# INTRODUCTION

In this study, a machine learning model is proposed to predict a customer churn on telecom industry to keep the customer on hold. During the development and evaluation of our model, we will show the code used for each step followed by its output. This will facilitate the reproducibility of our work.

## **Business Problem Framing:**

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.

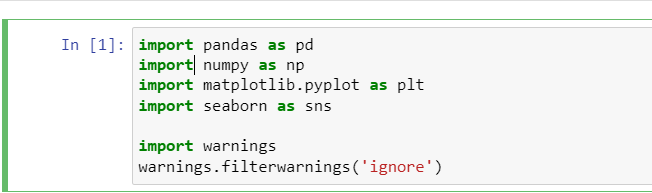
Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low

Overview:

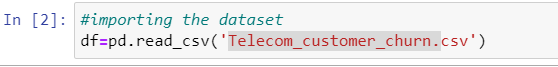
Data\_Info

In this model we have examine customer data from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models.

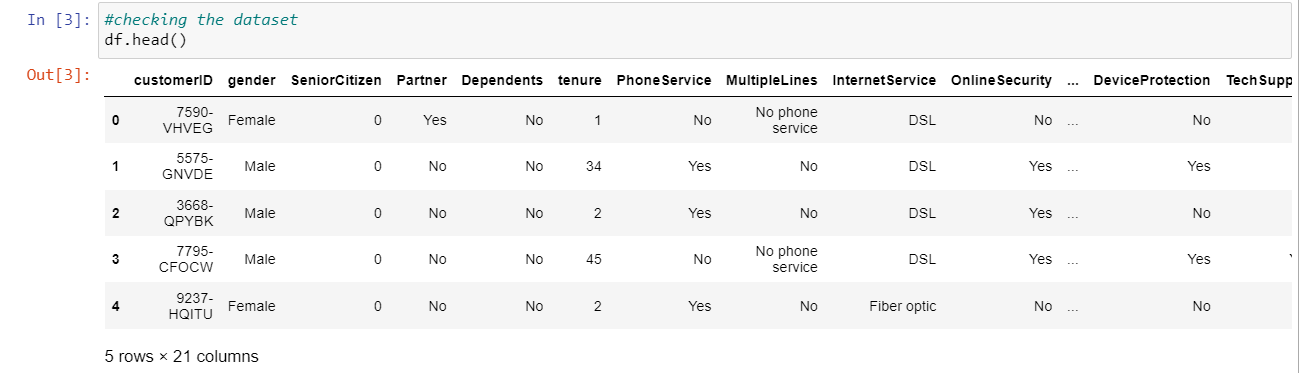
Importing the required Libraries:



Loading The Dataset:



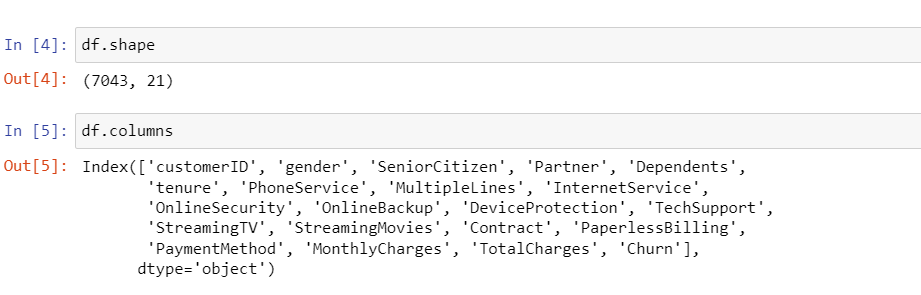
Checking Rows and Columns in Df:



Here we could observe that there is 21 columns in df ,We have both Categorical and Numerical data’s in it.

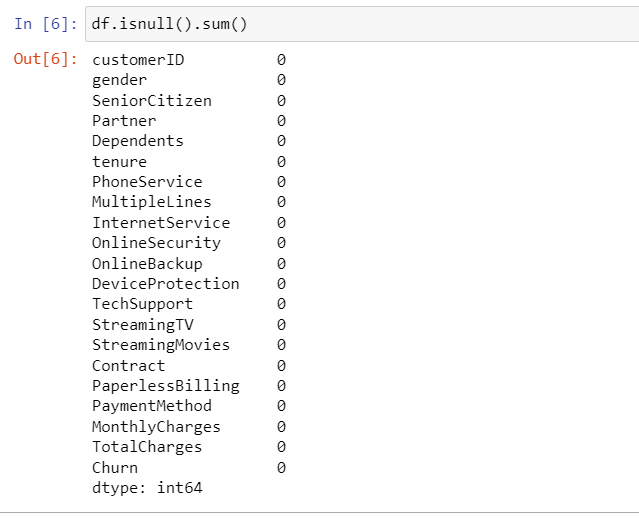
EDA:

Checking count of Rows and columns and Name of columns in df:



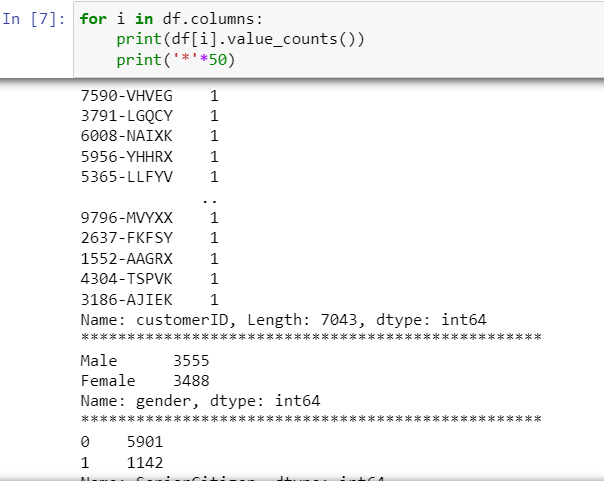
* The dataset conmtains almost 7043 Rowns and 21 columns
* The name of the columns are also mentioned above

Checking If any Null Va;lues present in Dataset:



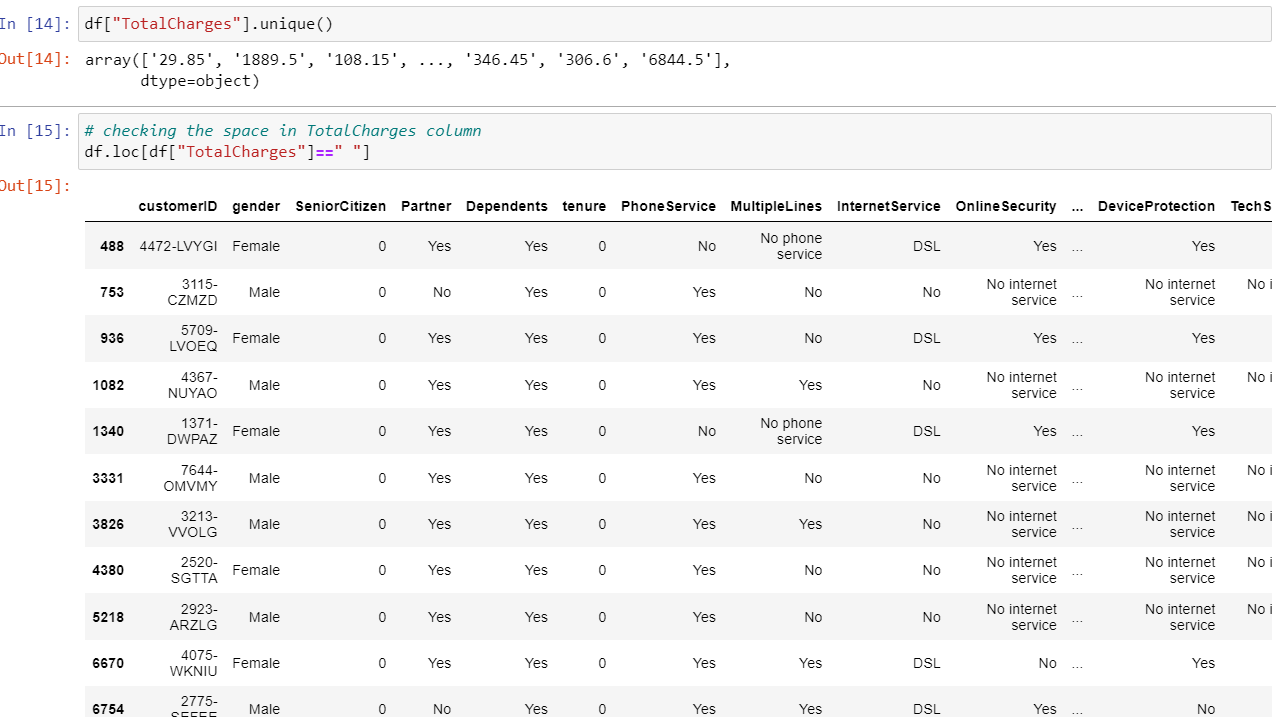
We could observe that there is no presence of null values in df ,Hence we can preoceed for further steps.

Checking Count of Each column:



* Here the obesevation is the data is distributed in to Continuous and categorical
* I could observe that the column-TotalCharges is continuous but the data shows it like Object DataType .
* Will check the count once again and check for spaces

Handling Spaces in DF:



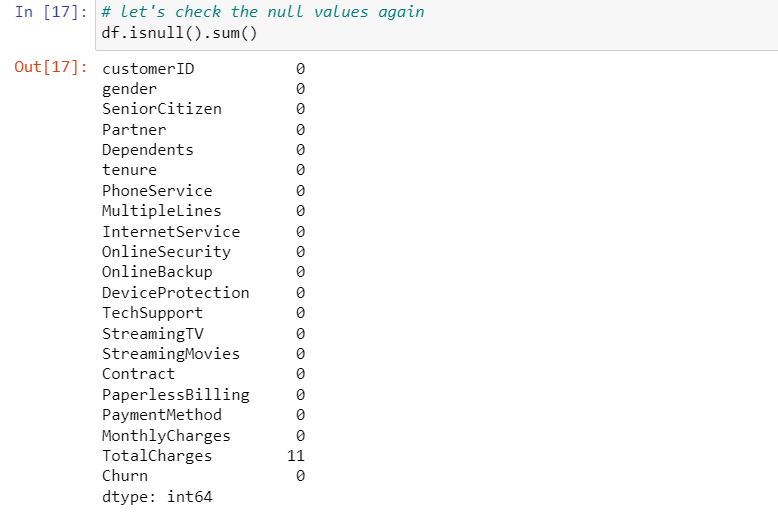
From above We could observe that the Column\_TotalCharges is in float type but due to spaces in rows it is showing like Object Datatype

The major reason to change the dtype is python can understand numbers than object

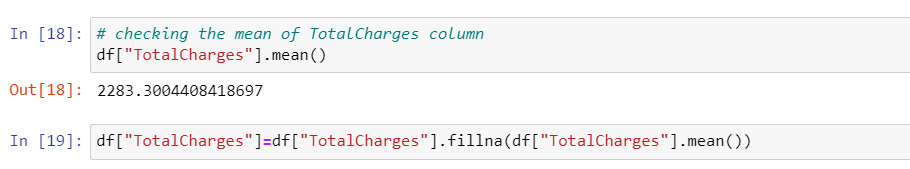
* Let's fill this column by some values and then we will convert this into float type.



Since the column Total Charges had all float numbers in the dataset but due to some reason it showed as object data type. For this reason, we have converted it into float data type.

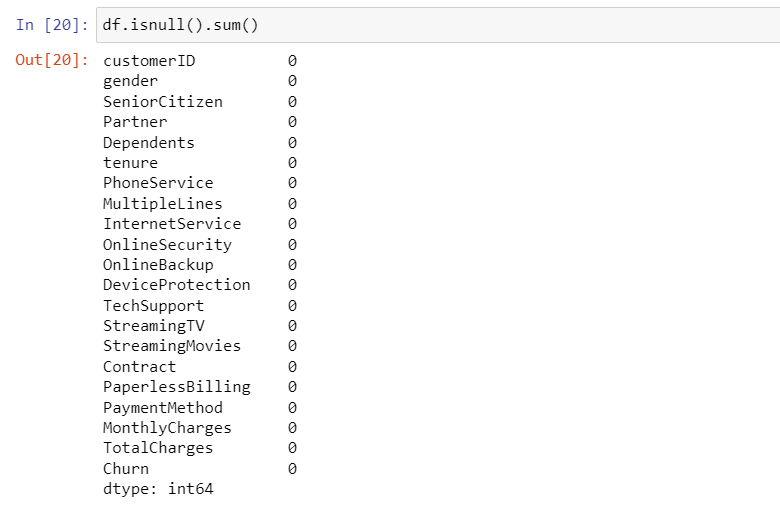


Since we have filled the blank spaces in TotalCharges column with nan values, it's showing 11 null values in that column. Replacing the NAN values using mean method as the column TotalCharges is continuous in nature. Let's handle the column



Here we have checked the mean value of TotalCharges column and replaced the missing values with its mean

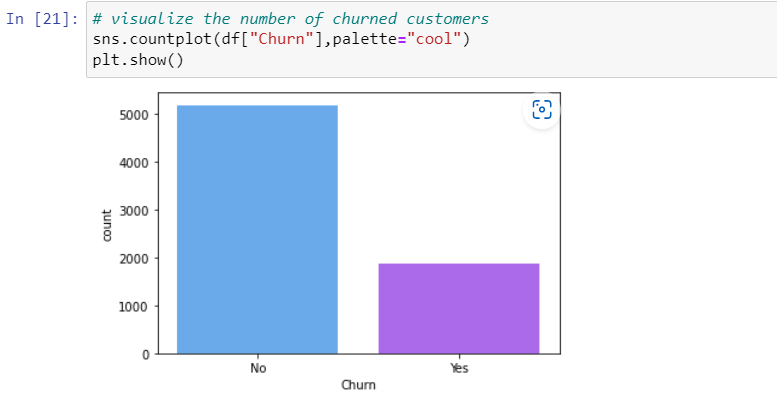
Checking For null values again:



Now Our Data is Free From Null values

## **Data Exploration**

Univarent Analysis:



From the count plot we can observe that the count of "No Churn" are high compared to the count of "Yes Churn". That is there are more number of customers who have not churned . This leads to class imbalance issue in the data, we will rectify it by using Oversampling method in later part.

Plotting features:

We will use for loop to visualize all the columns of the dataset and note down the observations



* From the plot we can observe the total number of male and female customers are almost same, but still the count of male is 3450 which is high compared to count of female which has 3400 counts.
* Here 0 represents senior citizen and 1 represent non senior citizen .The count of senior citizen is higher then non senior citizens. The count of senior citizen is 5901 and non senior citizen is 1142
* The data is almost equal ,yet the customers having partners are relatively higher compared to No.
* The customers who have dependents are very less in counts that means they do not have anyone dependent on them . Here around 70% of customers have dependents and only 29.96% have no dependents.
* The customers who have phone services are large in numbers and who do not own phone services are very less in number
* The customer who have phone services from single line have high counts compared to the customers having phone services from multiple lines, also the customers who do not have phone services have covered very less data compared to others.
* Most of the customers have chosen to get Fiber optic internet followed by DSL, but there are many customers who do not get an internet services.
* It is obvious that the customers who have internet services they needs online security and who do not own any internet services, they do not need any online security . But from the count plot we can observe the majority of customers who have internet services but they do not use any online security.
* From the plot we can see the majority of customers who own internet services they do not have Online backup and the customers who own internet services have very less online backup .
* From the count plot we can notice that the customers without any device protection have high counts as compared to the customers who have some kind of device protection. and the customers who do not have internet access they do not need any device protection.
* The customers who do not need any technical support are high in counts compared to the customers who need technical support . Around 49% of the people do not need any technical support and only 29% needs.
* The customers who do not use streaming TV have little bit high in numbers than the customers who do use Streaming TV and the customers who do not own internet they do not have this service much.
* Most of the customers prefer Month to Month contract compared to 1 year and 2 year contract
* Most of the customers prefer paperless billing and average number of customers who do not prefer paper less billing they may like to receive paper billing.
* Most of the customers prefer Electronic check payment method and the customers who prefer Mailed check, bank transfer and Credit card have average in count.

Splitting Numerical Columns and Plotting:



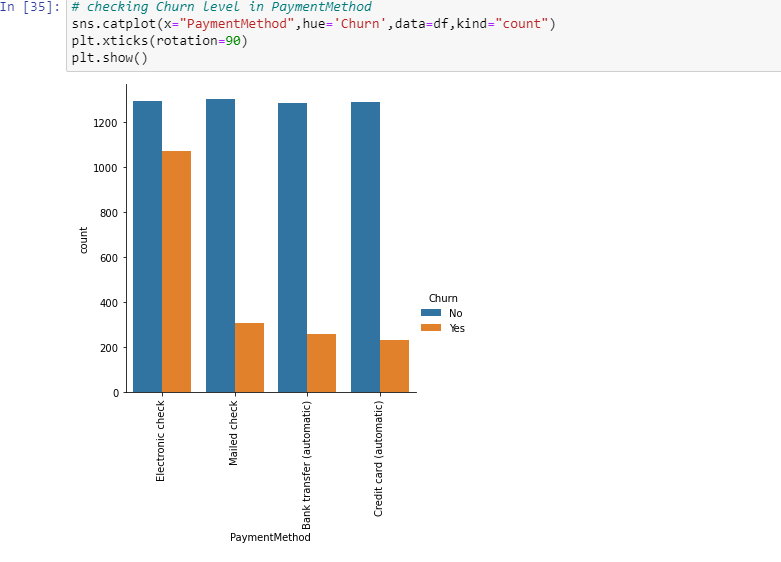
From the above distribution plots we can notice that the data almost looks normal in all the columns except SeniorCitizen and the data in the column TotalCharges is skewed to right Other two columns tenure and MonthlyCharges do not have skewness.

***Bivariant Analysis:***



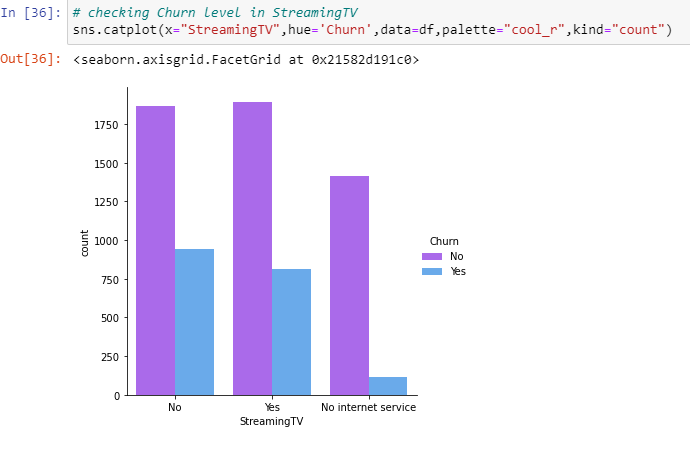
* In the first plot we can see there is no significance difference in the genders, both the genders have equal churn level
* In the second plot we can see the customers without partners have high churn rate compared to the customers with partners.
* The customers who do not have any dependency have high churn rate compared to the customers who have dependents.
* In the last plot we can notice the customers who have phone service have high tendency of getting churned
* The customers who have dependents with high tenure, then the churned level is high 80-110.
* There is no significant difference between the features, here both the feature are in equal length.
* The ratio of churn is high when the customers prefer Fiber optic internet services compared to other services, may be this type of service is bad and need to be focused on and the customers who own DSL service they have very less churn rate.

**Analysis on Payment Mode:**



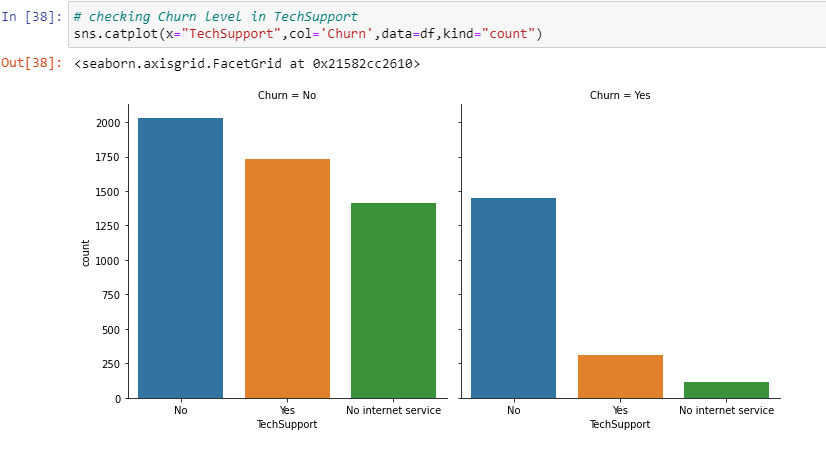
The customers who prefer Electronic check have high churn rate also the customers who existing in the company uses equal payment method.

Analysis on streaming TV:



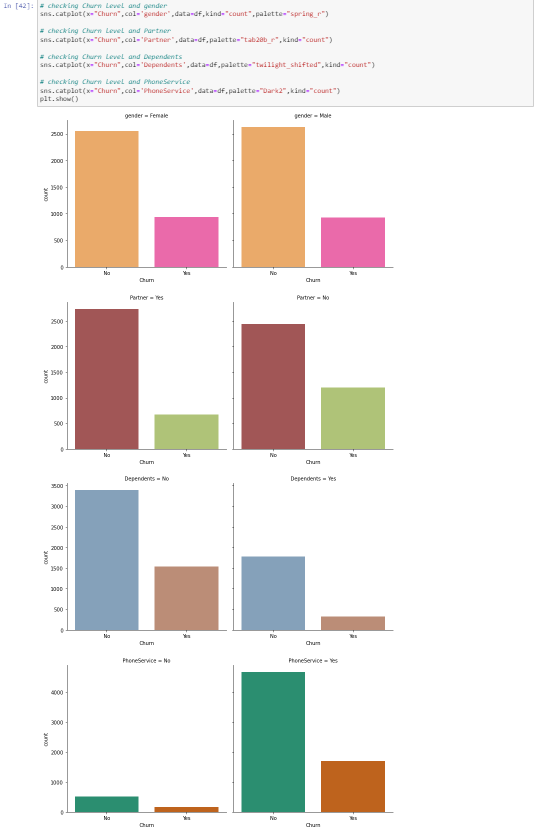
The churn rate is nearly same if the customer own StreamingTV or not.

Analysis On tech Support:



Here we can clearly see that the customers who do not have any techsupport then they have high churn ration

**Analysis of Churn With Gender,Partner,Dependent,PhoneService:**



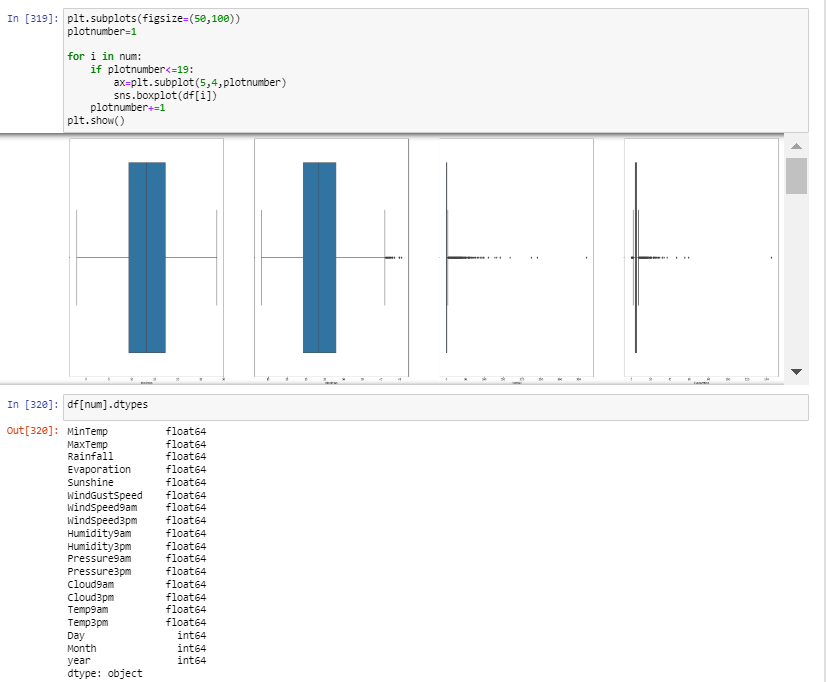
* In the first plot we can see there is no significance difference in the genders, both the genders have equal churn level
* In the second plot we can see the customers without partners have high churn rate compared to the customers with partners.
* The customers who do not have any dependency have high churn rate compared to the customers who have dependents.
* In the last plot we can notice the customers who have phone service have high tendency of getting churned

Multivariant Analysis:



* The pairplot gives the pairwise relation between the features on the basis of the target "Churn" On the diagonal we can notice the distribution plots.
* The features tenure and TotalCharges, MonthlyCharges and TotalCharges have strong linear relation with each other.
* There are no outliers in any of the columns but let's plot box plot to identify the outliers.

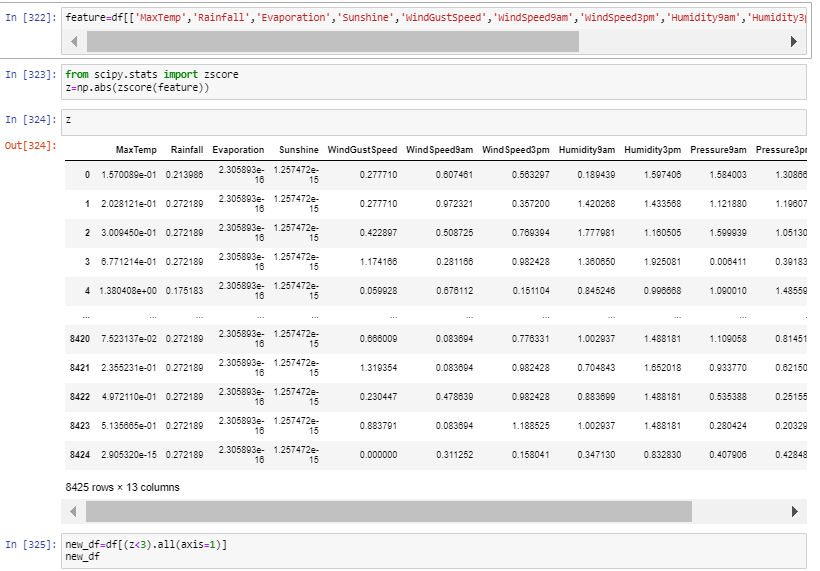
Outliers Handling:



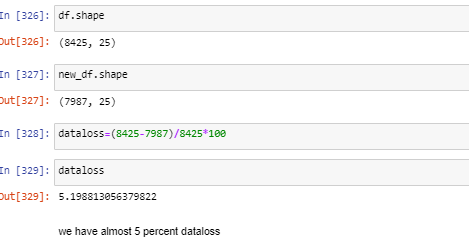
**observation:**

**Presence of outliers in :** Maxtemp,Evopration,Rainfall,windspeed3pm,windspeed9am,windgustspeed,sunshine,humidity9am,pressure9am,pressure3pm,temp3pm,cloud3pm

Removing Outliers Using Zscore and Creating The New Df:

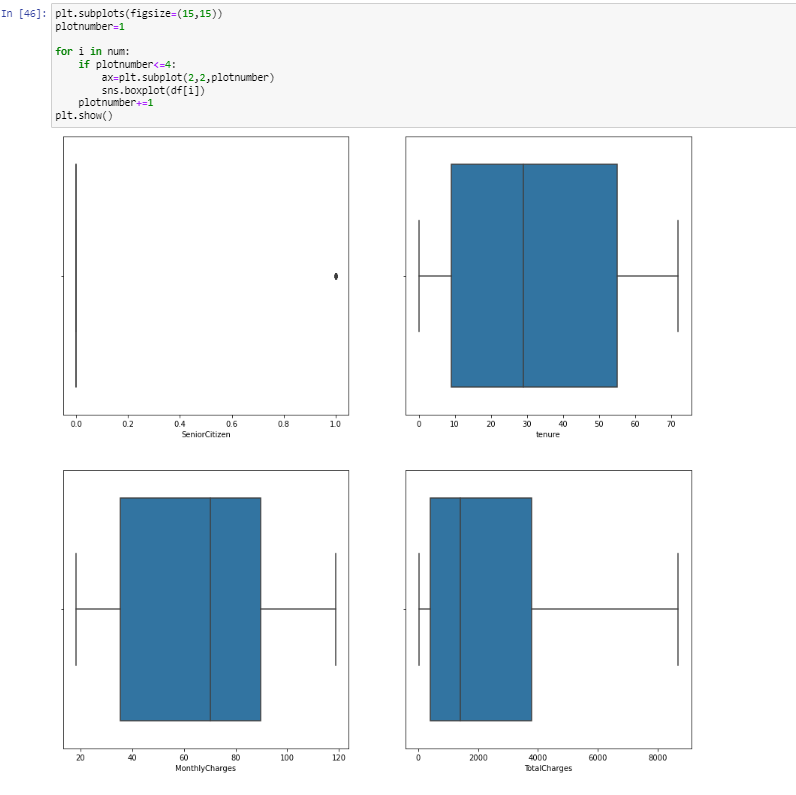


Checking For DataLoss:



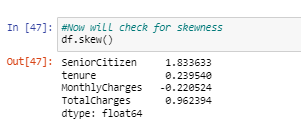
We Have Almost 5% percent Data Loss

Skewness handling:

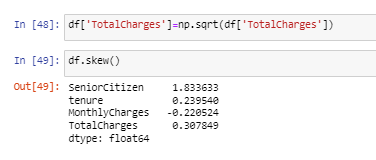


There is no presence of outliers in the Numerical columns

Removing skewness using Square Root Function:



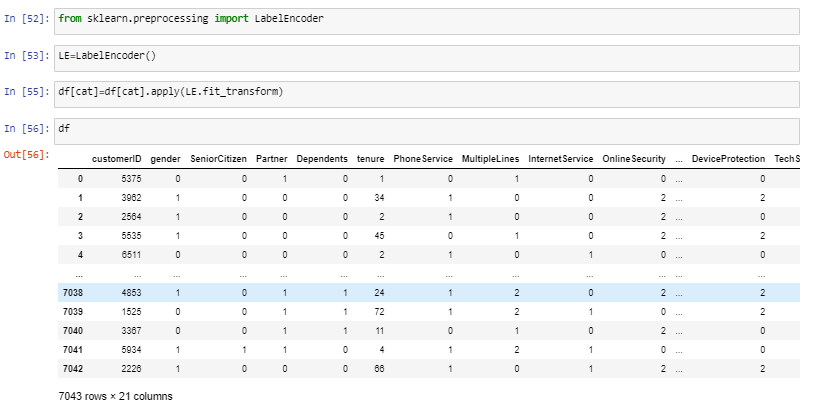
* Here senior citizen and tenure has skewness in data set hence seniorcitizen is catogorical we are not processing that
* We will proceed with Totalcharges for removal of skewness



Now skewness has been removed

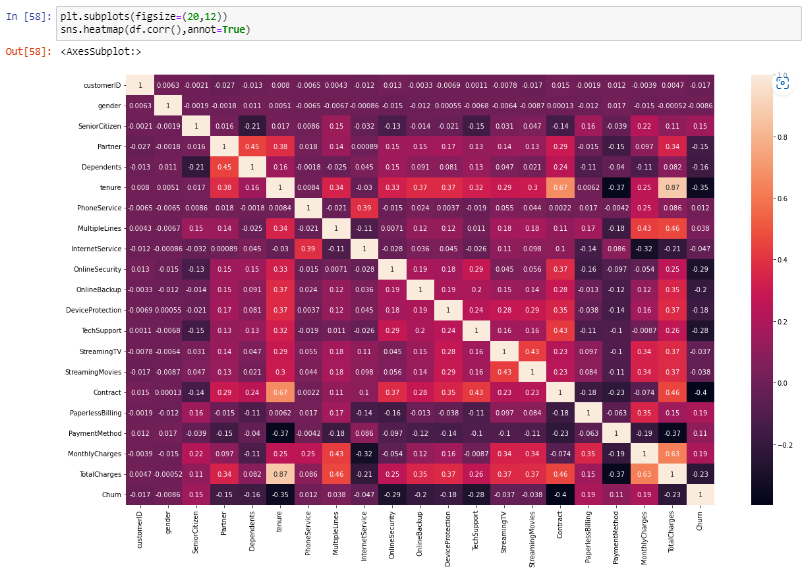
We will proceed further for enoing our categorical columns

Encoding:

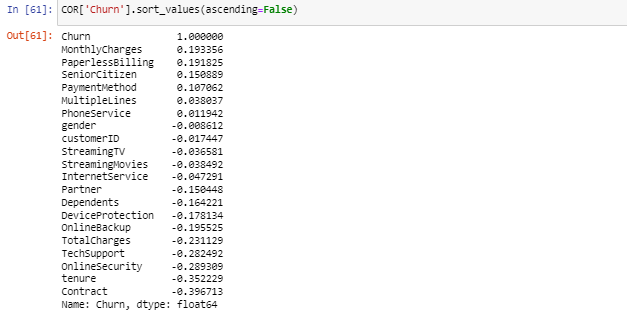


Now Our data is encoded with Numerical values.

# Correlation between the target variable and independent variables using HEAT map

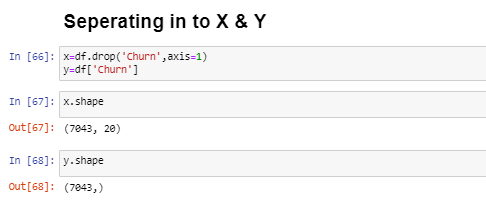


* This heatmap shows the correlation matrix by visualizing the data we can observe the relation between feature to feature and feature to label . this heat map contains both positive and negative correlation.
* There is no much positive correlation between the target and features.
* The columns MonthlyCharges,PaperlessBIlling,SeniorCitizen and PaymentMethod have positive correlation with the label Churn.
* The label is negatively correlated with Contract,tenure,OnlineSecurity,TechSupport,TotalCharges,DeviceProtection, Online Backup, Partner and Dependents.
* Also the column gender has very less correlation with the label we can drop it if necessary.
* The columns TotalCharges and tenure, Contract and tenure, TotalCharges and MonthlyCharges and many other columns have high correlation with each other.
* This leads to multicollinearity issue, to overcome with this problem we will check VIF values and then we will drop the columns having VIF above 10.

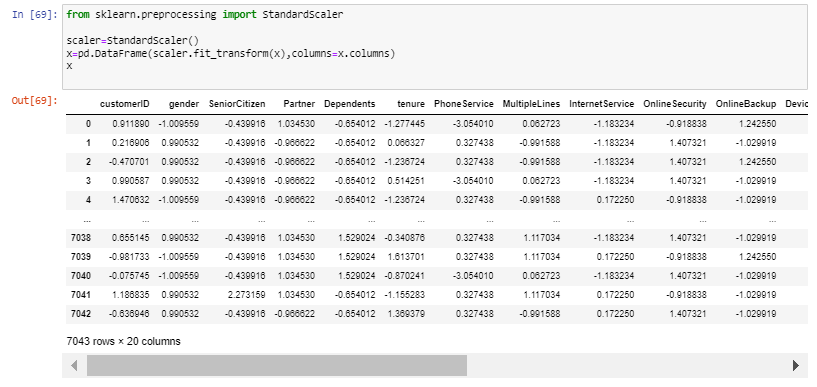


This shows the highly corr Feature with the Target

Splitting Feature And Target As X&Y:

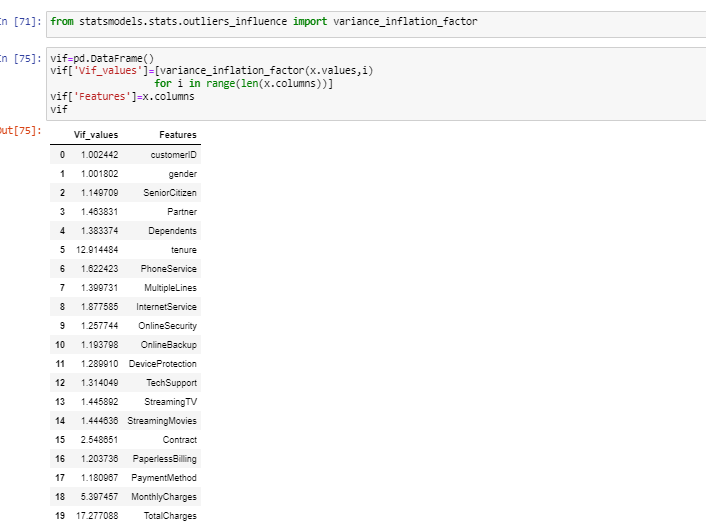


Scaling The Features with Standard Scaler:

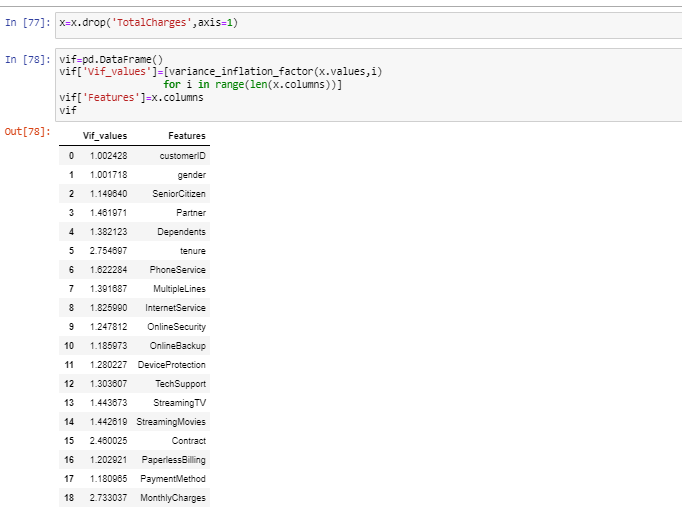


Now our features are scaled .

VIF:  
Now Checking If There is any Multicollinearity problem exist with in Features.

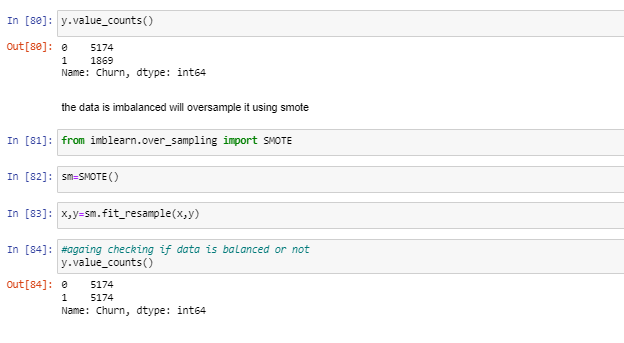


We could Observe that there is a presence of Multicollinearity issue present in Total Charges .I will drop the same and check for vif again



After removing we could observe that the vif has almost reduced and no multicollinearity issue in the features.

Imbalanced Data Set Handling:

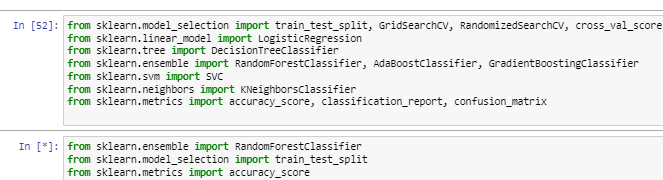


As we could observe initially that the Target data is imbalanced where o is 5174 and 1 is 1869 so with smote im balancing the Data .

Why balancing Target:

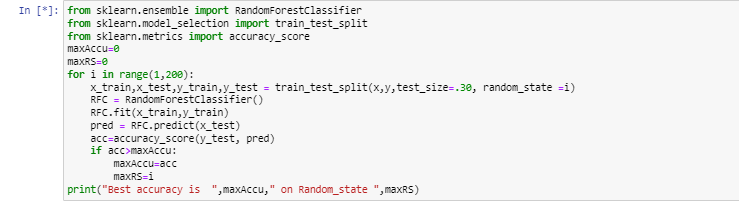
Most models trained on imbalanced data will have a bias towards predicting the larger class (es) and, in many cases, may ignore the smaller class (es) altogether. When a class imbalance exists within the training data, machine learning models will typically**over-classify the larger class (es) due to their increased prior probability**.

Model Building:



Importing necessary Libraries to build a respective model

Selecting Best Random State Splitting The X And Y as Training and Test Data:



Selecting Best random State using the Random forest classifier .Here maxACC is the accuracy accuracy score of the model and maxRS is RandomState



We Got around 87 percentage as model accuracy and Random\_state is 102

Training the model:



Here we could observe that I have create a model and used a for loop to execute each and every model to find the training score ,accuracy score, confusion matrix and classification report

Why Training score:

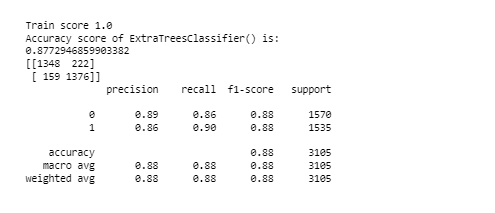
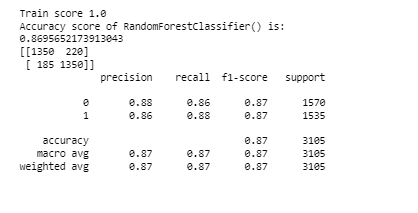
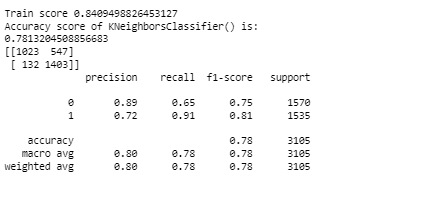
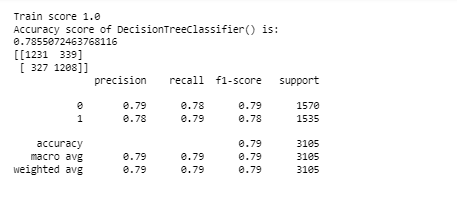
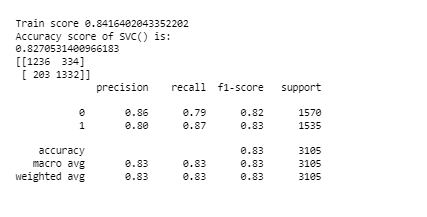
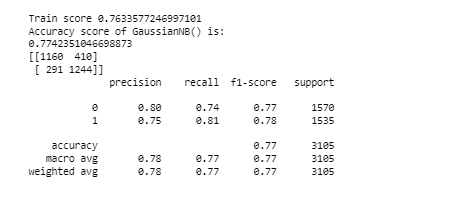
The Training score is important to find the model performance weather its overfitting or underfitting eg:Our model might Well learnt from training data and not perform well on Test data in such a case it may cause overfitting problem so it is important to know Training Score .

Accuracy score :

It is Calculated to measure the performance of model it is measured using the Formula: **(True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives)**.

Model Choosing :

Here we have Trained and Checked The Accuracy Score of the different Models we will check it from Below Images



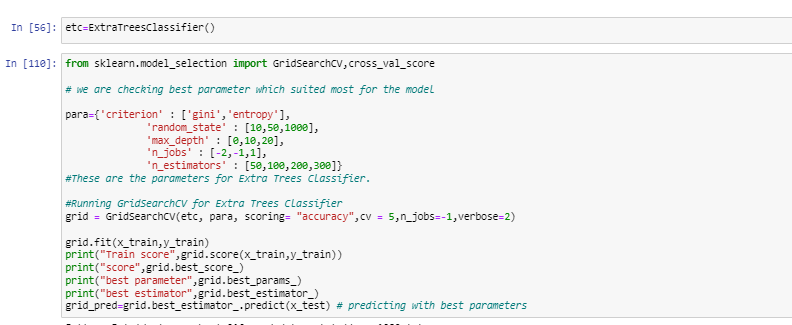
From The above Models I Think that the ExtraTreeClassifier is the best Performing. The Training Data is Overfitted hence it can be handled by parameter tuning also

The ExtraTreeClassifier gives the best score of 87.72 percentage .

Hyper Parameter Tuning:

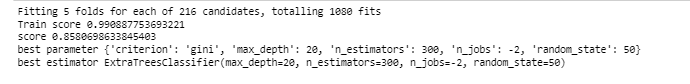
In general the model accuracy can be increased by tuning it with a proper parameter, where we will use GridSearchCV for Parameter Tuning .

Here I use the basic parameters which is acceptable for all the classification based algorithms

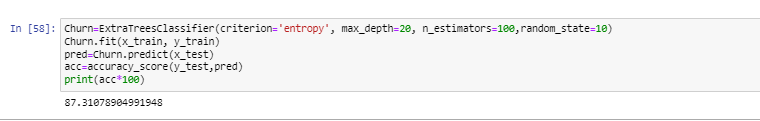


Here We are giving the minimal parameters where the GridsearchCV select the best parameters to train our model and we will fine tune it with the same parameters and predict the same

OUTPUT:



Now ,we almost got the best parameters we will fine tune with the same .

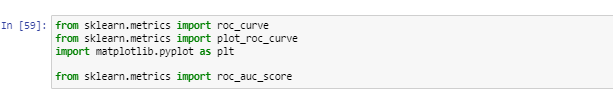


We could observe that The accuracy Score After Hyperparameter Tuning is 87.310 which means our model can predict 87% accuracy .

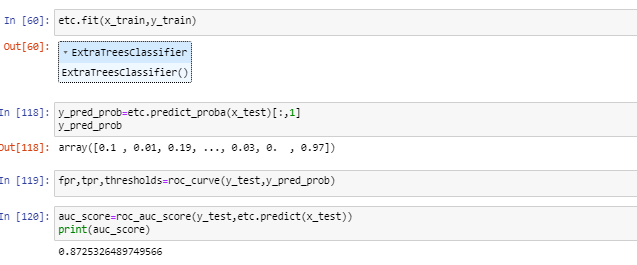
Now we will understand more With Roc\_Auc curve

ROC\_AUC:

Importing Necessary Libraries:



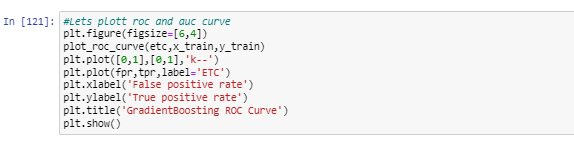
Here I have predicted the class probability also find the false positive rate ,true positive rate ,Threshold



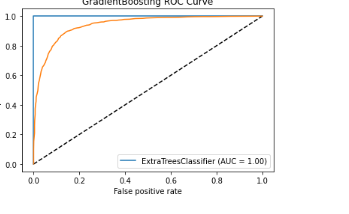
The roc\_auc score is 87.25%

Plotting roc\_auc curve:

Importing Necessary Libraries

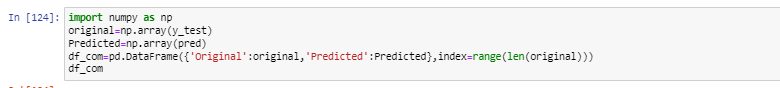


Plotting ROC\_AUC CURVE

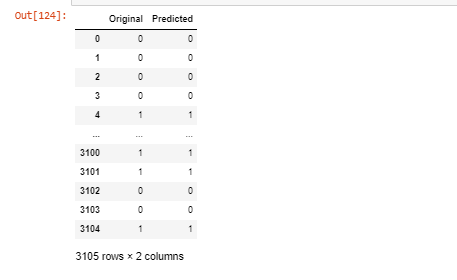


Original vs predicted Output:

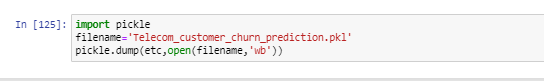
Input:



Output:



Saving The Model As Pickle(.pkl)



**Conclusion:**

From The above model we could observe that our predicted output and the original input are almost same

Hence we conclude that oru model is properly trained and tested with a accuracy score of 87%

**About Me:**

Greetings,Myself Arun Joshva Stephenson a Data Science Enthusiast working as a intern In FlipRobo Technologies Along with work, I’ve got an immense interest in the same field, i.e. Data Science, along with its other subsets of Artificial Intelligence such as Machine Learning, and Deep learning.