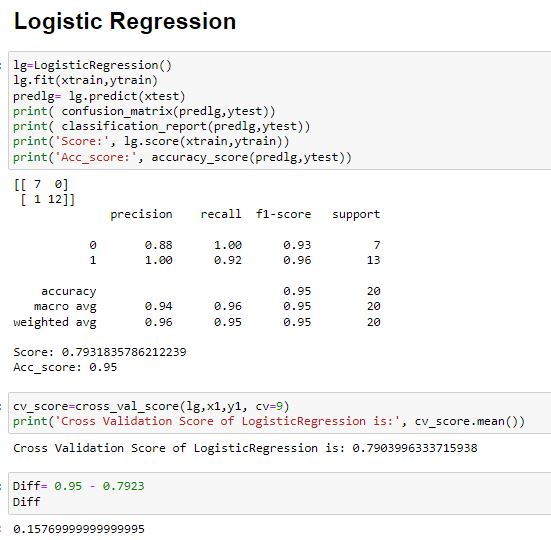
Next, we shall proceed to find the best CV score.

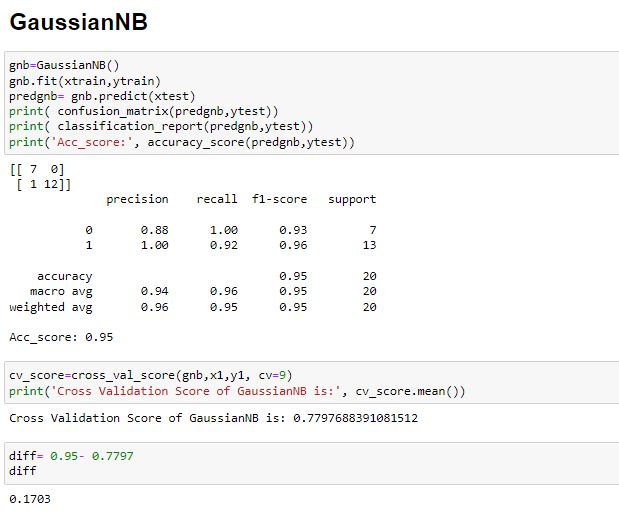


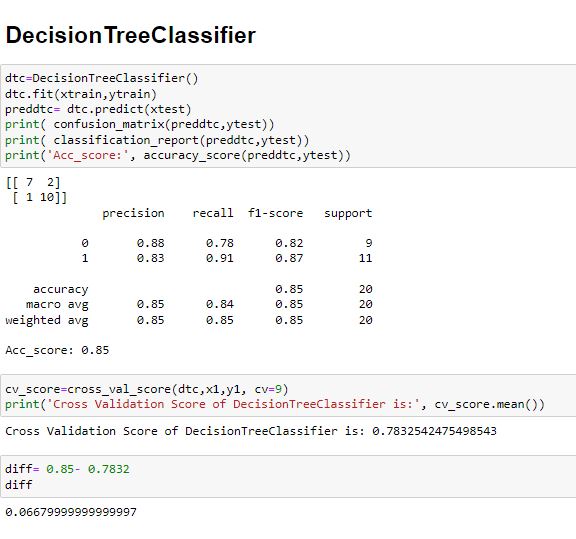
From the above code, best random state is found to be 3, best CV is CV=9. So we shall train the other models with this.

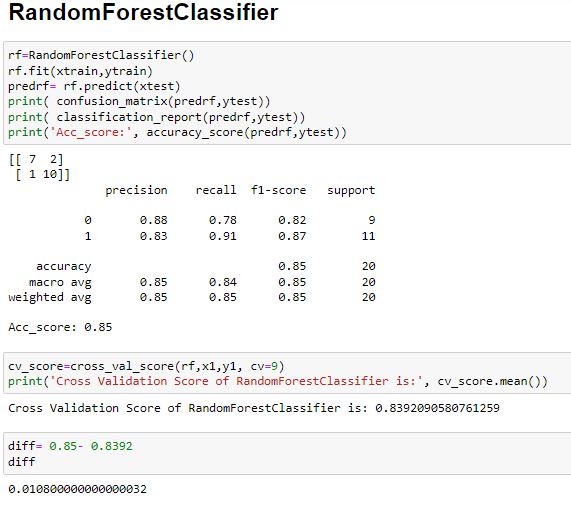
Capture.JPG

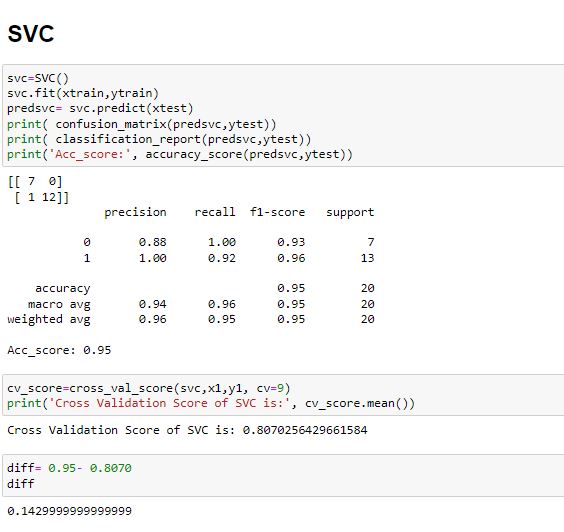
It is always recommended to build more than 5 machine learning models so that we can choose from the best performing model and then apply hyper parameter tuning to make it perform even better. Now we shall start with the basic model Logistic Regression.

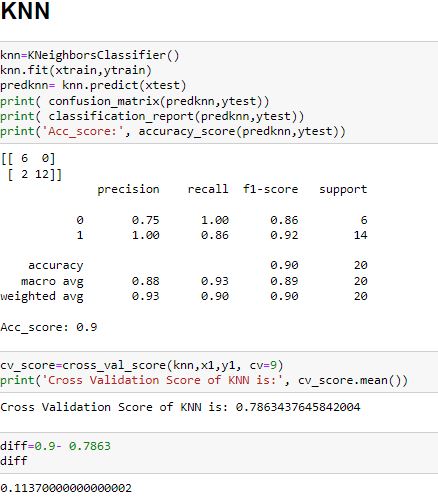








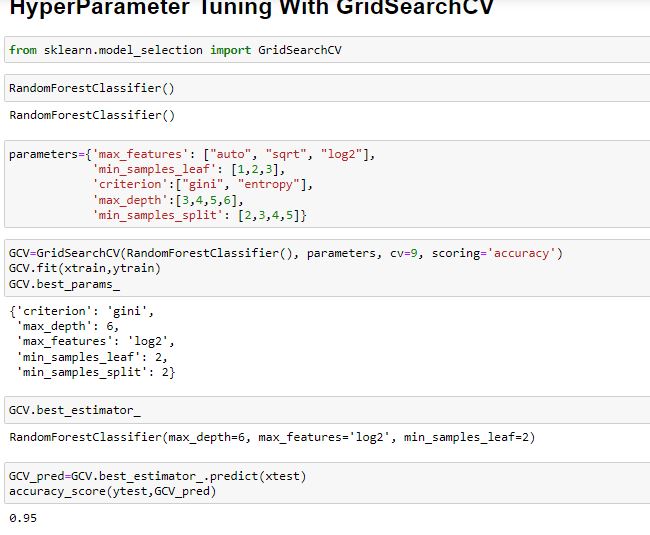




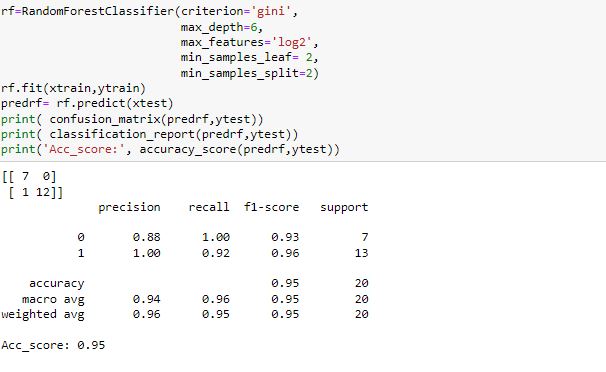
The below code shows the difference between accuracy score and CV score for each model.



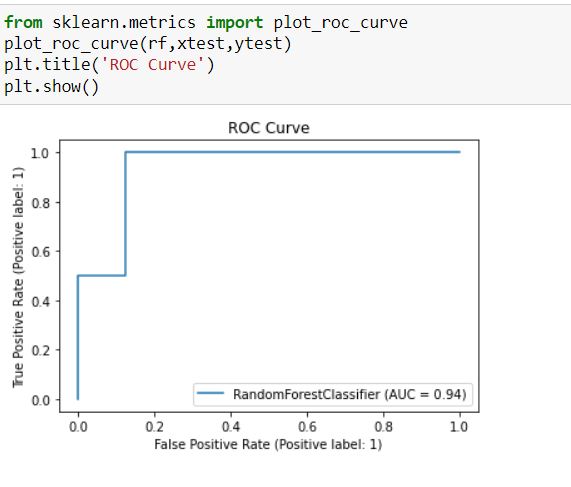
The difference between accuracy score and Cv score is less with RandomForestClassifier which reduces the overfitting of the model. So we shall improve the performance by tuning the Hyper Parameters with GridSearchCV.



After tuning the Hyper Parameter, Final model is run with the best parameters.



Now we shall create ROC curve for the final model- RandomForestClassifier.



After complete analyzing the data, training of the models and hyper Parameter tuning we have arrived at a satisfied result. So we shall save the model with pickle.



We shall compare the original results to the predicted results to get visually proved result of the models performance.



**Concluding Remarks**  
  
Dataset has 7043 rows and 21 feature including target variable, there is no outlier available in the numerical columns. I have applied label coding on categorical data to convert them into numerical dataset.

From correlation plot, we have found that an only contract column is highly negatively correlated. We have applied VIF to rule out multicollinearity. We have applied multiple model on the dataset and found that Random forest is working good with good accuracy score and CV score, have chosen it for saving the final model but before saving we have applied Gridsearch CV saved it.

We have seen that the Churn rate is high with people without Tech support and with the tenure period less than 20. Maintaining Good old customers are always beneficial to the organization than going for new customers from time to time.