**Customer Churn Analysis in Telecom Industry**

****

**Submitted by**

**Raj Sharma**

**Objective: -**

In this article, we are going to apply data analytics skills on the telecom customer churn dataset which contains customer level information and try to find insights about the key attributes which are causative for customer analysis, we try to find the hidden trends from customer behavior through visualization techniques, build several predictive classification models. This article concludes by providing successfully predict potential customer churn, Companies can use such ML pipelines to initiate retention strategies on those customers who are classified as likely targets of churn.

**Abstract: -**

With rapid development of telecommunication industry, the service providers are more inclined towards expansion of the customer base. Customer acquisition and retention has become a key concern for several industries and is particularly acute in fiercely competitive and growing business. Finding the Key factors which triggers the customer churn plays important role in early initiation of customer retention policies and cut back the churn. We will focus on analyzing the customer data, perform exploratory data analysis, to get insight about which variables are contributing to customer churn, implementing the machine learning algorithms to identify potential churn customers and label them based on usage patterns and visualize the results.

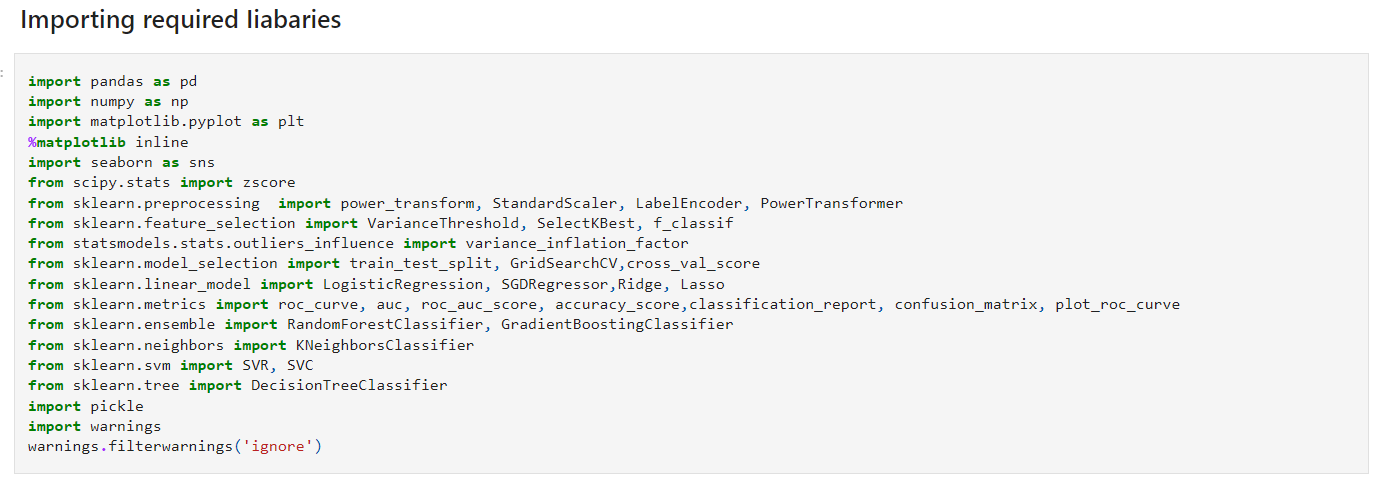
**Problem Statement: -**

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.

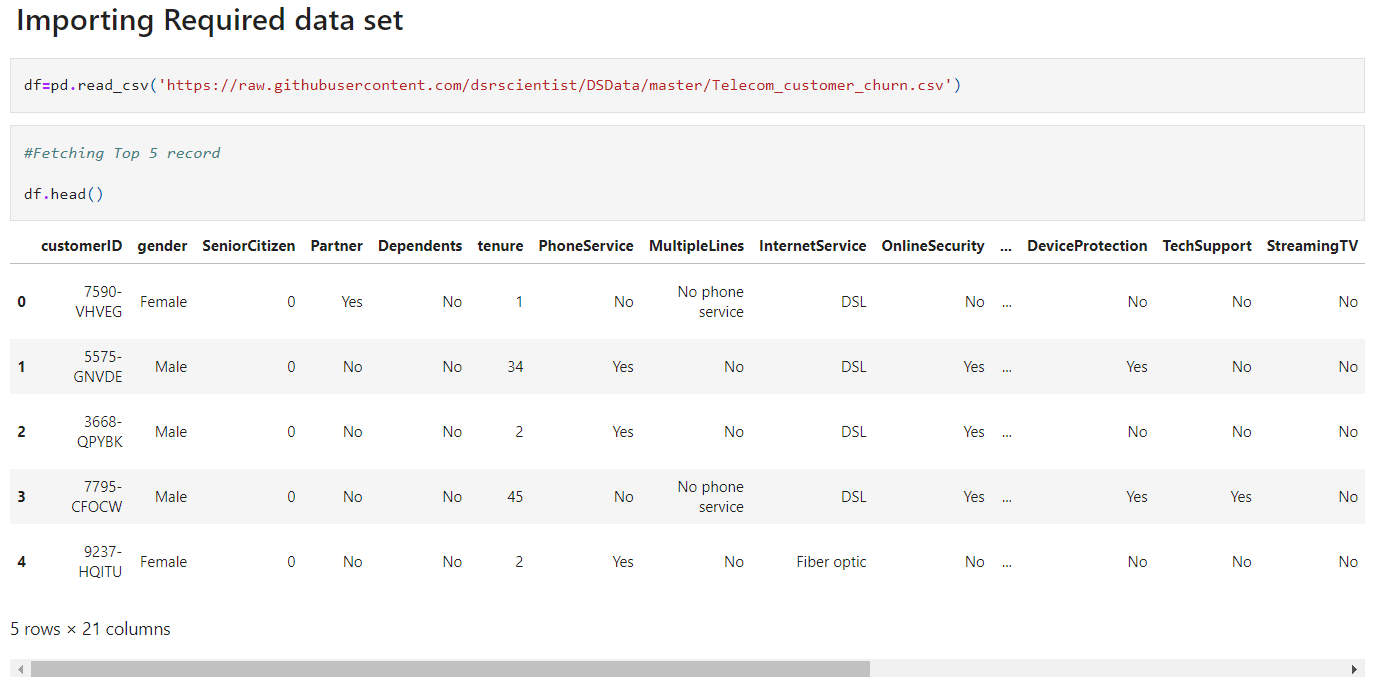
Customer retention can be achieved with good customer service and products. But the most effective way for a company to prevent attrition of customers is to truly know them. The vast volumes of data collected about customers can be used to build churn prediction models. Knowing who is most likely to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

**Data Analysis: -**

****

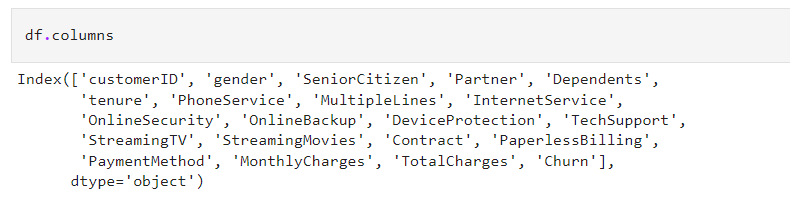
**Data Overview: -**

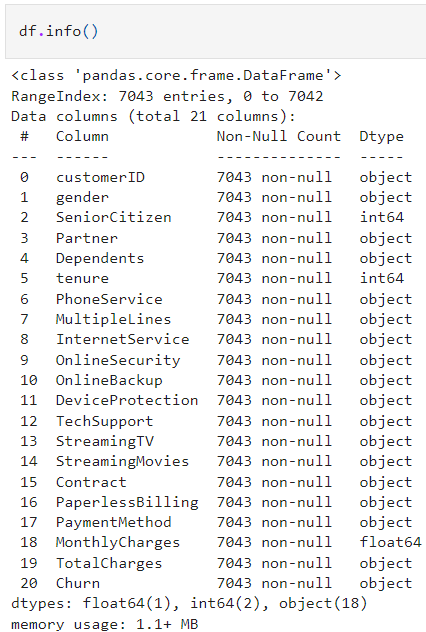
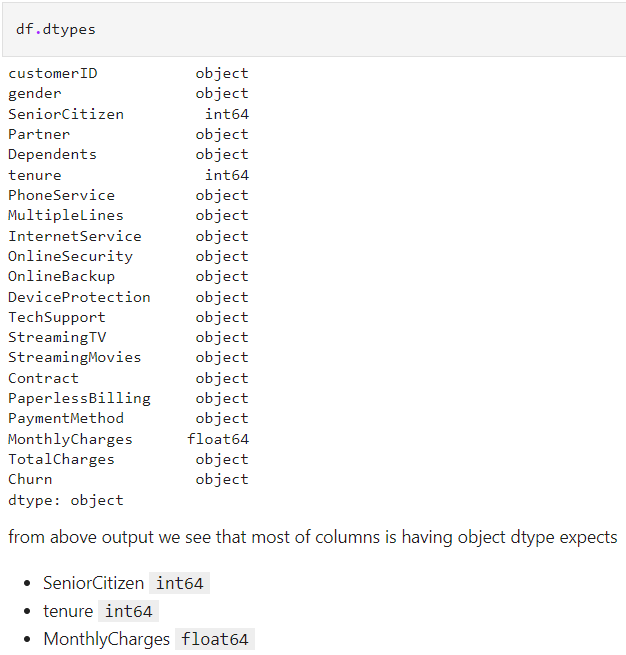
****

Here each column consists of details of customer information which is recorded by company during the service. There are 4 types of information

* Services used by customer.
* Customer Demographic information (basic details)
* Customer churn Details.
* Customer Account information.

Following are the attributes available from the datasets which contains all features and (Churn)target variable.

**

Observations: -

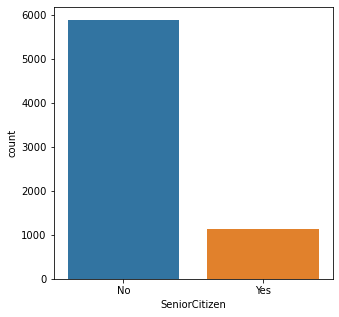
* The dataset consists of 7043 rows and 21 columns,
* The dataset contains 18 categorical and 3 numerical values represented in right table. Total charges should be numerical data but it is showing categorical data, let us check what is missing.
* There are no null values in the dataset.
* The descriptive statistical overview of the dataset is given in right table.
* The below table provides brief understanding of all the columns and types of data present

|  |  |  |
| --- | --- | --- |
| Column | Definition | Comments |
| Customer ID | Unique ID provided by company |  |
| gender | Gender of customer | Male, Female |
| Senior Citizen | If customer is Senior Citizen | 1= Yes, 0=No |
| Partner | If customer have partner | Yes, No |
| Dependents | If customer have any dependents | Yes, No |
| tenure | Since how many years customer is using service of company |  |
| Phone Service | If customer has phone service | Yes, No |
| Multiple Lines | If customer uses multiple line service | Yes, No, No phone service |
| Internet Service | Does customer have internet service | Fiber Optic, DSL, No |
| Online Security | Does customer use online security service | Yes, No, No internet service |
| Online Backup | Does customer use online backup service | Yes, No, No internet service |
| Device Protection | Does customer use device protection service | Yes, No, No internet service |
| Tech Support | Does customer use tech support of company | Yes, No, No internet service |
| Streaming TV | Does customer streams TV? | Yes, No, No internet service |
| Streaming Movies | Does customer Streams movies ? | Yes, No, No internet service |
| Contract | What type of contract does customer use? | Month-to-Month, One year, Two year |
| Paperless Billing | Does customer prefer paperless billing? | Yes, No |
| Payment Method | What is mode of payment customer opt for? | Payment Method |
| Monthly Charges | What is monthly charge of customer |  |
| Total Charges | Total charges since using the service |  |
| 'Churn' | Does customer leaves company or continues with the company service? | Yes, no |

**EDA: -**

Univariate Analysis:

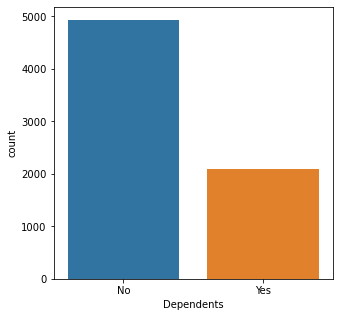
Gender Distribution and Senior Citizens:

* There are equal number of customers in our data set that male while the other half are female.
* There are 1142 Senior Citizens which comprise of only 16% and 5901 Non-Senior Citizens using the service.

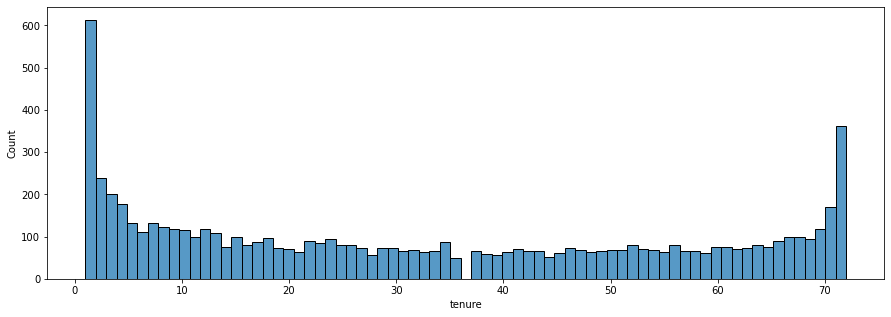
Partner and Dependents Distribution:

* There 3402, nearly 49-50% of customers having partners, and 70% of customers having dependents

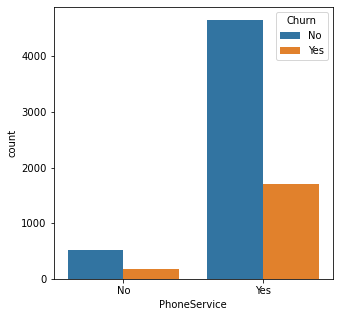
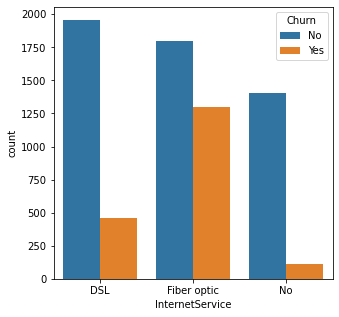
Tenure:

* The average number of people staying in the service is nearly 150, Most number of people churn is seen after 1-2 months gradually, we can see more number of people staying in service after 70months.



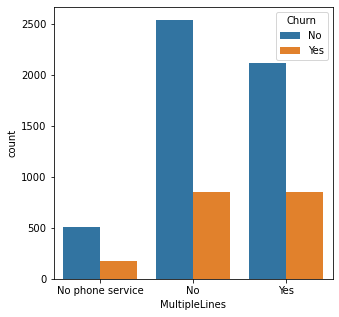
Phone Service and Internet Service:

* 21% of the customer data, do not have Phone Service, they are not contribution for any other service as well.
* 2. 79% of customer of overall has Internet Service, and 44% of them choose for Fiber Optics, and 35% choose for DSL Internet Service.

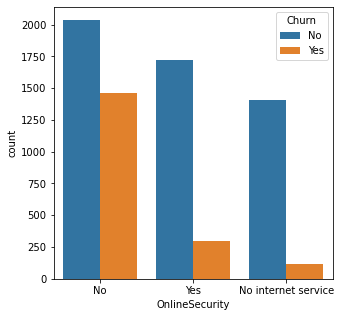
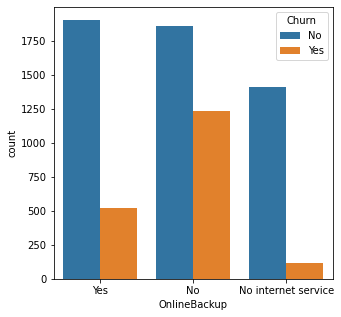
Multiple line Services:

* Only 40% of total customer has multiple lines, and 48% do not opt for multiple lines, 22% do not have phone service.



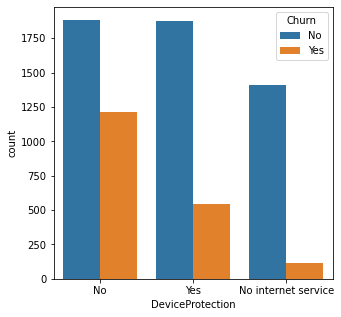
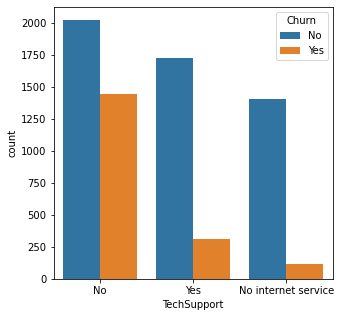
Online Security and Online Backup:

* Only 28% of total customer choose Online Security service, and 49% do not choose online Security service.
* And 34% of customer choose online backup service, 44% do not choose, and 22% do not have Internet Service.

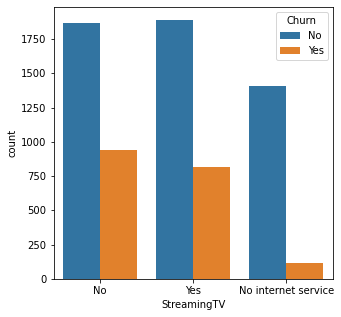
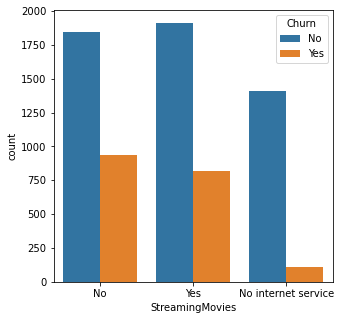
Device Protection and Tech Support:

* 29% of customer Choose Tech Support service from the company,
* 34% customer choose Device protection service.

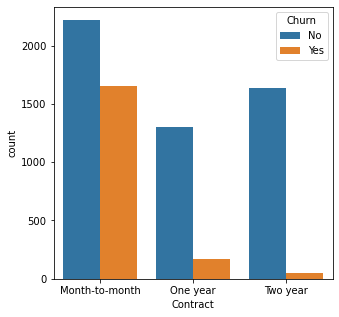
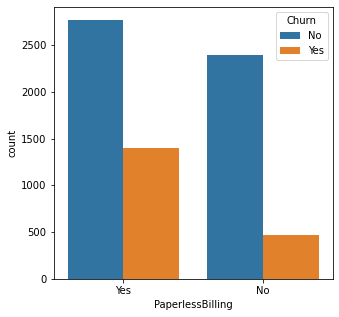
Streaming TV class and Streaming Movies class:

* Nearly 50% of customers using Internet stream TV.
* 48% of customers using internet will go for streaming Movies
* 21% of customers do not have Internet services.

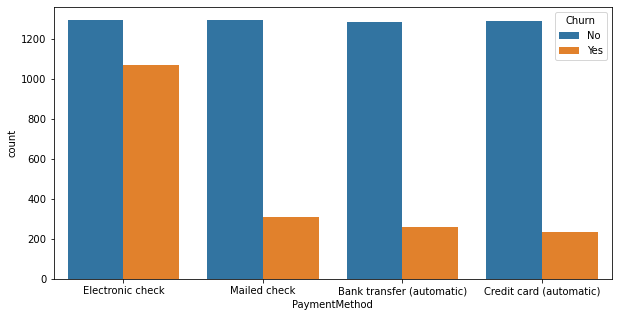
Contract class and Paperless Billing:

* 54% of customer choose Month-to-month contract, and 22% choose for 1 year contract,24% choose 2-year contract.
* 42% customer go for paper billing and 58% go for paperless billing.

Payment Methods:

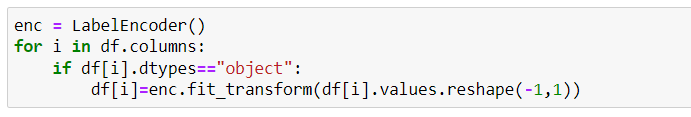
* Nearly 32% of payments have been done through electronic check, around 22-25% by other methods(Mailed check, Bank transfer, Credit Card)



**Data Preprocessing**

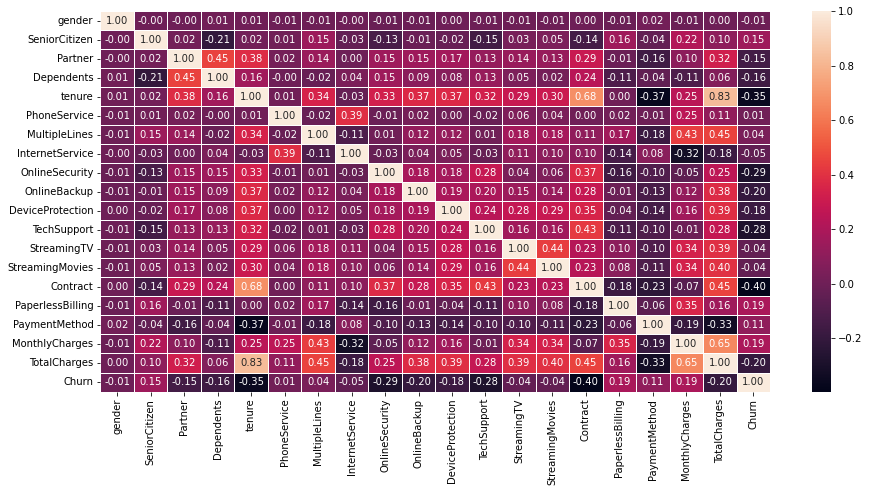
Data preprocessing is the most important phase in prediction models as the data consists of ambiguities, errors, redundancy which needs to be cleaned beforehand.

The following represents the columns which are being encoded.

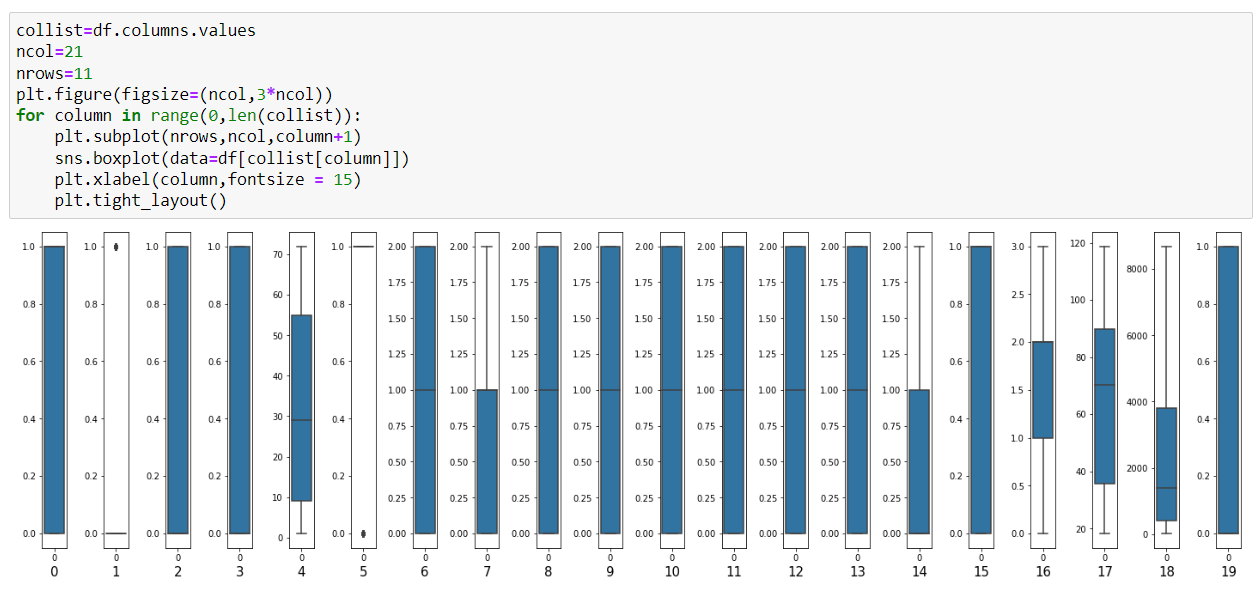


Correlation heatmap

Bright color indicates correlation values as 1 which indicates that there is stronger correlation, but dark color indicates there is negative correlation between the attributes.



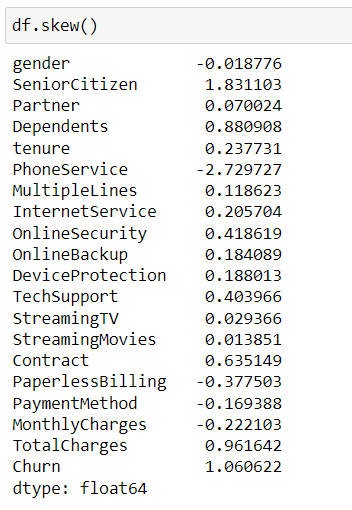
Checking for Outliers:



Observation: -

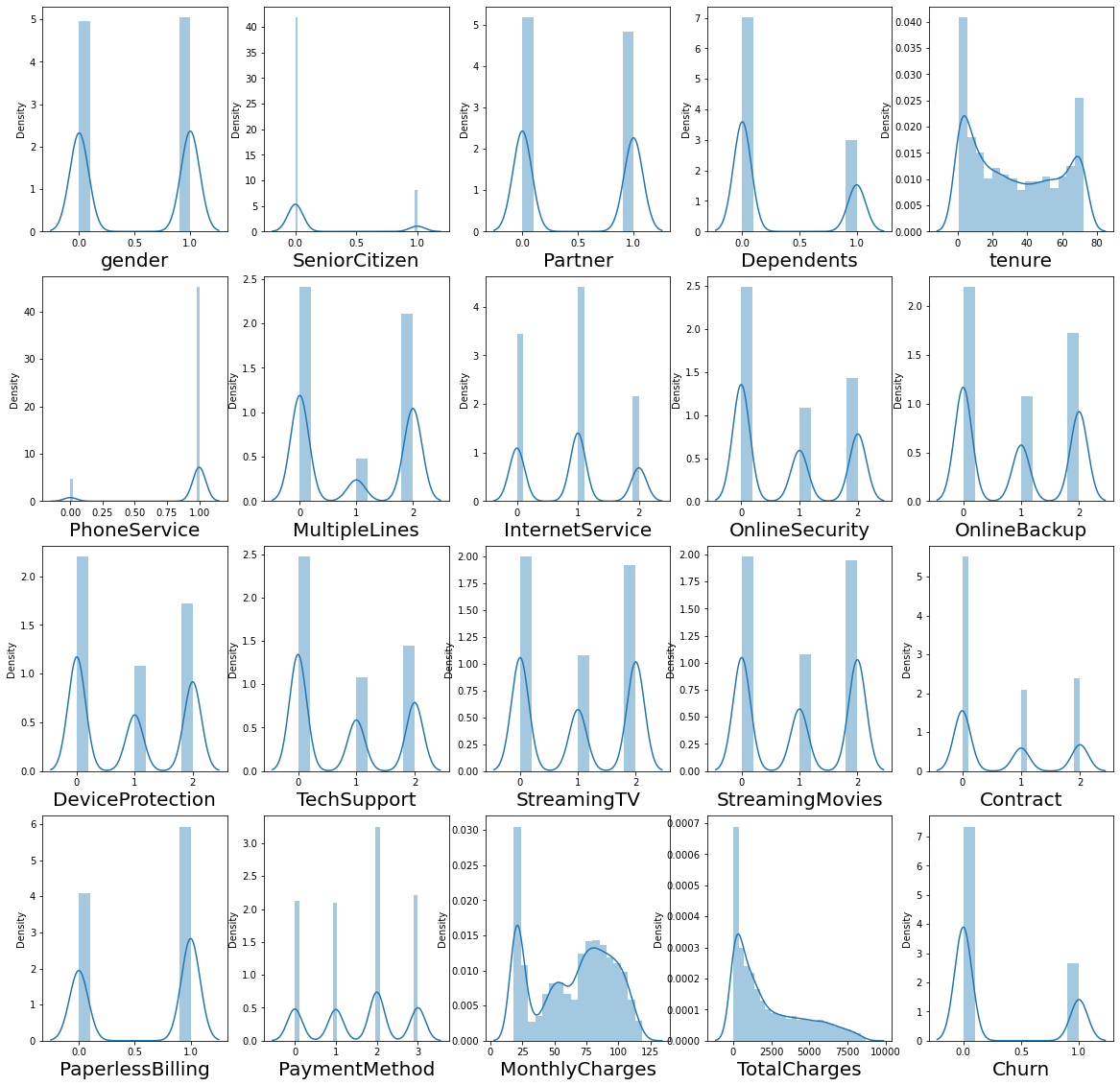
We can see Outliers are present only in 2 columns: "SeniorCitizen" and "PhoneService". But both column are categorical, so we will not remove outliers.

Checking for the skewness



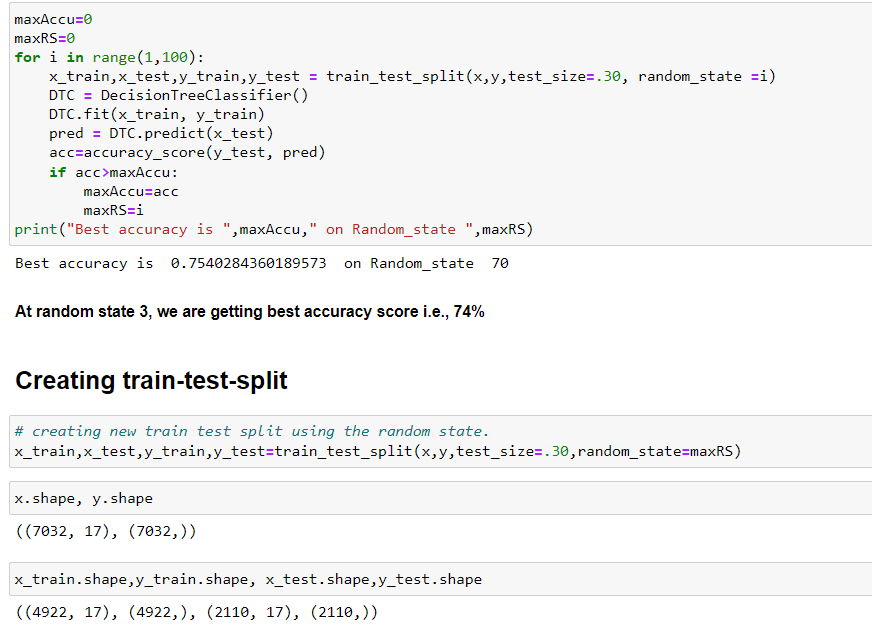
**Observation:**

* Skewness threshold taken is +/-0.25
* All the columns are not normallly distributed, they are skewed.
* Columns which are having skewness: SeniorCitizen, Dependents, PhoneService, OnlineSecurity, TechSupport, Contract, PaperlessBilling and TotalCharges.
* Since SeniorCitizen, Dependents, PhoneService, OnlineSecurity, TechSupport, Contract and PaperlessBilling are categorical column so we will not remove skewness from them.
* Only we will remove skewness from TotalCharges as this column contains continuous data.

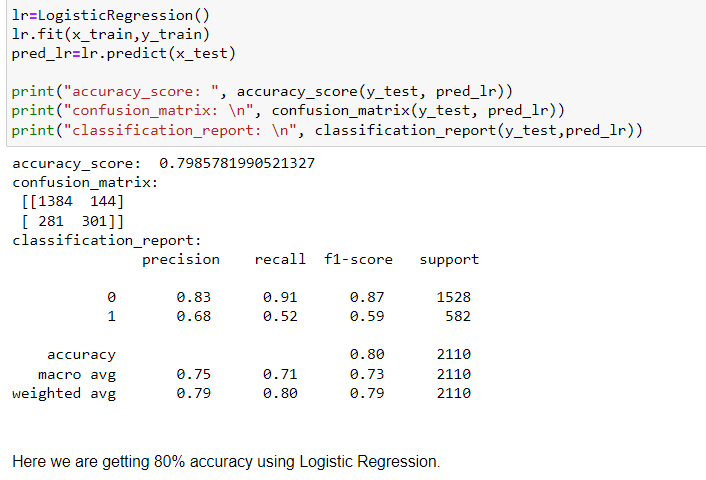


**Building Models:**

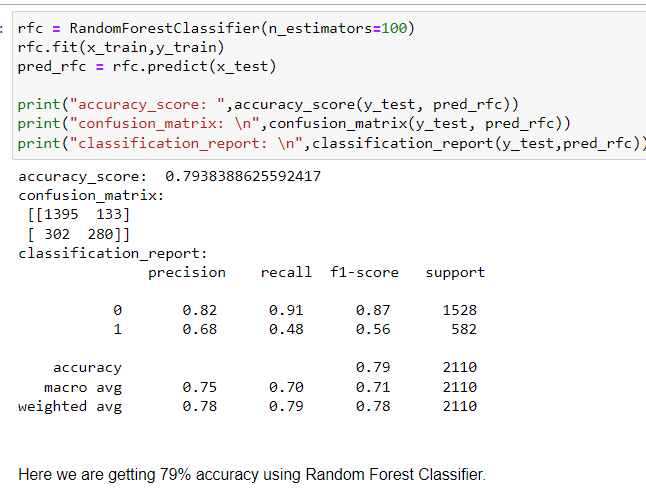
Let us standardize our data, before splitting the datasets.



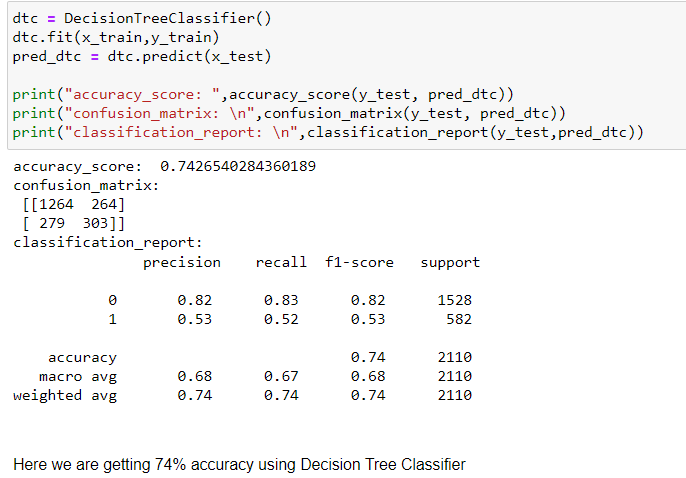
1. Logistic Regression



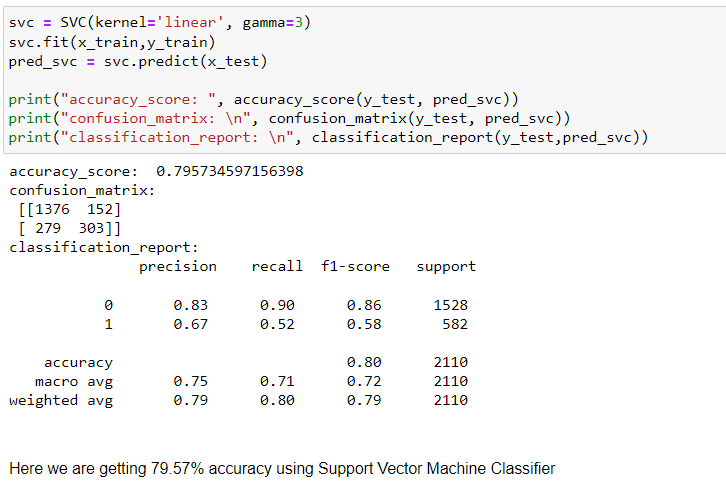
1. Random Forest Classifier



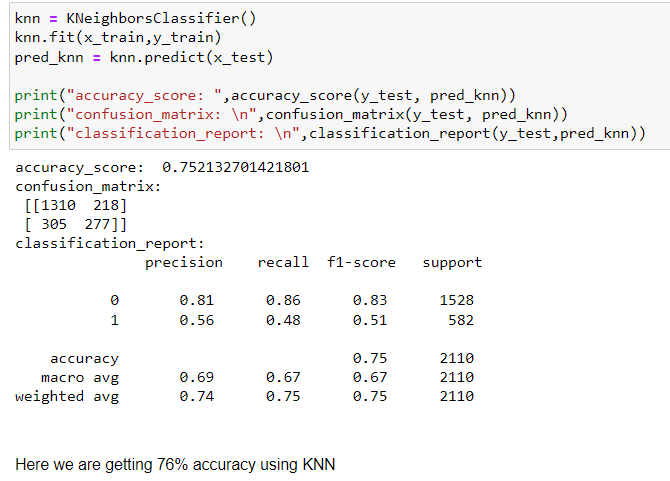
1. Decision Tree Classifier



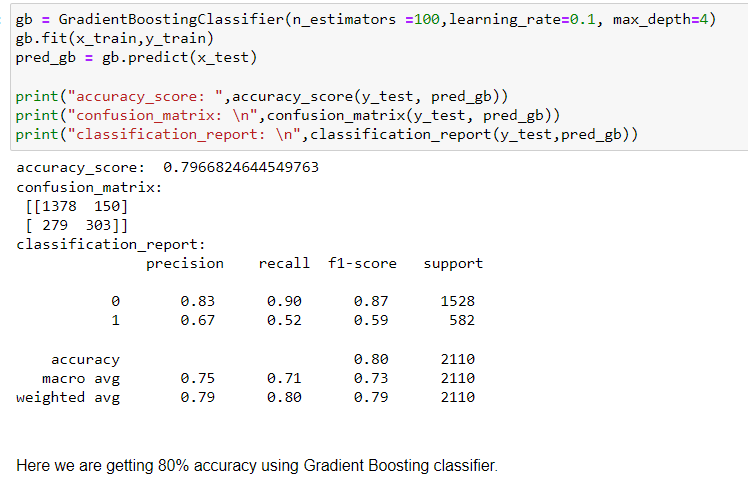
1. Support Vector Machine Classifier



1. KNN Classifier



1. Gradient Boosting Classifier

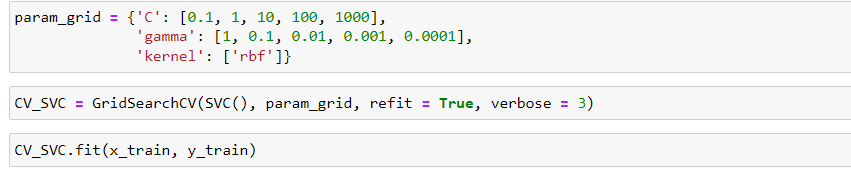


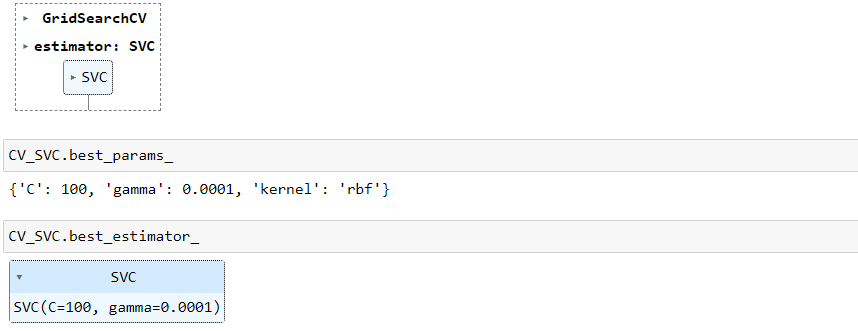
Cross Validation Score for all the model

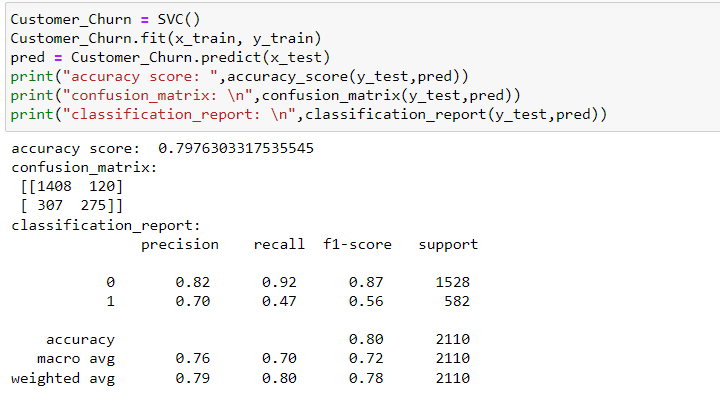


From the observation of accuracy and cross validation score and their difference we can predict that Support Vector Classifier is the best model

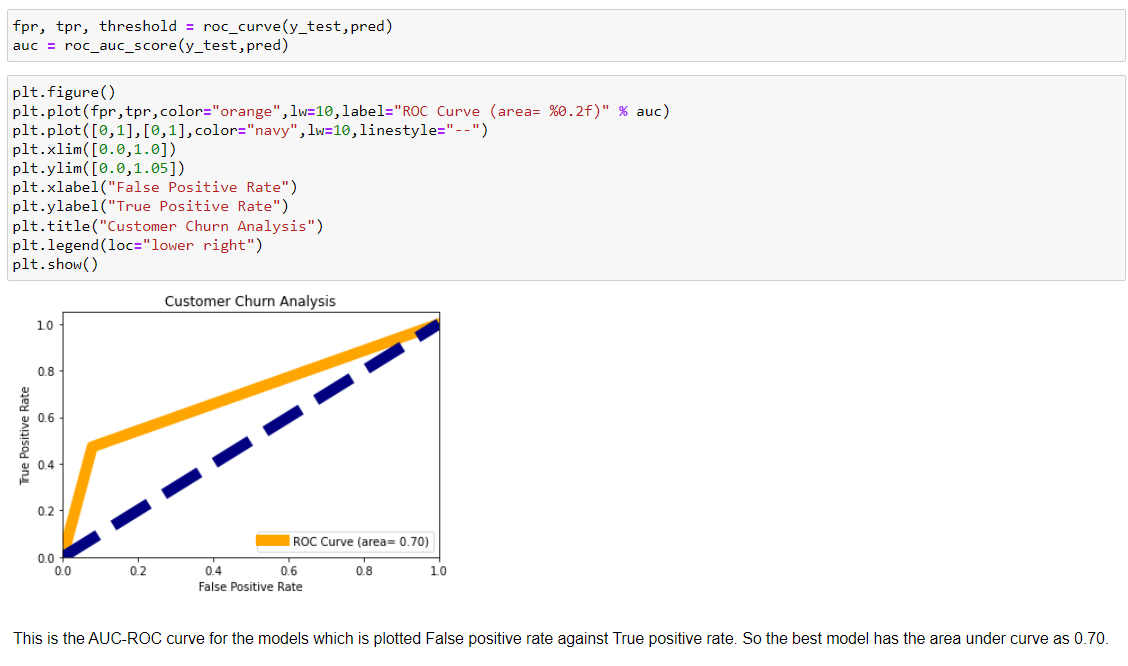
Hyper parameter tuning for best model using GridsearchCV







ROC-AUC Curve



Saving the Model

