**Customer Churn Analysis Based on Telecom Company**

****

INTRODUCTION

Telecom service suppliers in the world are facing difficult market conditions and revenue declines due to high subscriber’s churn rate. Telecommunications business derives its revenues/profits from the subscriptions from their customers. As it involves high cost of fixed network infrastructure establishments and its regular maintenance, customer loyalty is the key to their revenue generation. Nowadays, competition and government regulations have made easier for customers to switch from one network to other. People change their telecom service providers on minor of the issues like high billing, spam calls, emails, poor customer services to major issues like bad network connectivity, pathetic internet services speed, call drops etc. As a result, it is not surprising to learn that telecommunications companies have a high customer churn rate which is particularly problematic in this industry.

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 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.

**Objective:**

In this project, we build a model to predict how likely a customer will churn by analyzing its characteristics like information of customer details and information of services. The objective is to obtain a 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.

Preventing customer churn is critically important to the telecommunications sector Here, with the help of the given data set, we would be able to predict whether a customer is going to churn or not.

**Steps of the project:**

The project consists of the following process:

* **Importing Libraries and Data Loading**
* **Exploratory Data Analysis**
* **Data Cleaning and Preprocessing**
* **Univariate and Bivariate data Visualization**
* **Feature Engineering**
* **Splitting the data in training and testing sets**
* **Assessing multiple algorithms**
* **Algorithm selected and Hyperparameter tuning**
* **Saving and loading of selected the model**
* **Conclusions and Remarked**

**Data Description (Features):**

The dataset has total 21 columns and 7043 rows which has customer records of telecom company.

1. Customer ID---customer unique id
2. Gender — M/F
3. Senior Citizen — Whether the customer is a senior citizen or not (1, 0)
4. Partner — Whether customer has a partner or not (Yes, No)
5. Dependents — Whether customer has dependents or not (Yes, No)
6. Phone Service — Whether the customer has a phone service or not (Yes, No)
7. Multiplies — Whether the customer has multiple lines or not (Yes, No, No Phone Service)
8. Internet Service — Customer’s internet service type (DSL, Fiber Optic, None)
9. Online Security — Whether the customer has Online Security add-on (Yes, No, No Internet Service)
10. Online Backup — Whether the customer has Online Backup add-on (Yes, No, No Internet Service)
11. Device Protection — Whether the customer has Device Protection add-on (Yes, No, No Internet Service)
12. Tech Support — Whether the customer has Tech Support add-on (Yes, No, No Internet Service)
13. Streaming TV — Whether the customer has streaming TV or not (Yes, No, No Internet Service)
14. Streaming Movies — Whether the customer has streaming movies or not (Yes, No, No Internet Service)
15. Contract — Term of the customer’s contract (Monthly, 1-Year, 2-Year)
16. Paperless Billing — Whether the customer has paperless billing or not (Yes, No)
17. Payment Method — The customer’s payment method (E-Check, Mailed Check, Bank Transfer (Auto), Credit Card (Auto))
18. Tenure— Number of months the customer has stayed with the company.
19. Total Charges-- The amount charged to the customer monthly.
20. Monthly Charges-- The total amount charged to the customer.
21. Churn-- The Churn column (target variable) indicates whether the customer departed within the last month or not.

**Data Analysis: -**

Data Analysis is a Process of Obtaining raw data and converting it into useful information by using charts, graphs, map and other visualization technique.

**Importing Libraries**

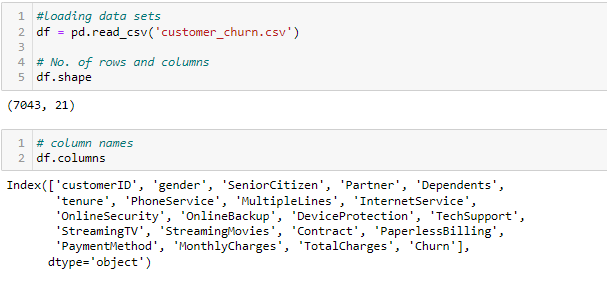
Import all the libraries at a time which are required for the process of Exploratory data analysis and building machine learning model.



**Extracting dataset**

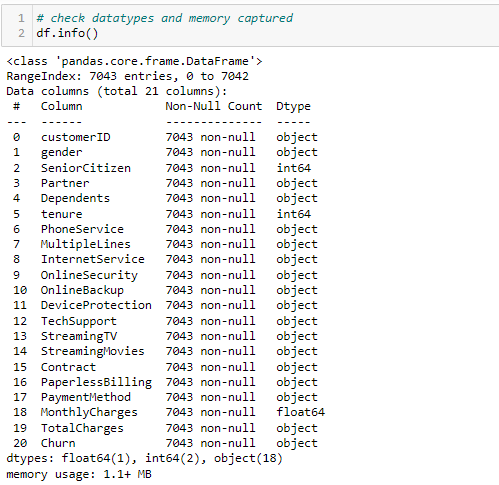
Dataset contain 7043 rows and 21 columns. We will use this data for prediction of our machine learning model from which dataset 70% of data is used for training a model and 30% for testing a model.

* Loading and saving the given Dataset in a variable.
* Find out the shapes of our dataset.
* Display the columns of dataset.



**Checking datatypes of columns**

The **info** method is used to obtain summary of the dataset, including the column names and their data types, the number of non-null values, and the amount of memory used by the dataset.

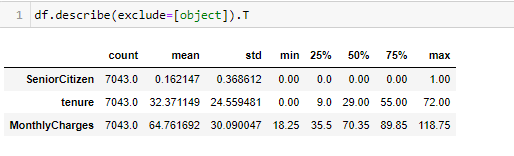
****

**Observation**

* There is both numerical and categorical type of data is present.
* We observed that Total charges is floating value but here shows the object that mean any space value may be in total charges we have to check.

**Statistical analysis**

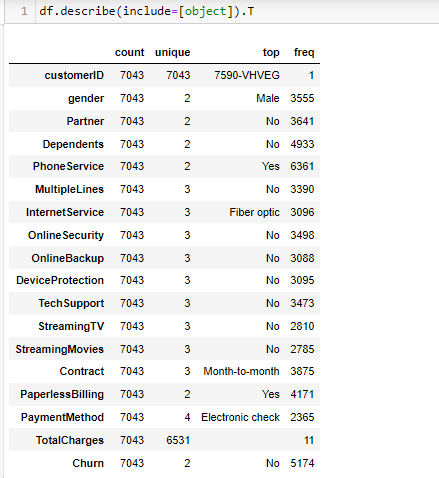
**Statistical analysis of Numerical Features** represents the count, min, max, median ,25% and 75% value of each feature which is help us to find out the presence of Nan value, outliers and skewness in our datasets.



**Observations:**

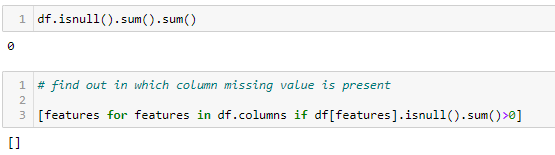
* There is no column with single unique values
* Null values is not present because the value of count for all columns are same (7043).
* Skewness are present because higher difference between 75% and max value of Monthly charges and
* Total charges is descriptive type data but did not came in descriptive analysis because may be some incorrect variable is present in this column.

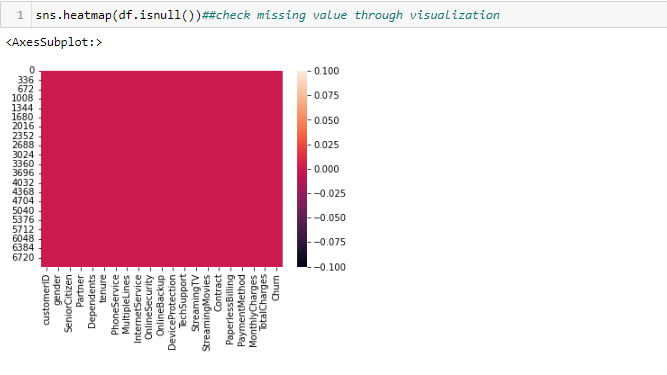
**Statistical analysis of categorical Features** represents the top, count, unique and frequent value of each column. which is help us to find out uniqueness and mode value of each column in our datasets.

****

**Check null value**

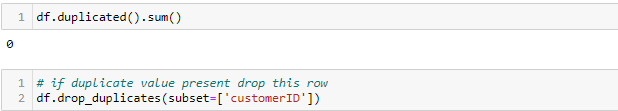
Check missing value present in any columns by coding and visualization.

.



We observed that no missing value present in our dataset.

**Check duplicate value**



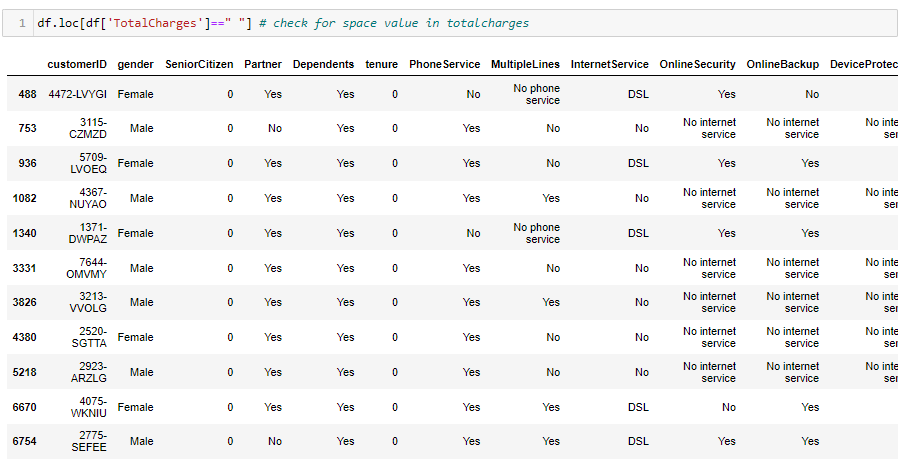
No Duplicate value present there if any duplicates present then drop this particular row.

**Data cleaning**

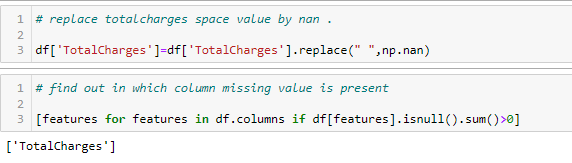
In the process of data cleaning, we fixing the missing, incorrect, incomplete, and duplicate value.

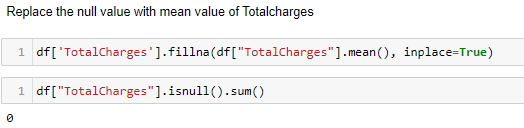
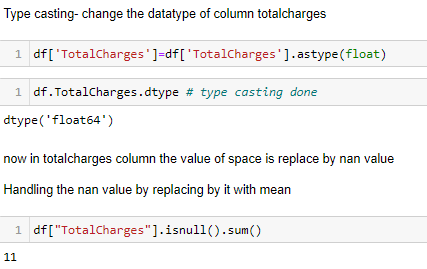
Total Charges

we observe that the column Total Charges was wrongly detected as an object. This column represents the total amount charged to the customer and it is, therefore, a numeric variable. For further analysis, we need to transform this column into a numeric data type.



Observed that we have 11 rows that contain space value firstly we will replace it with nan value then impute this nan value by using mean value after that change datatype of column object to float.

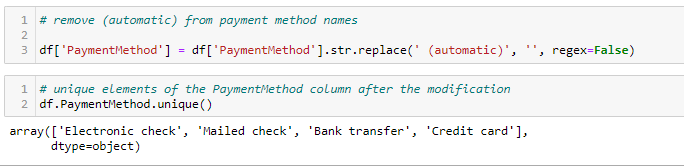




Now Total Charges is float and no null value present.

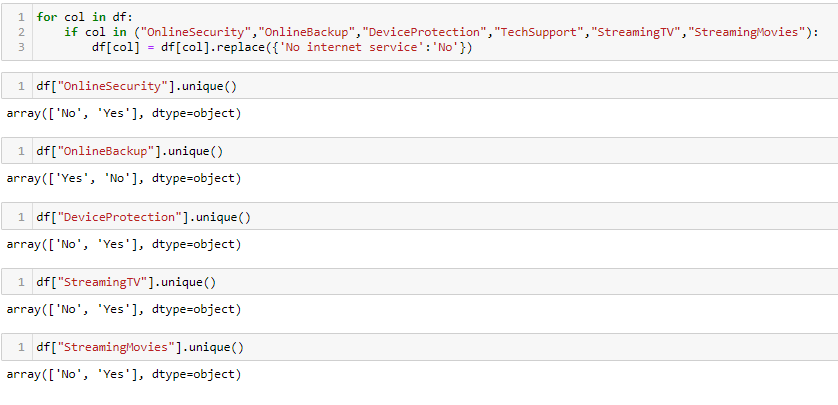
Payment Method

Some payment method denominations contain in parenthesis the word automatic. These denominations are too long to be used as for visualizations. Therefore, we remove this clarification in parenthesis from the entries of the Payment Method column.



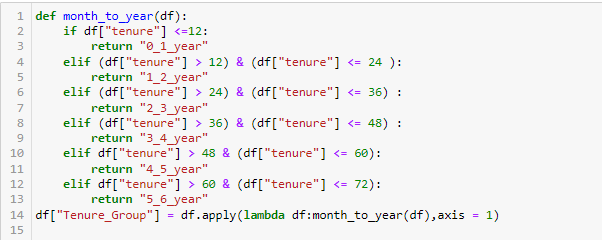
No Internet Services

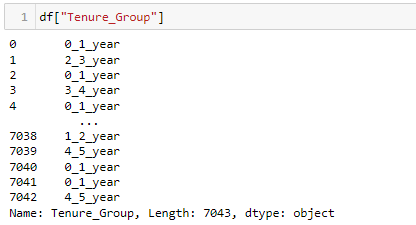
The customer who don’t have Internet services they are also not able to do any facility by using internet like OnlineSecurity, OnlineBackup, StreamingTV, StreamingMovies, and deviceProtection, So we can change the columns which has No internet service into 'No' .



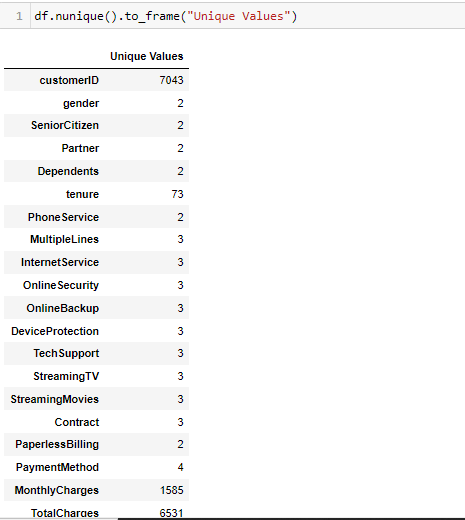
Tenure

change tenure into group of years





**Check unique values present in each Column**

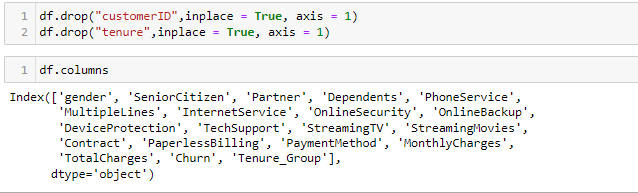
****

**Observation:**

* Customer Id, monthly charges and total charges has higher unique value.
* All value of customer id is unique and not useful for prediction so we can drop this column.
* No anyone unique value present there.

**Drop unnecessary columns**

* Drop column customerID which has lots of unique value and no need for model prediction.
* Drop column tenure because I added new column tenure group which is more understandable.



**Segregate the Numerical and Categorical features**

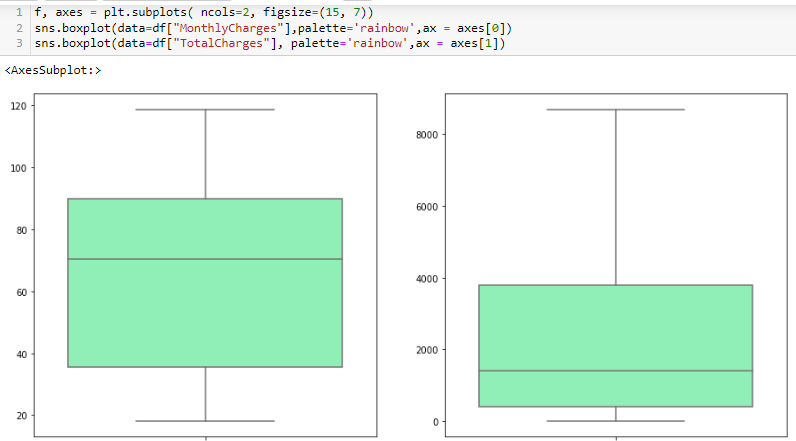
Now we separate numerical and categorical columns so that we can visualize our data easily.

****

**Univariate Analysis**

Check Outliers

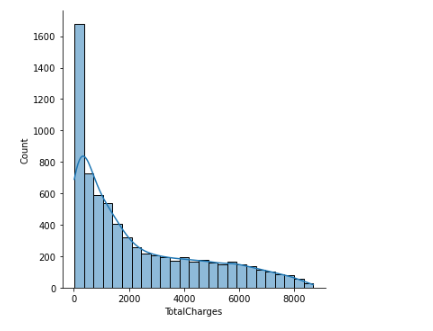
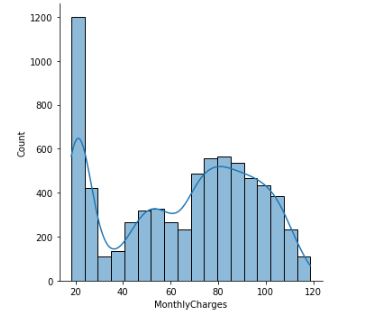
We always check outliers for numerical features. Here Senior citizen is numerical value but nature is categorical type of data so we can check outliers for Monthly Charges and Total Charges features only.



**Observation:**

* No outliers present in Total charges and monthly charges of data sets.
* If any outliers are present then it will be removed by using Z-score or IQR technique.

Check Skewness

Let’s we will check skewness is present in our dataset by using distribution plot.

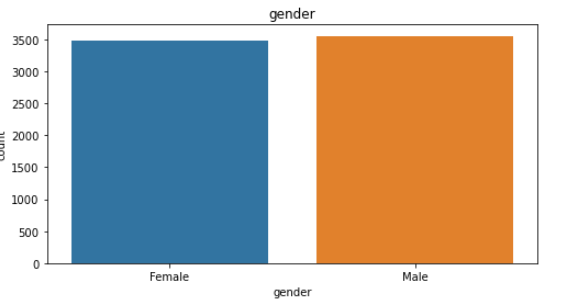
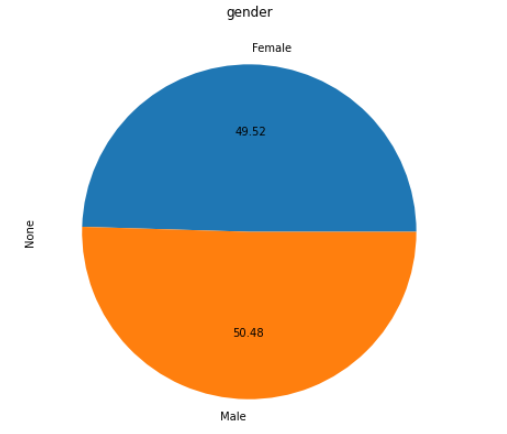
**Observation:**

Skewness is present in total charges and monthly charges so later we have to remove it by using any transform technique.

**Visual Analysis of Categorical Variable**

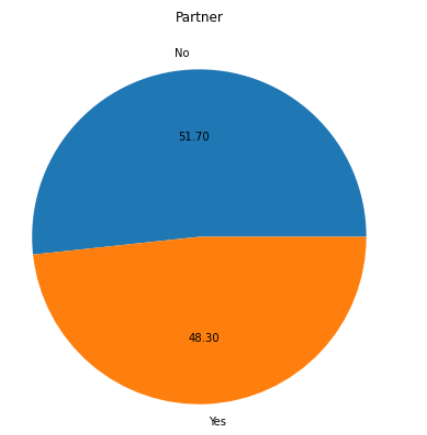
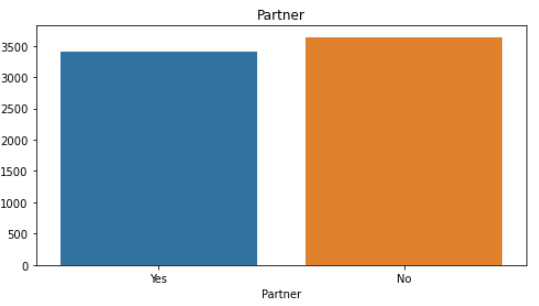
We will visualize all categorical variable by using count plot and pie chart.

* Gender

Count and percentage of male and female user almost same.****

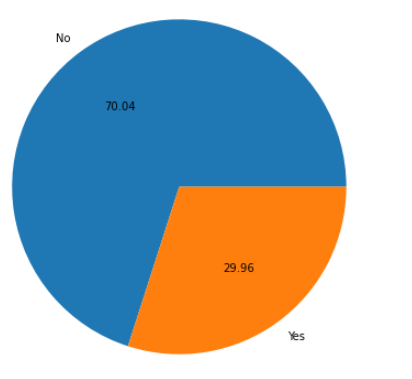
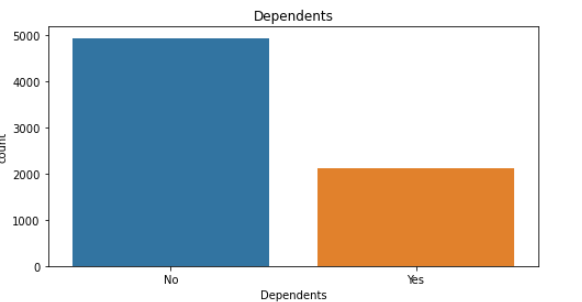
* Partner

Count and percentage of partner is little bit difference.



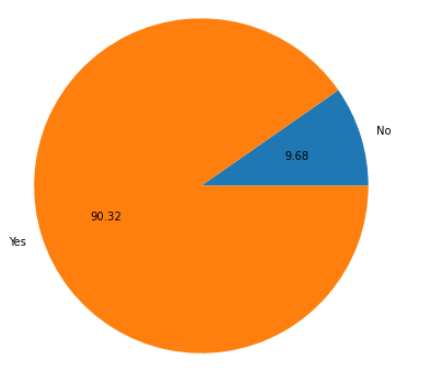
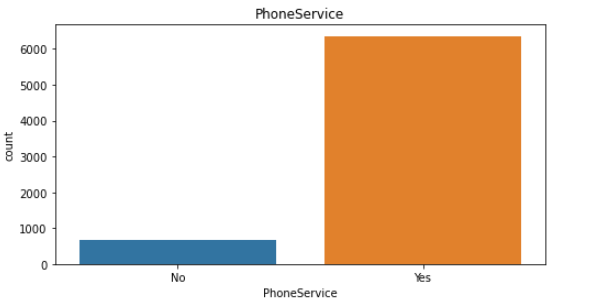
* Dependence

The percentage of independence customer are 40% higher than dependent customer.



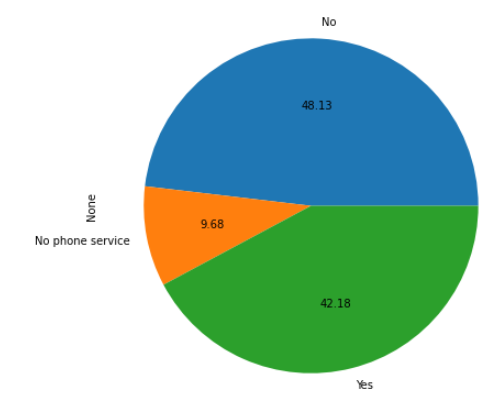
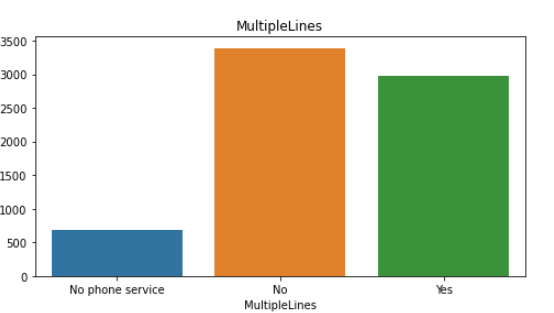
* Phone service

90% of customer preferred for phone service.



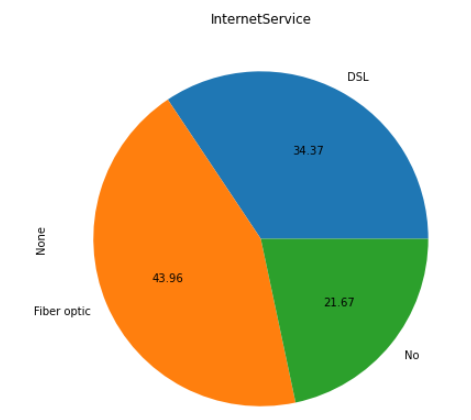
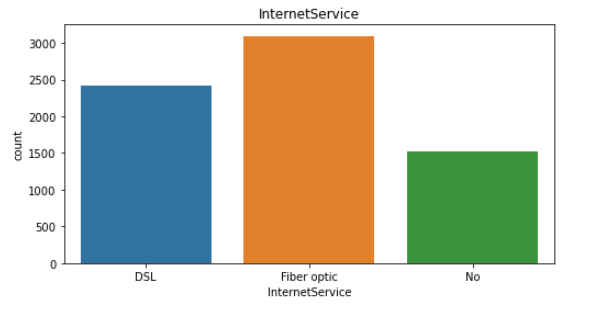
* Multiple lines

some customers are preferred for multiple lines but maximum customer is preferring for single line and very few people do not have any services.

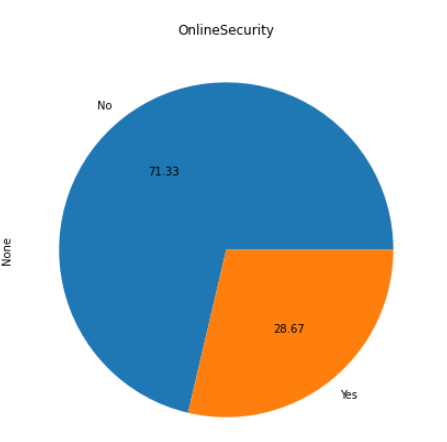
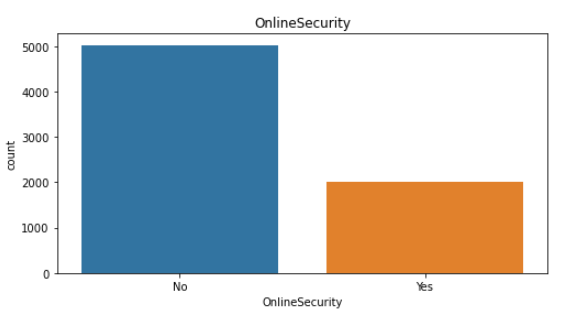


* Internet Service

The customer who had Internet services almost preferred for fiber optics.

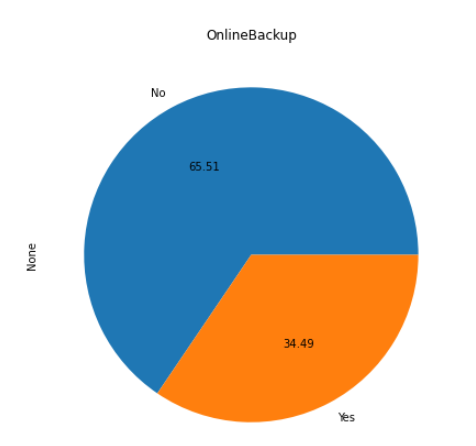
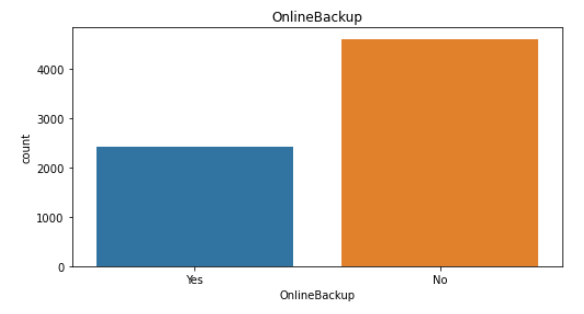


* Online security

71% customer doesn’t have online security because the customer doesn’t have internet service is not able to do this service.

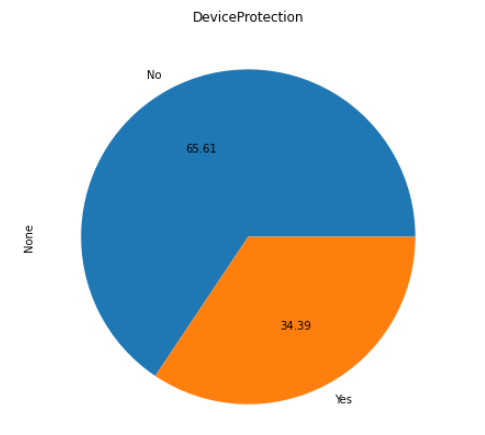
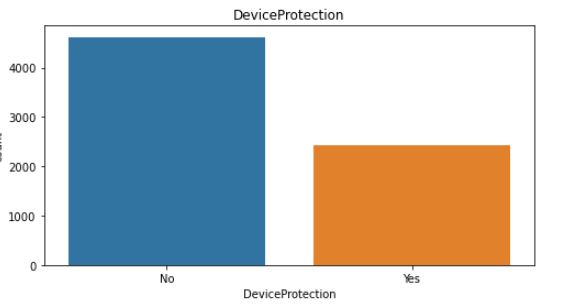
* Online Backup

65% customer doesn’t have online backup facility because the customer doesn’t have internet service is not able to do this service.



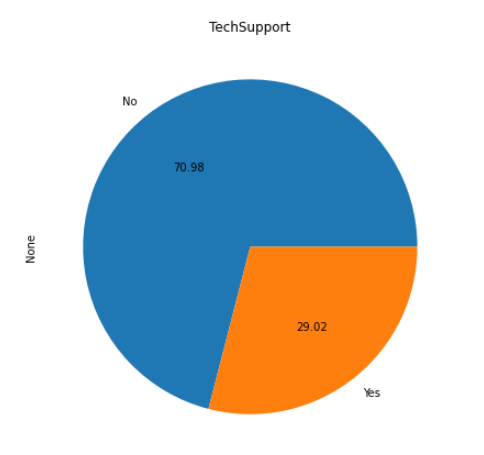
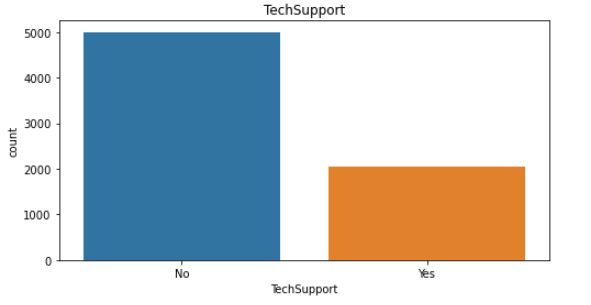
* Device Protection

The number of customers who has Device protection is less because the customer doesn’t have internet service is not able to do this service.



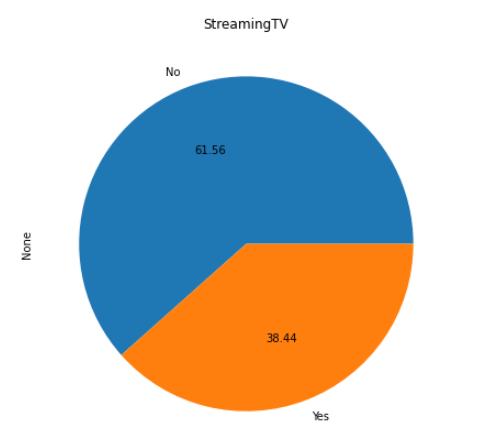
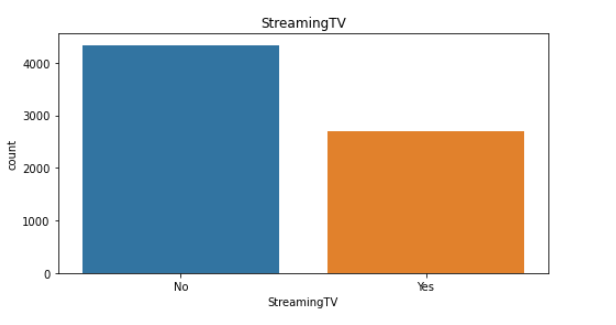
* Tech Support

The number of customers who has tech Support facility is less because the customer doesn’t have internet service is not able to do this facility.



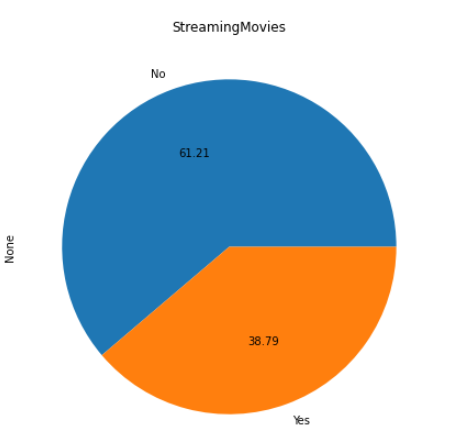
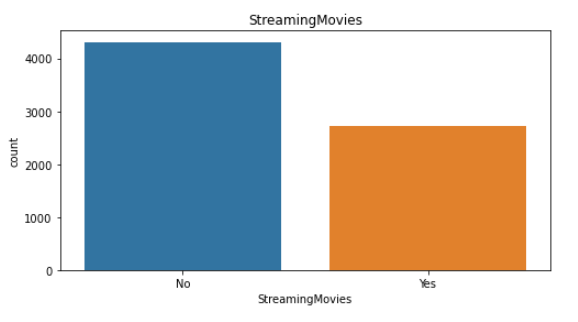
* Streaming TV

The No. of customer who had no streaming TV is higher because the customer doesn’t have internet service is not able to do Streaming TV.



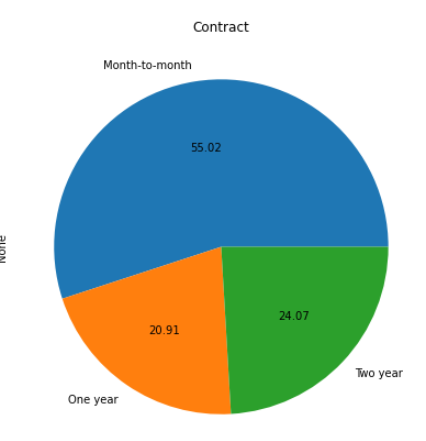
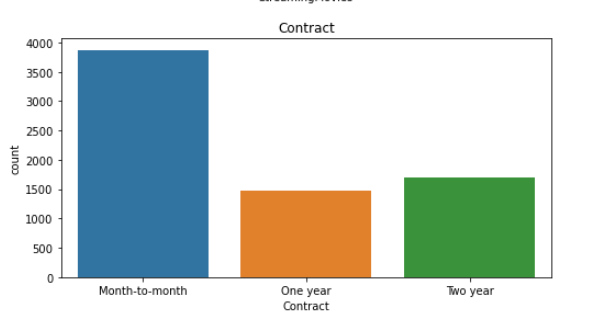
* Streaming Movie

The No. of customer who had no streaming Movie is higher because the customer doesn’t have internet service is not able to do Streaming Movie.



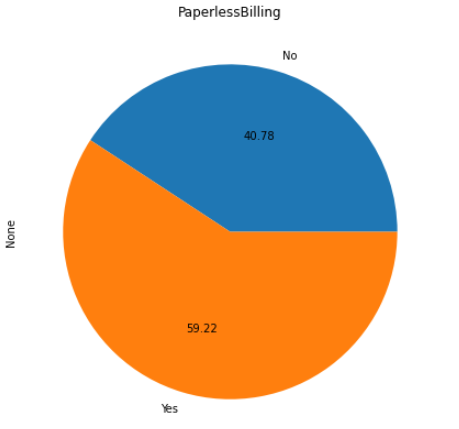
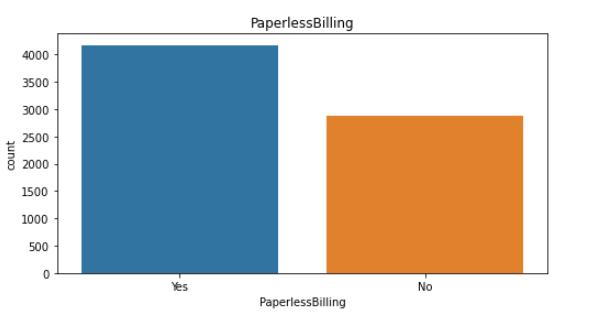
* Contract

The customer having mouth-to-mouth contract will have higher chance of churn. So, the telecom company should focus this type of customer first.



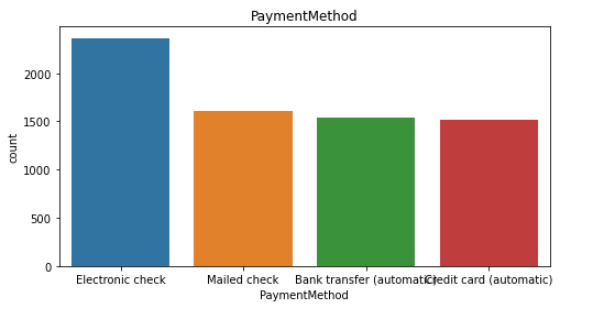
* Paperless Billing

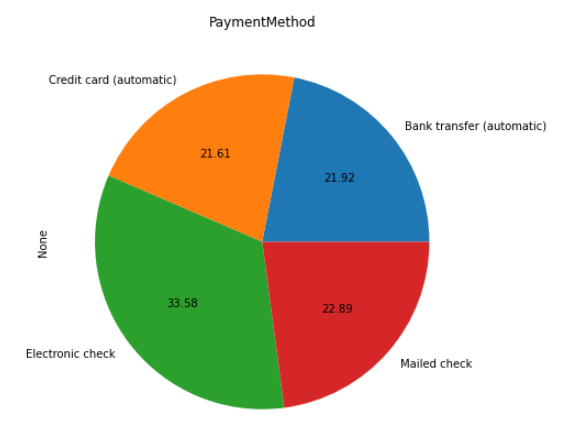
Maximum customer preferred paperless billing, so telecom operator should provide online payment mode.



* Payment method

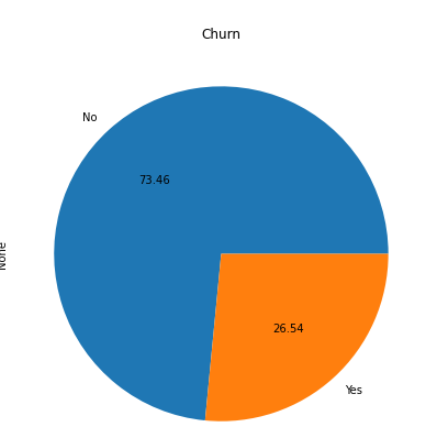
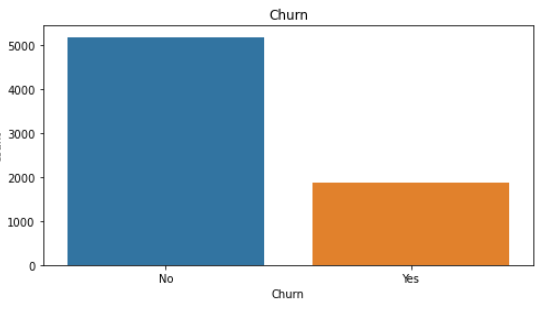
Maximum customer preferred payment by electronic cheque.





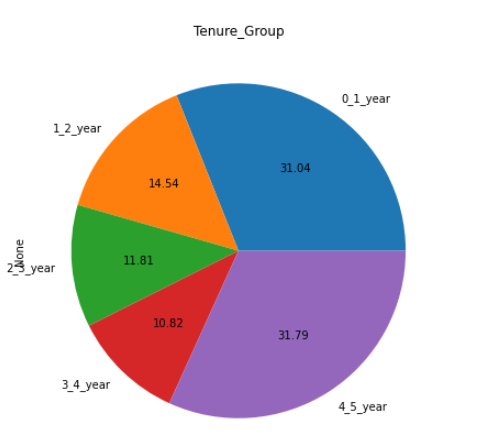
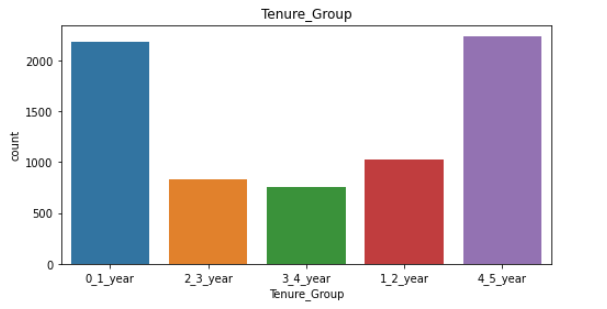
* Churn

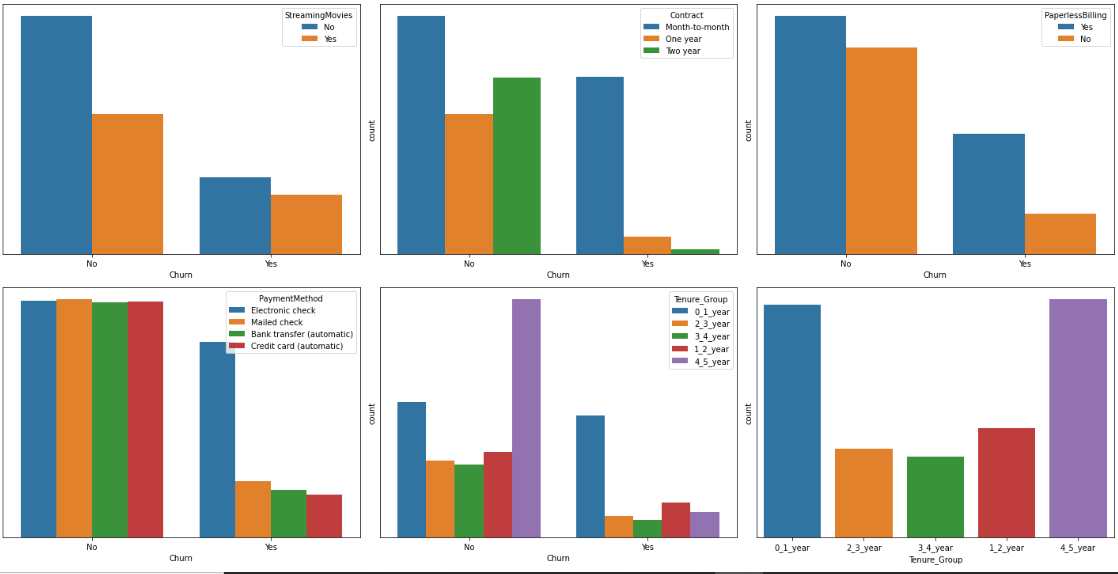
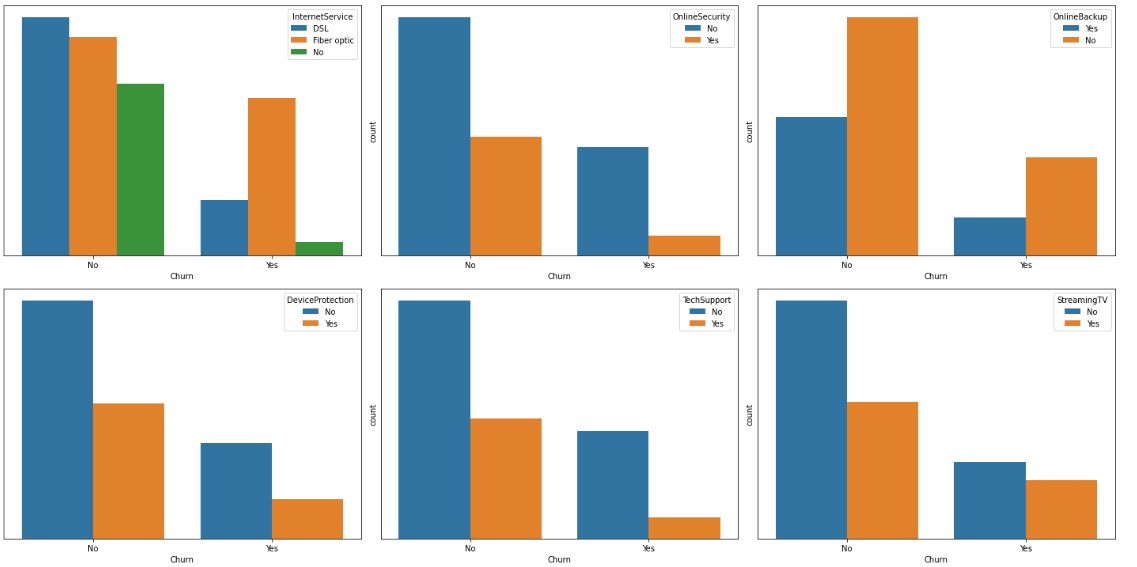
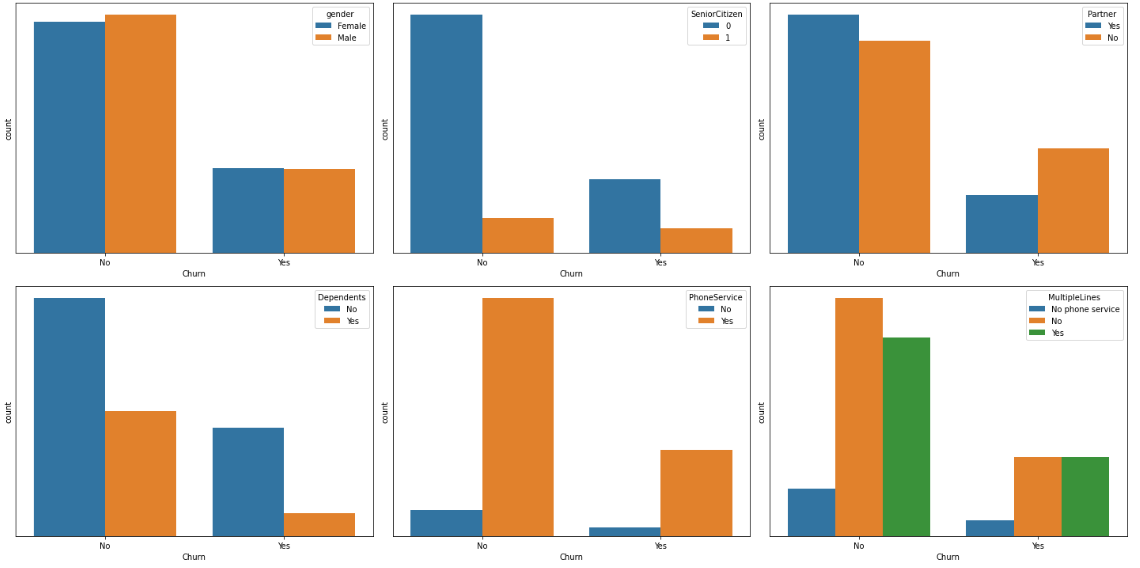
The customer churn rate is 26% in our dataset.



* Tenure Group

The customer who belongs from tenure group medium tenure group has chance to churn higher.

Bivariate Analysis:

Visualize all categorical features with respect to target variable churn by using count plot.Observation From above graph

1.The count of churn is no for maximum customer but that is not too much dependent on gender means churn rate is same for both male and female So, we can drop the gender column because it is not useful for model prediction.

2.We saw that in above pie chart only 16% Senior Citizen exist in telecom operator but now I observed that the churn rate is higher as compare to younger customers may be its possible the older people is not used phone services above the age of 70.

There is Less Senior Citizens among the group and within the senior citizens churning is less, hence we would need to focus on the younger group to reduce the number of churns through various technologies and innovations they can more relate to.

3.In given data set 48% customer preferred for partnership but the people who is partner the churn rate is less. Churning is occurring more among the customers who aren’t partners. Hence more amount of marketing would need to brough among non-partner to make them a partner or build their trust and affinity to the product.

4.70 % people are not dependent. There are more people with no dependents, If the customer has a dependent, there can be seen a higher chance to churn.

5.We saw 90% of them have phone services. People who don’t have phone service is more likely to churn

6.Multiple lines are present among 42%, Only single lines are present among 47% and No phone service is among 9.7% Highest people are churning if they have multiple line. There is almost equal churning in single line and No phone service

7.Fibre optics is used by maximum customers, Highest churning is seen for People Using Fiber Optic

8.Online backup is not present among 43% users; Highest churning can be seen when there is no online backup.

9.Device protection is not present in 44%, There is highest churning when there is no Device Protection. Hence Having a Device Protection is more important than having internet service.

10.Tech support is not present among 49%, The highest churning when there is no tech support. Hence Having a tech support is more important as compare to having internet service.

11.Streaming TV is not present among 40%. Having or not having streaming TV does not contribute much towards churning. Both who had and didn’t have Streaming TV have almost equally churned

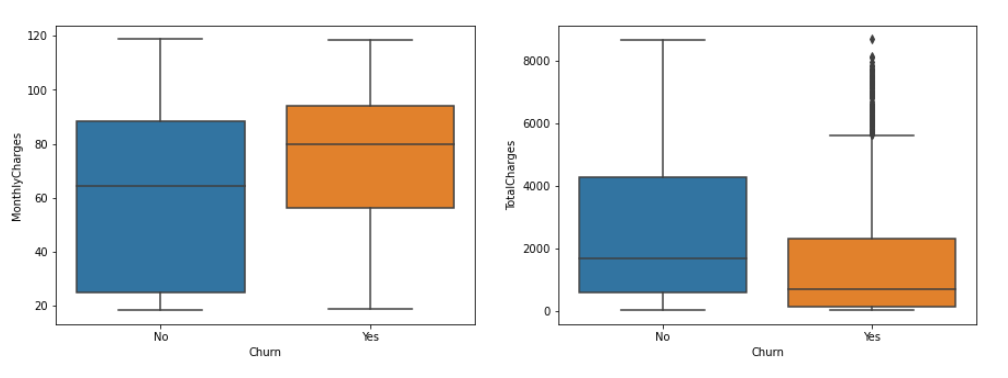
12.55% customers have month to month contract. Churning is highest among month-to-month contracts So need to focus who had among month-to-month contracts.

13.The Customer who preferred Paperless billing is among 60% employees and churning is highest among them.

14.Electronic cheque is used by 34% employees and churning is highest among them.

15.The people whose tenure is higher 4-5 years is less chance to churn.

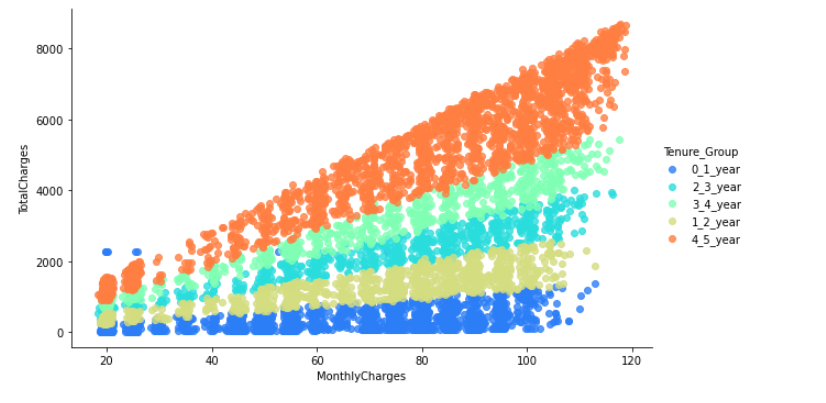
**Bivariate Analysis of numerical variable with respect to customer churn**



**Observation:**

When the total charges increase the churn rate is also increases. When monthly charges are less then less chance to churn the no. of users.

**Relation between Total Charges Monthly Charges and Tenure Group**

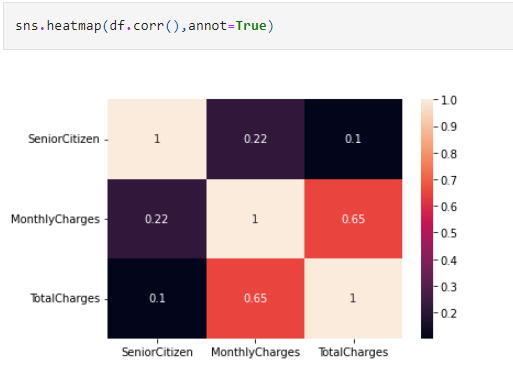


**Observation**:

when the monthly charges are increases obviously Total charges is also increases but who belongs from 4–5-year tenure group pay maximum bill and less chance to churn so, need to be focus the customer who belongs from tenure-group 1-2 because they played minimum bill and more chance of churn.

Correlation

Now we check correlation of numerical variable



**Observation:**

Senior citizen is categorical type data but the monthly charges and Total charges is highly Correlated to each other’s.

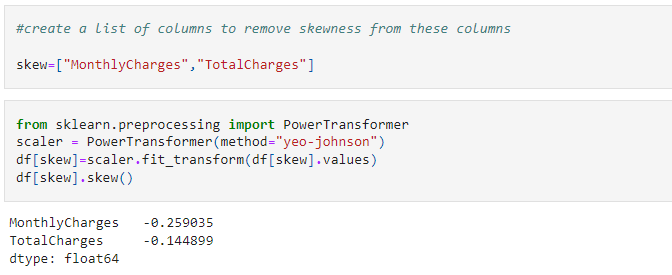
**Data Preprocessing**

After understanding the nuances of our dataset and the main issues in the data through EDA, data preprocessing comes into play by preparing our dataset for use in the model.

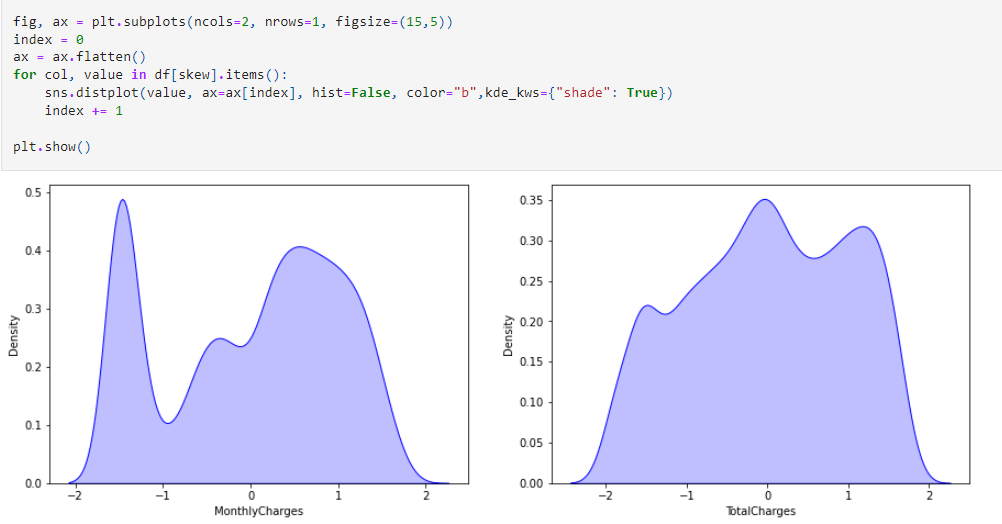
Some preprocessing techniques in order to make the data more usable. These techniques will facilitate its use in machine learning (ML) algorithms, reduce the complexity to prevent overfitting, and result in a better model.

**Remove Skewness**

In above distribution plot we notice that skewness is present in Totalcharges and monthlycharges column now we remove that skewness by using Transform technique.



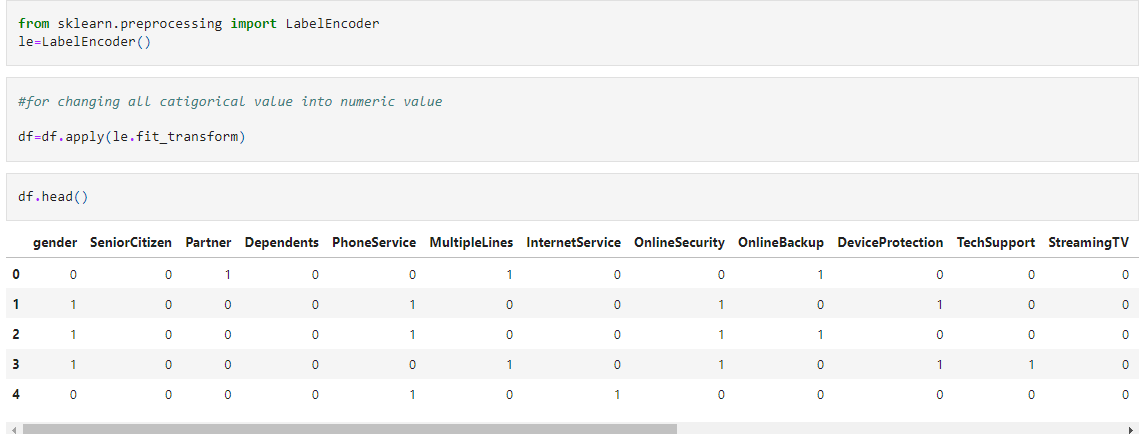
**Check skewness are removed are not**

****

Now we have successfully removed the skewness from above two columns.

**Label Encoding**

convert categorical data into numerical variable by using Label encoding technique.

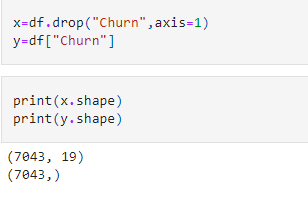
****

**Model Building**

Churn is a target variable and it is binary type so, we have to take classification models to predict

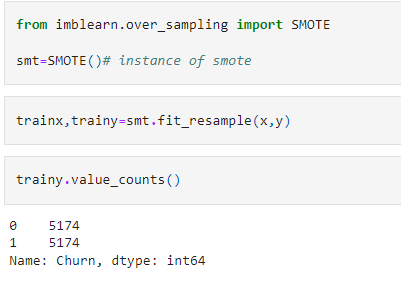
**Feature and Label Selection:**

First, we create a variable **x** to store the **independent attributes** of the dataset. Additionally, we create a variable **y** to store only the **target variable** (**Churn**).



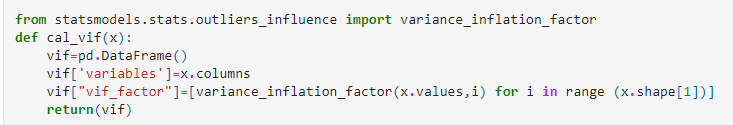
**Balancing the target dataset:**

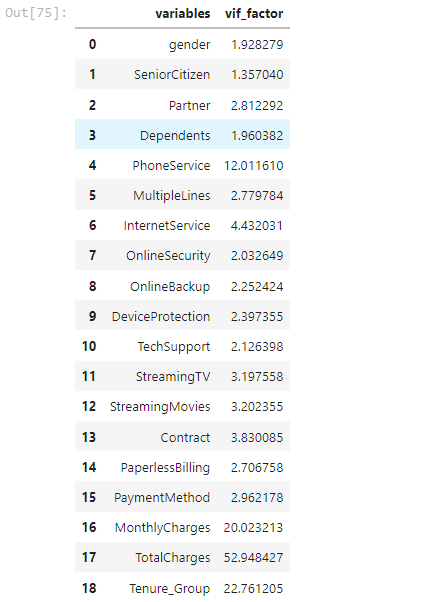
Given dataset is unbalanced so by using oversampling method we will balance our data.



**Multicollinearity:**

By using vector inflammation factor, we will check if any multicollinearity presents in our columns.



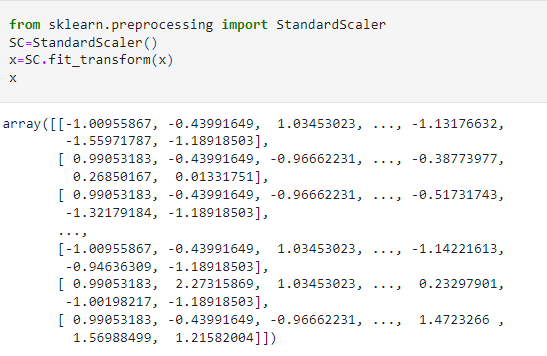
****

We observed that Total charges column is suffering from multicollinearity so we can delete it.

**Data normalization:**

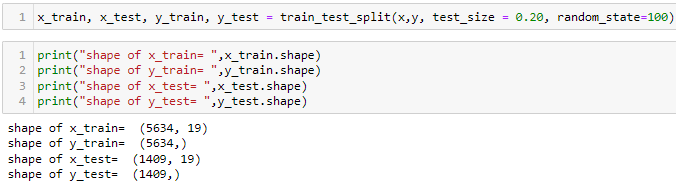
**Data normalization** transforms multiscale data to the same scale. After normalization, all variables have a **similar influence** on the model, improving the stability and performance of the learning algorithm.

Scaling our data by using standard scaler.



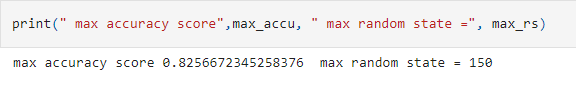
Now our data is in range between 0 to 1.

**Splitting the data in training and testing sets**

****

The first step when building a model is to **split the data into two groups**, which are typically referred to as**training and testing sets**. The training set is used by the machine learning algorithm to build the model. The test set contains samples that are not part of the learning process and is used to evaluate the model’s performance.

**Finding Best Random State and splitting the data:**

****

**Classification model selection**

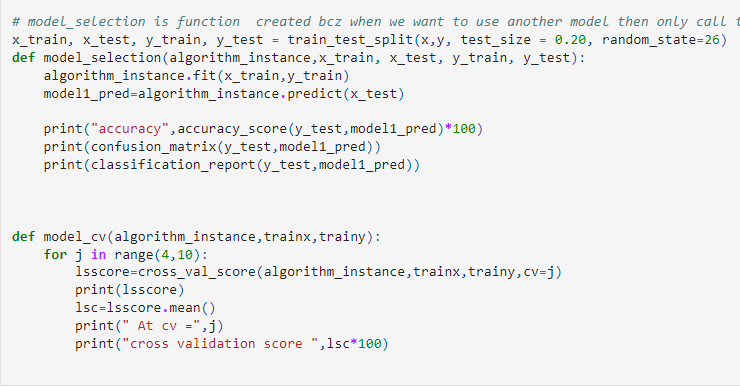
Churn is a target variable and it is binary type so, we have to take classification models to predict.

**Algorithm selection is a key challenge in any machine learning project** since there is not an algorithm that is the best across all projects. Generally, we need to evaluate a set of potential candidates and select for further evaluation those that provide better performance.

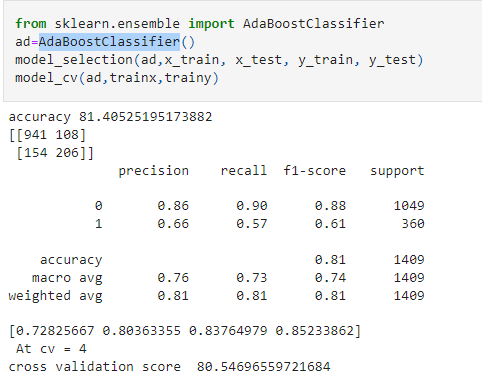
In this project, we compare **6 different algorithms**

1. **Ada booster**
2. **K Nearest Neighbors**
3. **Logistic Regression**
4. **Support Vector Machines**
5. **Random Forest**
6. **Decision Tree**
7. **Gradient Boosting**

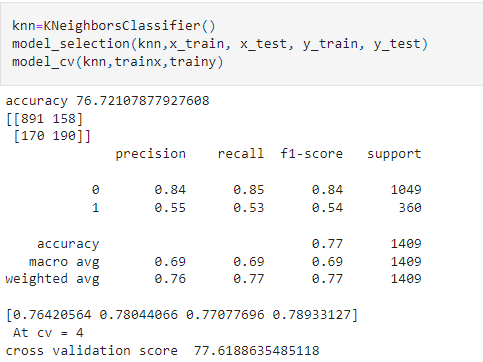
Now we will define a function for model accuracy and cross validation and called it.



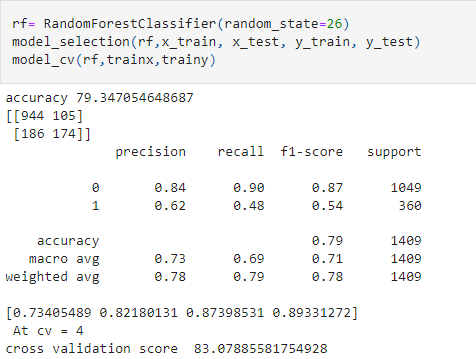
**AdaBoostClassifier**

****

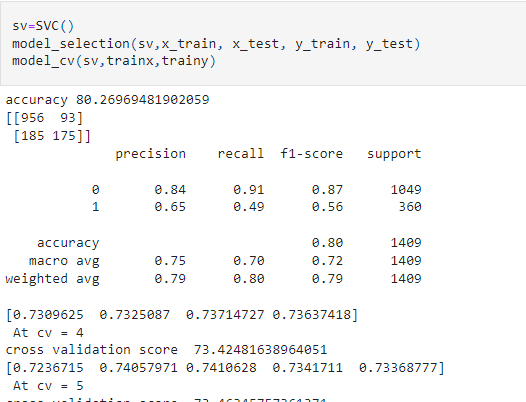
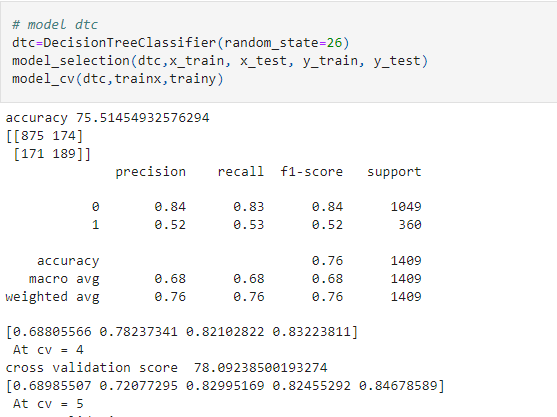
**KNeighborsClassifier**

****

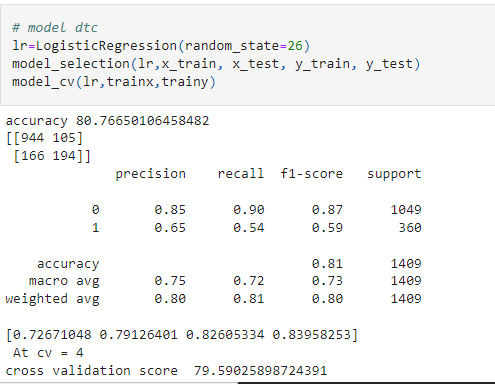
**RandomForestClassifier**

****

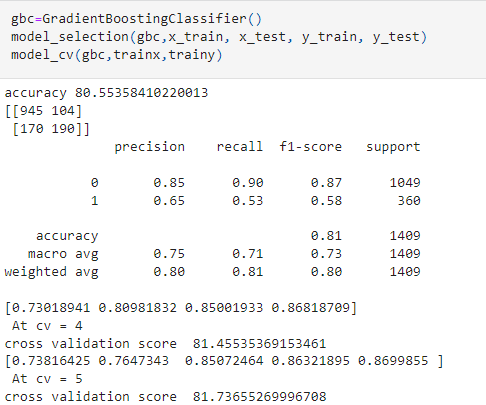
**Support Vector Classifier**

**DecisionTreeClassifier**

**LogisticRegression**

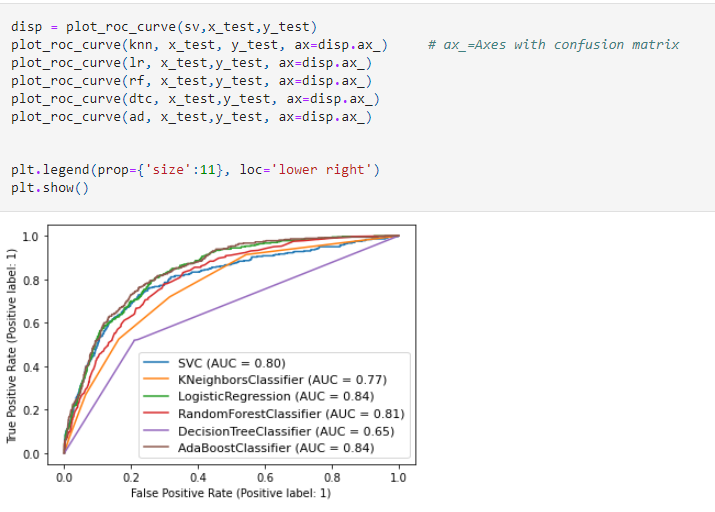


**GradientBoostingClassifier**

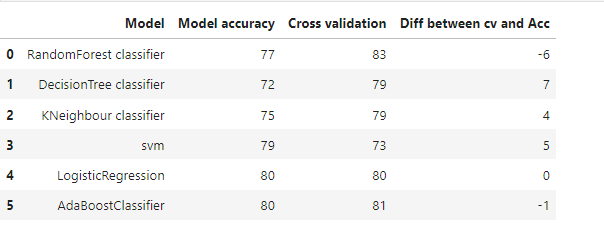
****

**ROC\_AUC\_Curve:**

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings, Logistic Regression has ROC AUC highest score, hence choosing it is our best model.

****

**Compare Performance of All Model:**

****

**Hyper Parameter Tuning for Best Mode**

It is important to bear in mind thatwe have trained all the algorithms using the default hyperparameter**s**. The accuracy of many machine learning algorithms is highly sensitive to the hyperparameters chosen for training the model. A more in-depth analysis will include an evaluation of a wider range of hyperparameters (not only default values) before choosing a model (or models) for hyperparameter tuning.

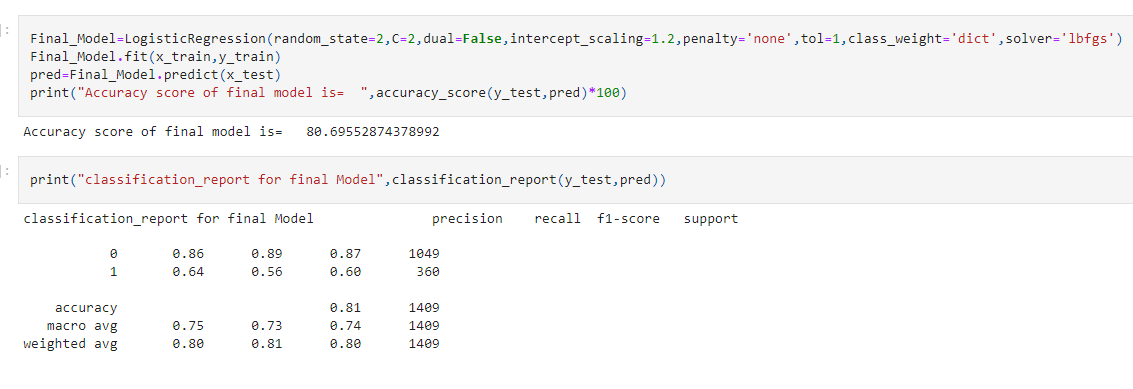
* Choosing best model as Logistic Regression" Model
* minimum difference between cross validation and Accuracy score is 0
* Roc\_Auc score is maximum

****

**Applying Best Parameters for our best model**

The next step in the machine learning process is to perform hyperparameter tuning. The selection of hyperparameters consists of testing the performance of the model against different combinations of hyperparameters, selecting those that perform best according to a chosen metric and a validation method.

There are multiple techniques to find the best hyperparameters for a model. The most popular methods are (1) grid search, (2) random search, and (3) Bayesian optimization. Grid search test all combinations of hyperparameters and select the best performing one. It is a really time-consuming method, particularly when the number of hyperparameters and values to try are really high. Now we are using grid search for hyperparameter tuning.

****

**Confusion Matrix of final model:**

The confusion matrix, also known as the error matrix, is used to evaluate the performance of a machine learning model by examining the number of observations that are correctly and incorrectly classified. Each column of the matrix contains the predicted classes while each row represents the actual classes or vice versa. In a perfect classification, the confusion matrix will be all zeros except for the diagonal. All the elements out of the main diagonal represent misclassifications**.** It is important to bear in mind that the confusion matrix allows us to observe patterns of misclassification (which classes and to which extend they were incorrectly classified).

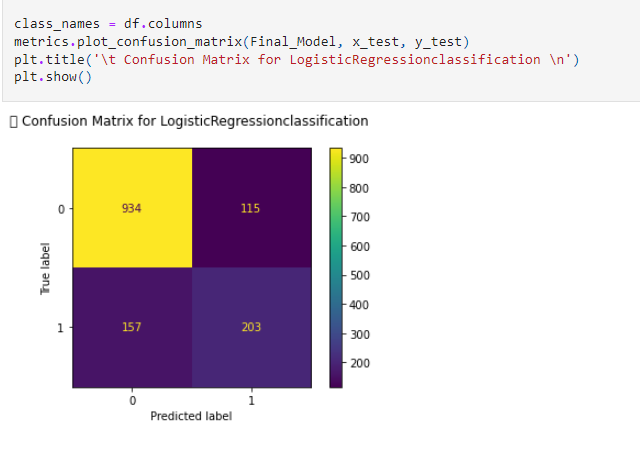
Inbinary classification problems, theconfusion matrix is a2-by-2 matrix composed of 4 elements:

**TP (True Positive)**: number of patients with spine problems that are correctly classified as sick.

**TN (True Negative)**: number of patients without pathologies who are correctly classified as healthy.

**FP (False Positive)**: number of healthy patients that are wrongly classified as sick.

**FN (False Negative)**: number of patients with spine diseases that are misclassified as healthy.

**Saving, Loading and Prediction of the best Classification ML model:**



**Our model is now ready to predict customer churn or not**

|  |
| --- |
| ***Accuracy Score: 0.80***  ***Precision: 0.84***  ***Recall: 0.90***  ***F1 score: 0.87***  ***ROC\_AUC\_SCORE: 0.81*** |

**Conclusion and Remarked**

In this project, we have walked through a complete end-to-end machine learning project using the telecommunications customers dataset. We started by cleaning the data and analyzing it with visualization. Then, to be able to build a machine learning model, we transformed the categorical data into numeric variables. After transforming the data, we tried 7 different machine learning algorithms using default parameters. Finally, we tuned the hyperparameters of the **Logistic Regression** (best performance model) for model optimization, obtaining an**accuracy of nearly 80%** .

For more and clear steps please use the below links to access the codes and results:

https://github.com/raganeeverma/Blog/blob/main/Telecom\_customer\_churn-checkpoint.ipynb