**Step 1. 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.

Here we will examine customer data and predict whether the existed customer will churn or not by building classification models and applying classification machine learning algorithm.

**Features information extracted from Dataset:**

**Target Column or Target Variable**:

**Churn**: This column has binary labels which says whether the customer will churn or not.

**Independent Variables:**

**1.CustomerID**: Customer Unique ID.

**2.Gender**: Gender of a customer male or a female.

**3.SeniorCitizen**: Does the customer comes under senior citizen category.

**4.Partner**: Any partner of the customer.

**5.Dependents**: Any Dependent of the customer.

**6.PhoneService**: Does the customer avail phone service or not.

**7.MultipleLines**: Does the customer has multiple lines.

**8.InternetService**: The type of customer internet service provider.

**9.OnlineSecurity**: Does the customer has online security feature with internet provider.

**10.OnlineBackup**: Does the customer has online backup benefits features with the provider.

**11.DeviceProtection**: Does the customer avail the service for device protection.

**12.TechSupport**: Does the customer avail the service for tech support.

**13.StreamingTV**: Does the customer avail the service for streaming TV.

**14.StreamingMovies**: Does the customer avail the service for streaming movies.

**15.Tenure**: Duration of time the customer linked with service provider company.

**16.Contract**: Terms & Condition of the contract of the customer.

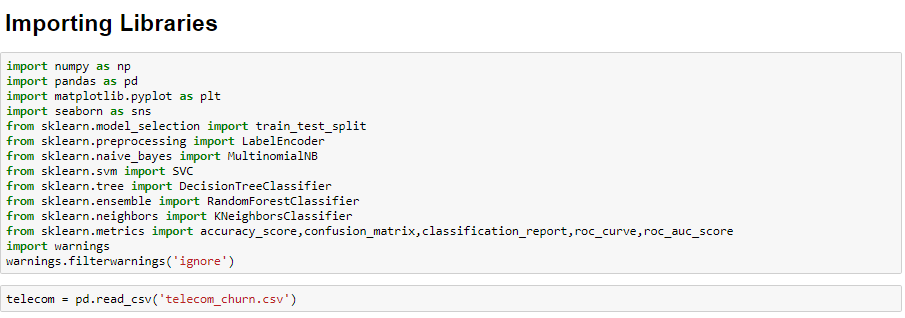
**17.PaperlessBilling**: Does the customer preferred paperless billing.

**18.PaymentMethod**: Preferred payment method used by the customer’s.**19**.**MonthlyCharges**: This feature indicates monthly charge of the customer.

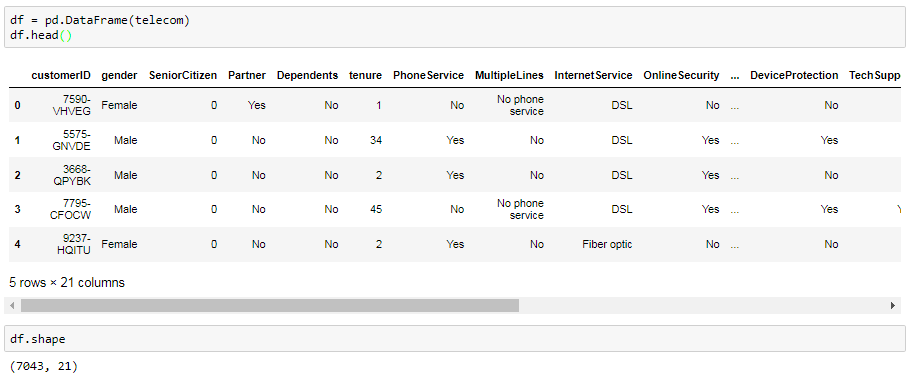
**20.TotalCharges**: This feature indicates the total amount charged to the customer.

**Step 2 Data Analysis:**

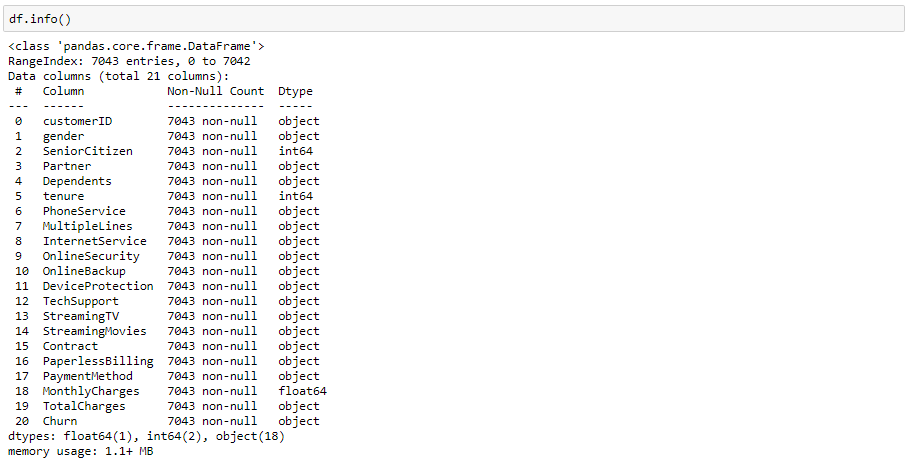
* Importing Neccessary libraries for model preparation -Numpy, Pandas, Matplotlib & Seaborn for plots
* Five Machine Learning algorithm used from sklearn library Logistic,SVC,Random Forest Classifier, Decision tree classifier and KNN
* As the metrics is common for all classification problems so imported common metrics all at once (Confusion Metrics, Accuracy Score, Classification Report) from Sklearn metrics module.



Loading Dataset and Creating dataframe:



* New dataframe for easy analysis.
* The dataset has objects,float and Int Input variables.
* Churn is a binary class target variable.



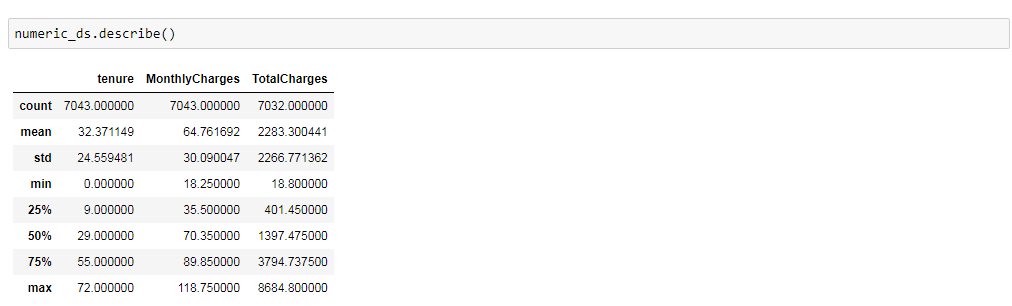
* There are 21 variable with int, float & object data types.
* There are 18 variable with object data type , 2 int data type variables and 1 float data type variable.
* index 0 = customerID & index 24 = Churn column

Dataframe is divided into numeric & objects dataframe for better Graphical Analysis of categorical variables.



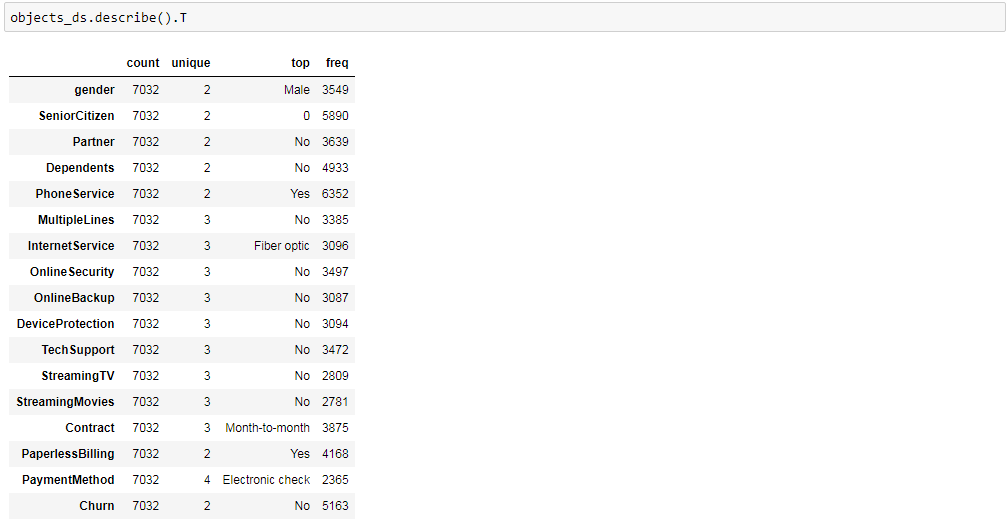
In numeric dataframe there are only three variables where

* Monthly charges shows big diff between max & 75th Percentile.
* There is missing values as the number of row count in TotalCharges is less than the row of 7043 value.
* Total charges also show big difference between max & 75th Percentile.
* Standard deviation is high for Total Charges & Mean is also greater then 50th percentile
* As per above observation it seems data is skewed and spreaded.



**In object dataframe there are 17 variables:**

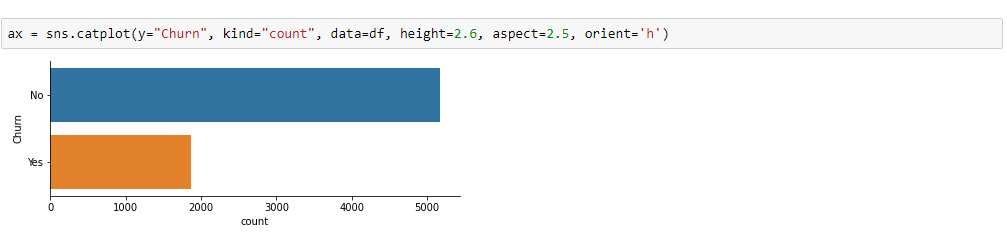
* cutomerID column is dropped which is not so significant for analysis.
* Paymentmethod column has highest 4 unique class where payment mode through cheque is top frequency count.
* Similar observation can be drawn with other object variables.

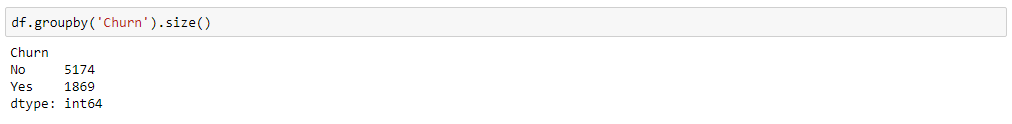


**Graphical Analysis:**

**Target Variable:**

We have a slightly unbalanced target: Churn: No - 72.4% Churn: Yes - 27.6%

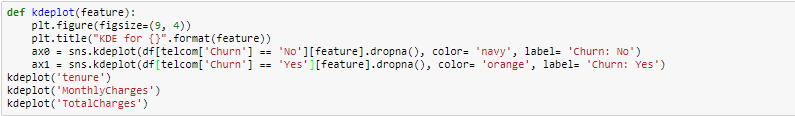


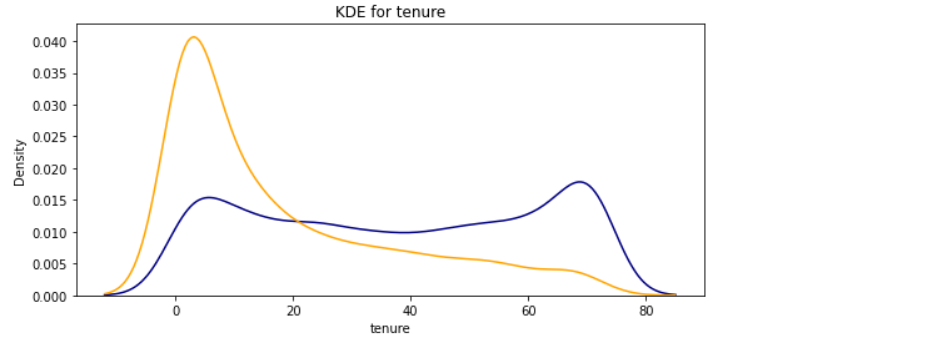


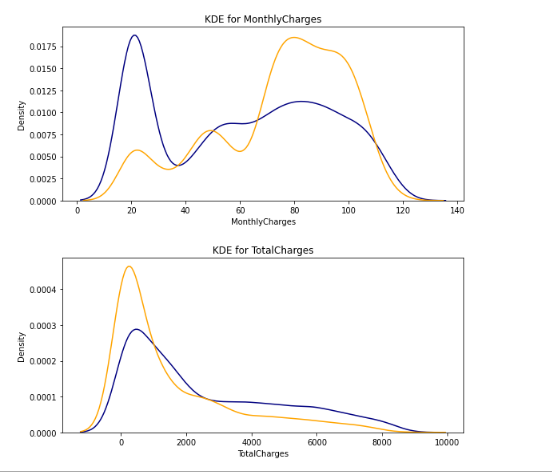
As we can see target column class is not balanced so the same is balanced with upsampling method later below.

**A.Numerical variables:** In this part we will look into our numerical variables, how they are distributed, how they relate with each other and how they can help us to predict the ‘Churn’ variable.

There are only three numerical columns: tenure, monthly charges and total charges. The probability density distribution can be estimate using the **seaborn kdeplot function.**



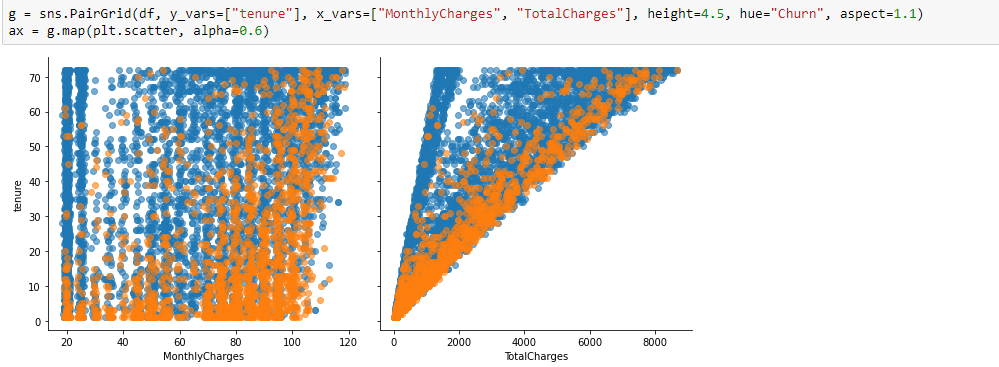




From the plots above we can **conclude** that:

* Recent clients are more likely to churn
* Clients with higher MonthlyCharges are also more likely to churn
* Tenure and MonthlyCharges are probably important features

**Pairgrid Scatter Plot** : We can come to the same conclusions when we use scatterplots below.



**B. Plot for Categorical Variables:**

**Categorical variables.**

This dataset has 16 categorical features:

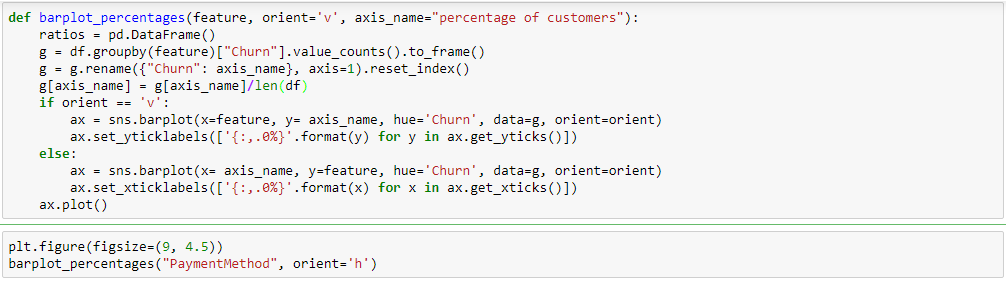
* Six binary features (Yes/No)
* Nine features with three unique values each (categories)
* One feature with four unique values

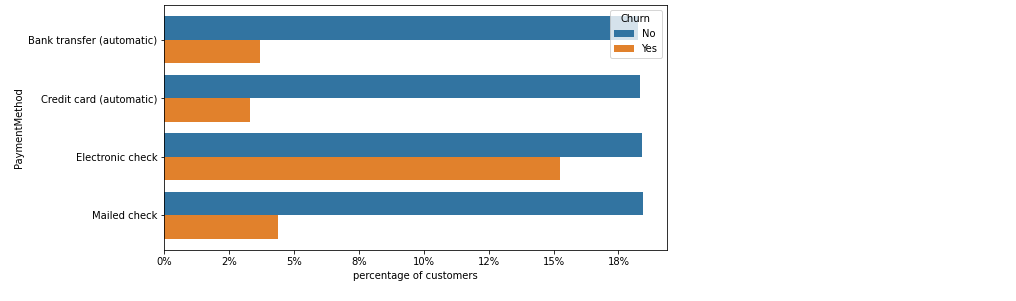
**Contract and Payment Variable Relationship:**

A few **observations**:

* Customers with paperless billing are more probable to churn
* The preferred payment method is Electronic check with around 35% of customers. This method also has a very high churn rate
* Short term contracts have higher churn rates
* One and two year contracts probably have contractual fines and therefore customers have to wait until the end of contract to churn.

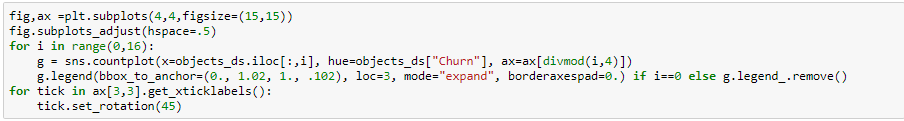
These observations are important for when we design the retention campaigns so that we know where we can focus

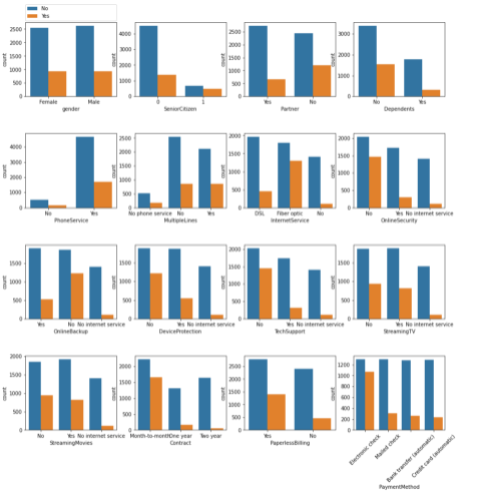




#### Plot for all categorical variables as a bar plot:

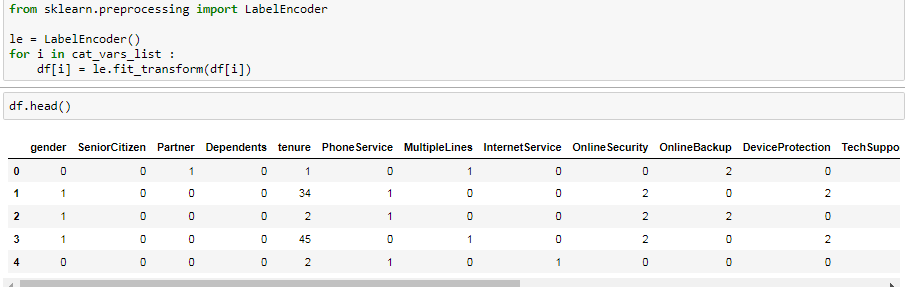
Now we have a better picture of the variables that are more important to us, for example, having Month-to-month contract is a strong indicator if the client might leave soon, so is the Electronic check payment method, being a senior citizen on the other hand is a good predictor but only represents a small amount of the companies clients so you might prefer to focus on the variables that delivers the best results first before tackling it.





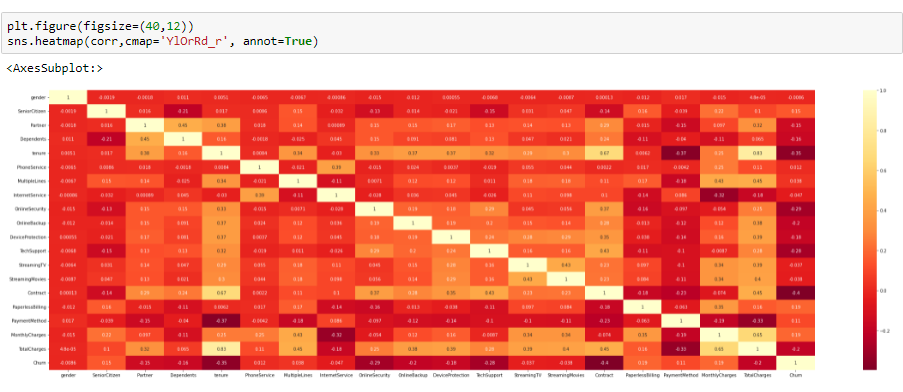
**Labelencoder:**

We have applied label encoder to object or categorical variables which is merged with df dataframe where we can see the encoded categorical variables.



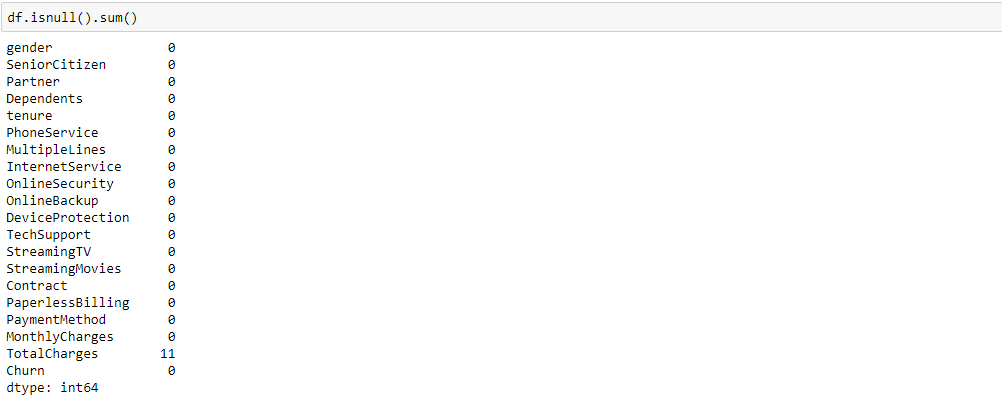
**Heatmap Correlation :**

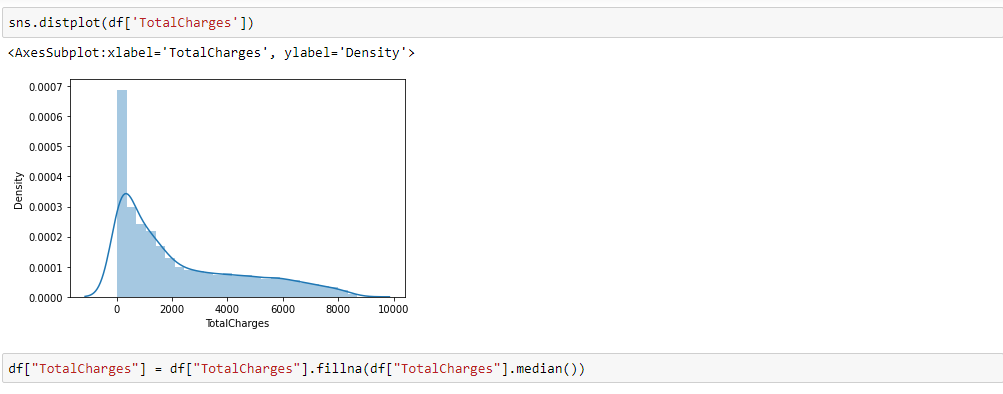
* Churn doesnt show corr with almost any variables.
* Tenure has corr with TotalCharges & contract
* Monthlycharges & totalcharges shows some corr



**Missing Value**:

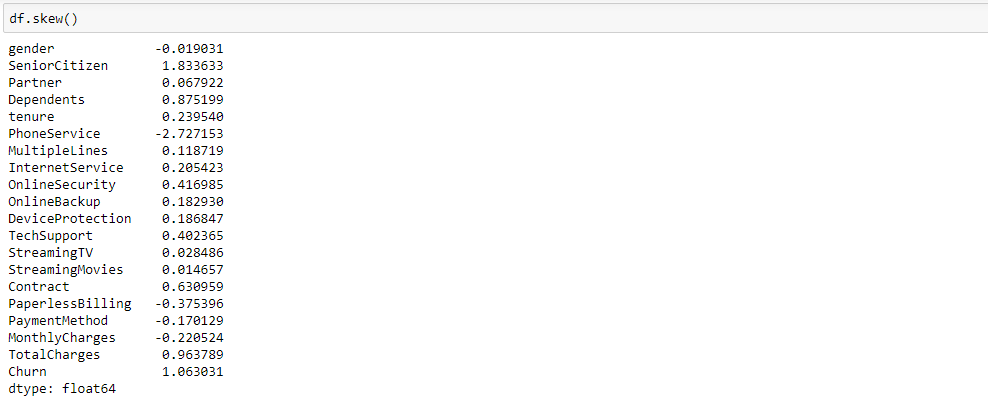
As we can see only TotalCharges column has eleven missing value, which is replaced by median as the data distribution in TotalCharges column is not normal.

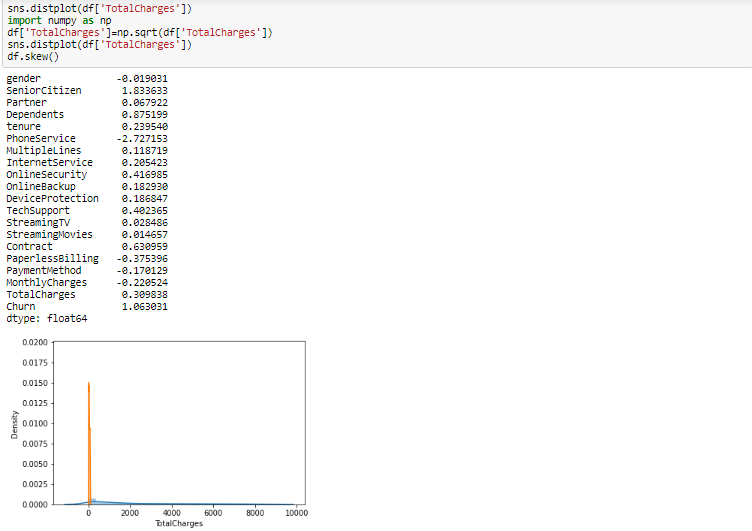




**Removing Skewness** :

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





**Step 3. EDA Concluding Remarks:**

EDA is an important step in churn prediction as it helps us to find the important and hidden pattern in the dataset as we already seen above in graphical analysis.

In churn as well with all other prediction EDA helps us to know the important pattern and relationship of the input features with target variable also help us to know which input features is highly influencing the target variable present in existing dataset**.**

Step 4. **Pre-processing**

**Class Imbalance:**

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

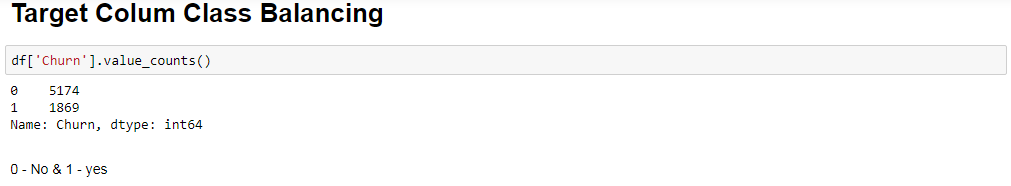
Imbalance Dataset is related with Classification problem having binary class, linear regression there is no issue with imbalance Dataset:

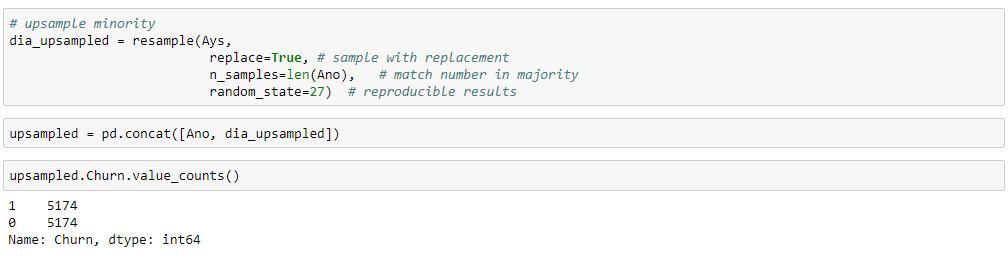
There are two methods to balance the ration of Target Variables:

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

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

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



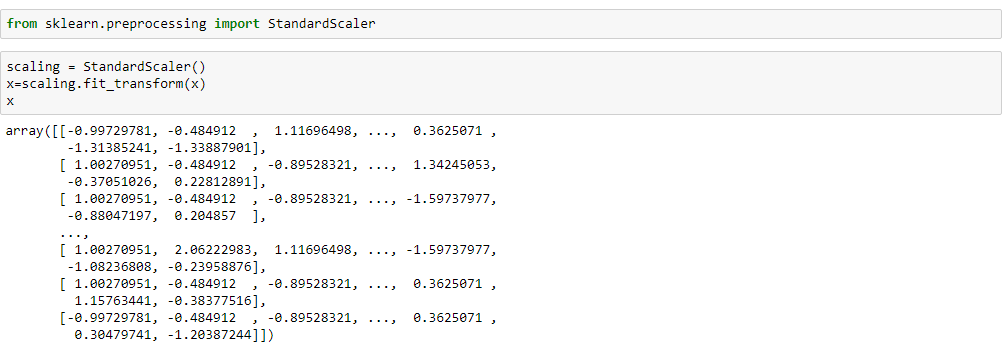


**Scaling:**

Standard Scalar used to build better model, Standard Scalar is a Feature Scaling Technique for the features having Larger number, In order to prevent model to be biased By doing this we can able to build better generalised model.

So when we apply standard scalar we will get the data as normally distributed where mean = 0 & standard deviation = 1.

after label encoding we can see three variables that is monthlycharges, totalcharges and tenure having higher value as compared to other variables so standard scaler is used to scale the values of all variables in similar values.

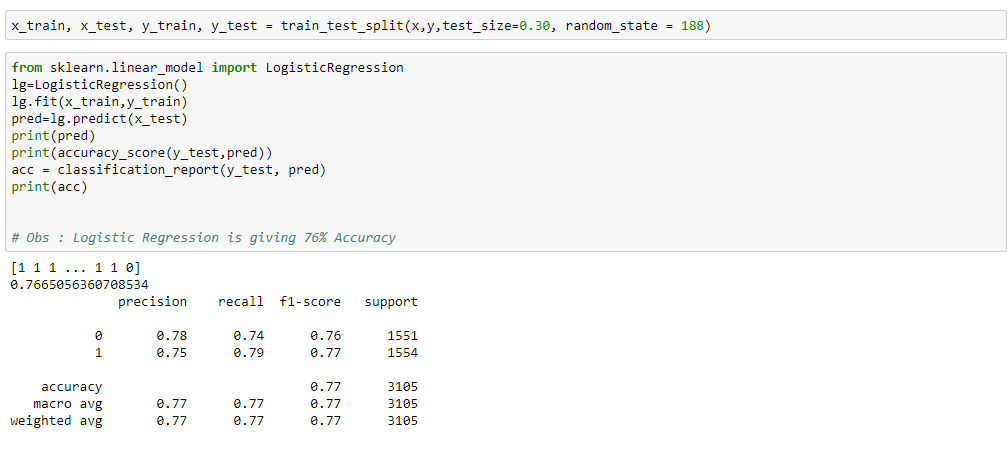


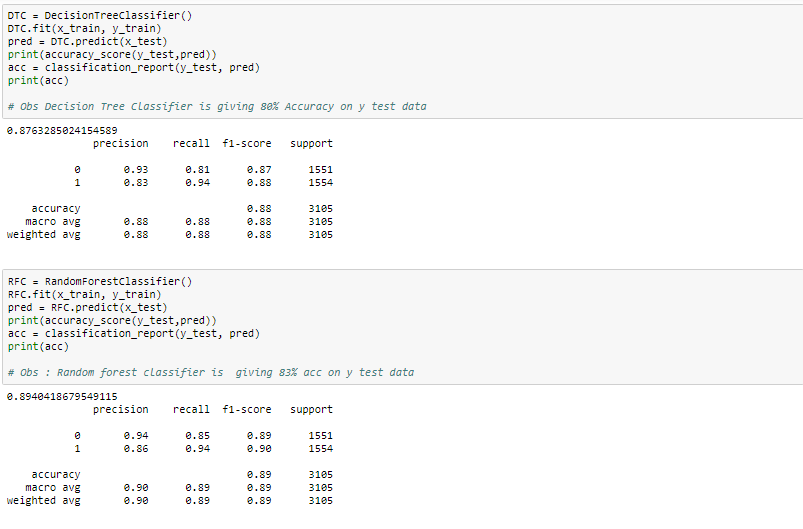
**Step 5. Building Machine Learning Models:**

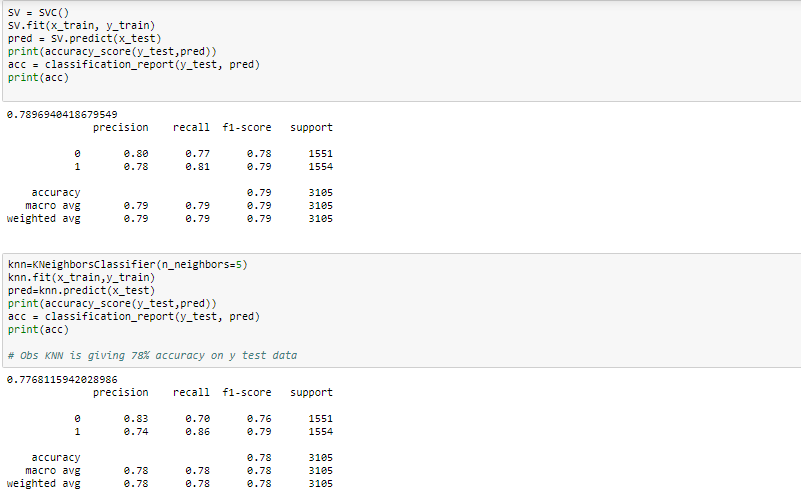
After the independent variables are scaled and target variable class is balanced then the x and y is divided in training and test set where 30% of dataset is used for testing and 70% for training.

Then the following classification machine learning algorithm(logistic, decision tree classifier, random forest, support vector and KNN classifier ) is applied to x train & y train set, then the same algo is applied to x test in order to get the predicted y, after which accuracy and classification report is generated on y\_test and predicted y.

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







**K-Fold Cross Validation:**

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

In K-fold, Model will not be biased as the model has seen almost each data in a dataset with k-fold cross validation where the whole dataset is divided into 'k' sets probably of equal sizes in which the first set is selected as the test set and the rest k-1 sets are used to train the data. Error is calculated for this particular dataset. then the steps are repeated, i.e the second set is selected as the test data, and the remaining k-1 sets are used as the traning data. Again, the error is claculated. similary, the process continues for k times in the end, the CV error is given as the mean of the total erros calculated individually.

The variance in error decreases with the increase in K.

In this example I have selected KNearestNeighbors as the best model because I got highest AUC RUC Score and sharp RUC curve with KNN as compared with RFC & others, also the difference value between KNN Model accuracy with KNN CV score (78-77=1) and RFC has negative difference value of -.72

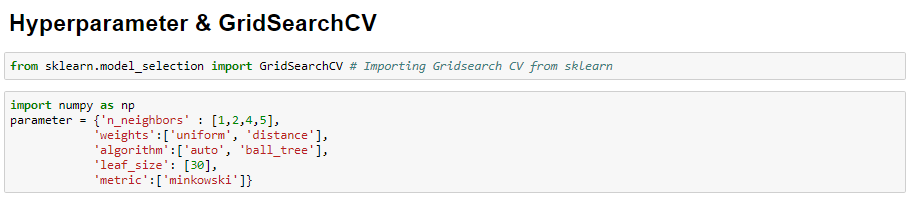


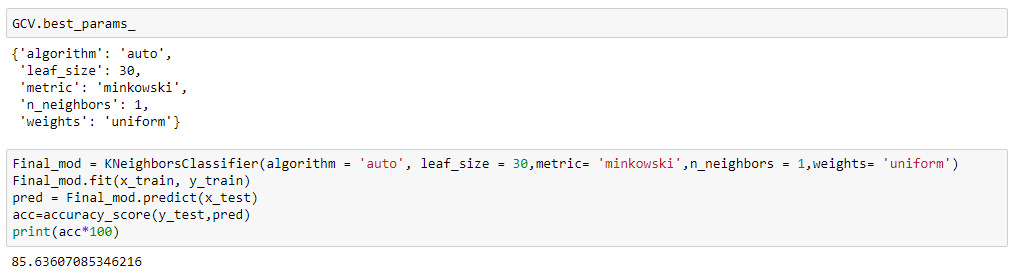
**KNN Hyperparameters & GridSearchCV :**

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

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

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



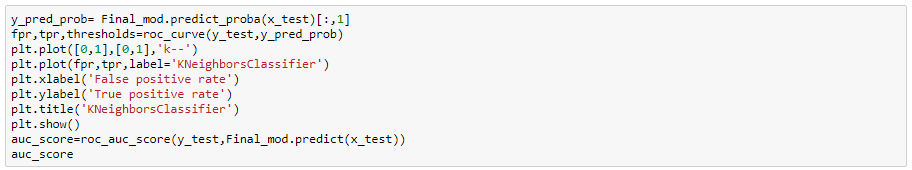


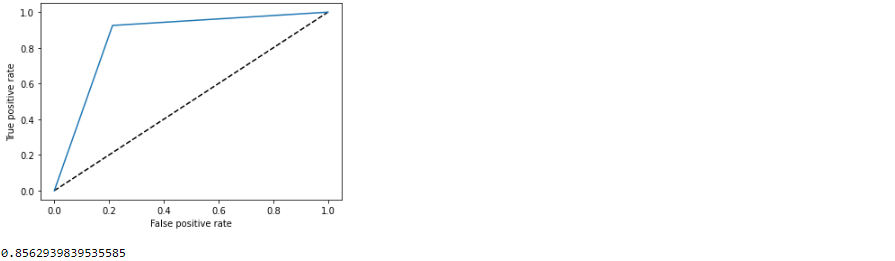
**AUC ROC Curve:**

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

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

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





Step 6. Concluding Remarks:

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