**INTRODUCTION**

Customer churn is one of the most important concerns for large companies. It is a considerable concern in telecom sectors with high competitive services provided by competing industries . On the other hand, predicting the customers who are likely to churn will allow company to minimize losses and save time.

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.

**PROBLEM DEFINITION**

There is an increase of competition in telecom industry where switching from one service to another is so easy for customers.

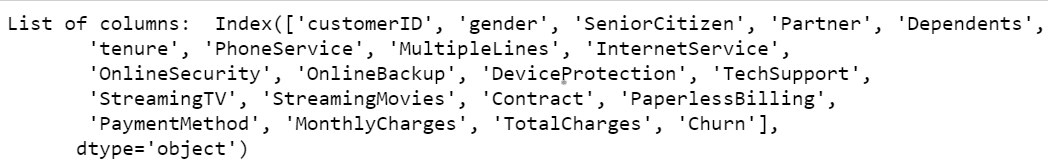
To retain existing customers is a big concern for companies because keeping an existing customer is far less expensive than acquiring a new customer.

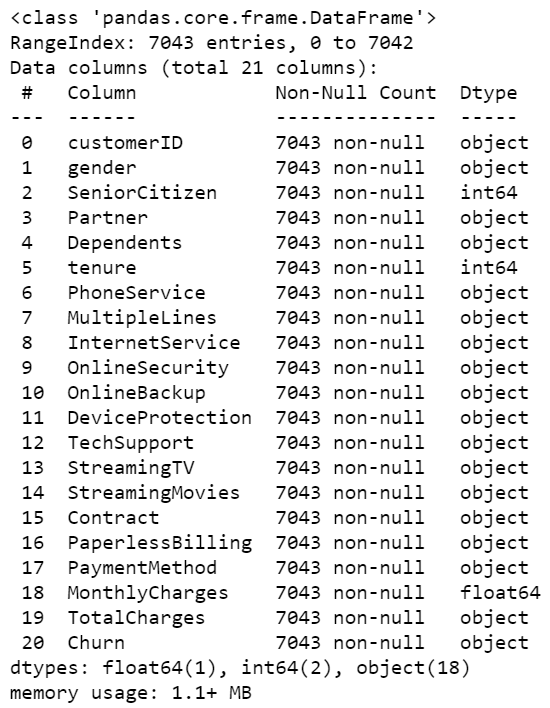
Company has provided dataset of its existing customers and through that dataset we have to analyse and build models to predict who is most vulnerable to churn. So that the company can prioritise focused marketing efforts on that subset of their customer base.

**DATA ANALYSIS**

The dataset provided by the client contains 7043 rows and 21 columns.  
 It has 21 attributes (20 features and 1 target variable).

C:\Users\Nitish\OneDrive\Desktop\churn1\Screenshot 2021-09-02 105511.png



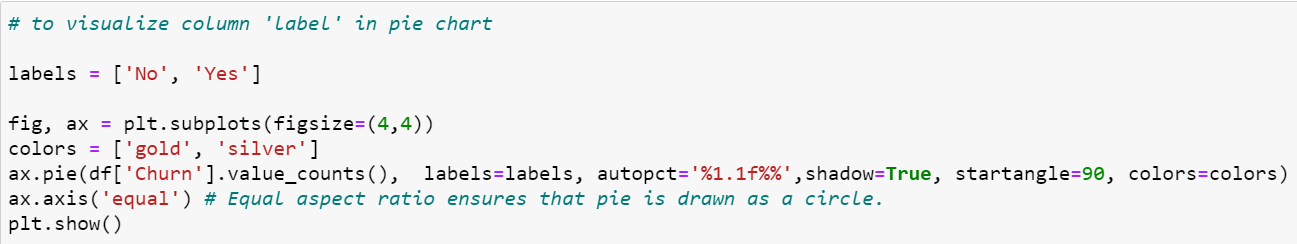
C:\Users\Nitish\OneDrive\Desktop\churn1\Screenshot 2021-09-02 105607.png

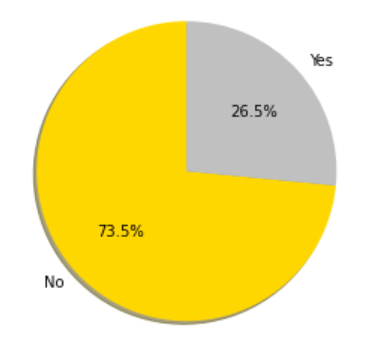
Using df.describe() function to get high understanding of dataset or to get overview/stats of the dataset



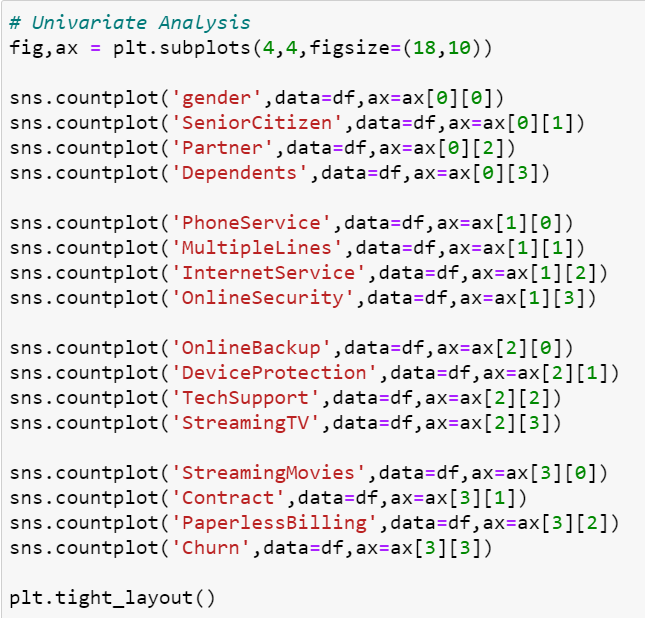
* Count is not same for Total Charges, so null values will be present in the dataset.
* Data is messed up for Total Charges because standard deviation value is close to its mean value.

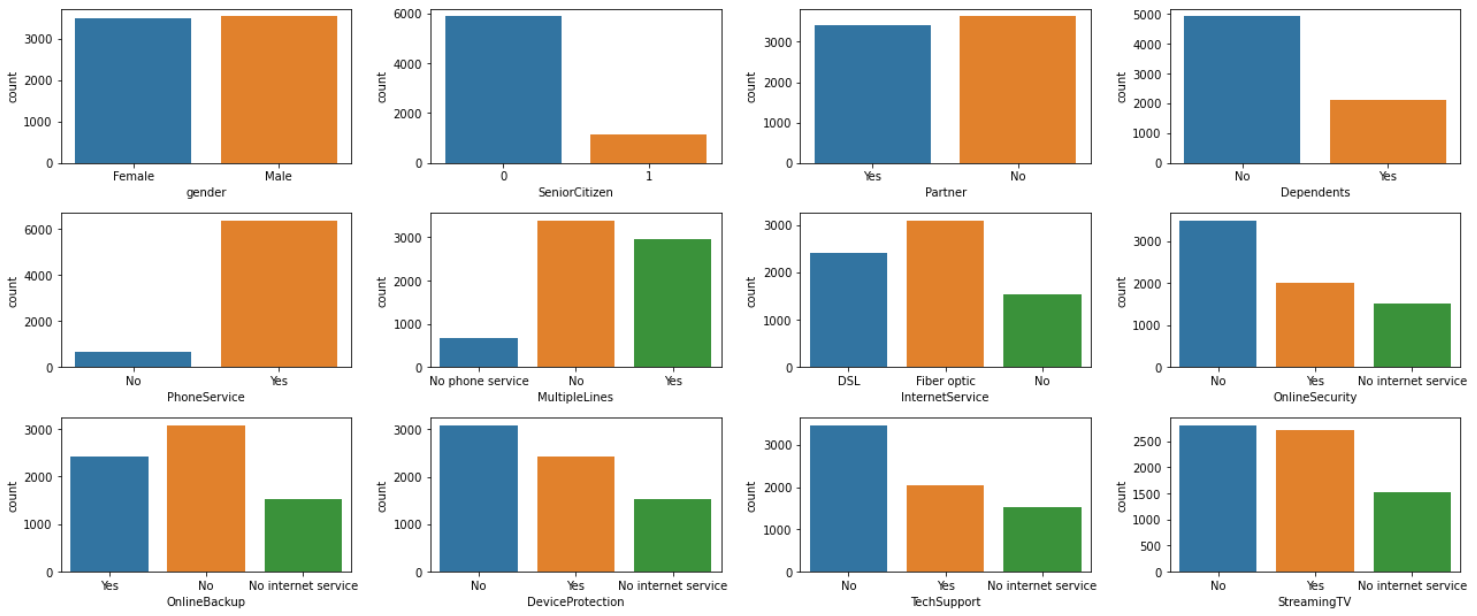
**EXPLORATORY DATA ANALYSIS (EDA)**

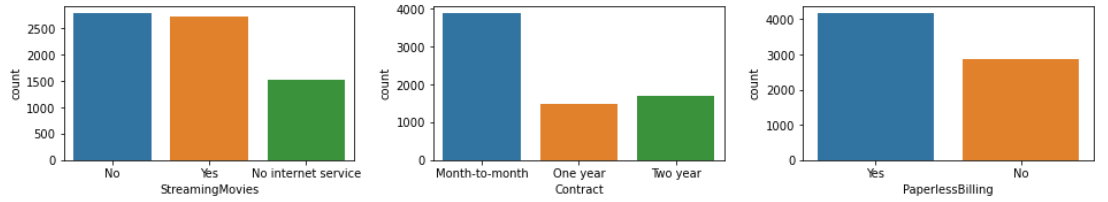
****

It’s a imbalanced dataset where we can see that the customer vulnerable to churn is only 26.5% whereas 73.5% customers did not change their service provider.

The pie chart here depicts the % of churn in telecom industry.

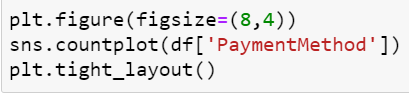


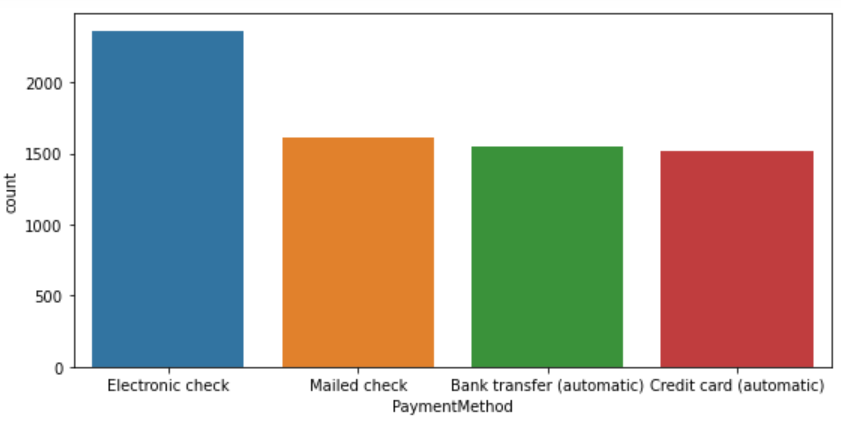
****

****

Inferences from above plot :

* Count of male customers is approximately same to female customers.
* Only 16% customers are there who are senior citizen.
* Customers having partners or not is approximately same.
* Customers having dependents is half to that of not having.
* 90% customers have phone service.
* Customers having multiple line connections is 42%.
* 44% customers have fiber optic connection and 34% have DSL connection for internet.
* Lots of customers online security is at risk as count of no is more than yes.
* Customers not having online backup is more than having backup.
* For security, customers having device protection is 34%.
* Lots of customers does not have tech support.
* Customers streaming and not streaming TV is approximately equal.
* Customers streaming and not streaming movies is approximately equal.
* 55% customers have opted for monthly subscription plan, 24% for one year and 21% for two years.
* Count of customers doing online payments is more.
* Visualizing the Payment Method column





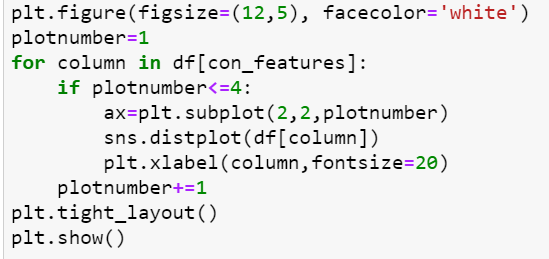
Customers using different modes of payments (in %) :

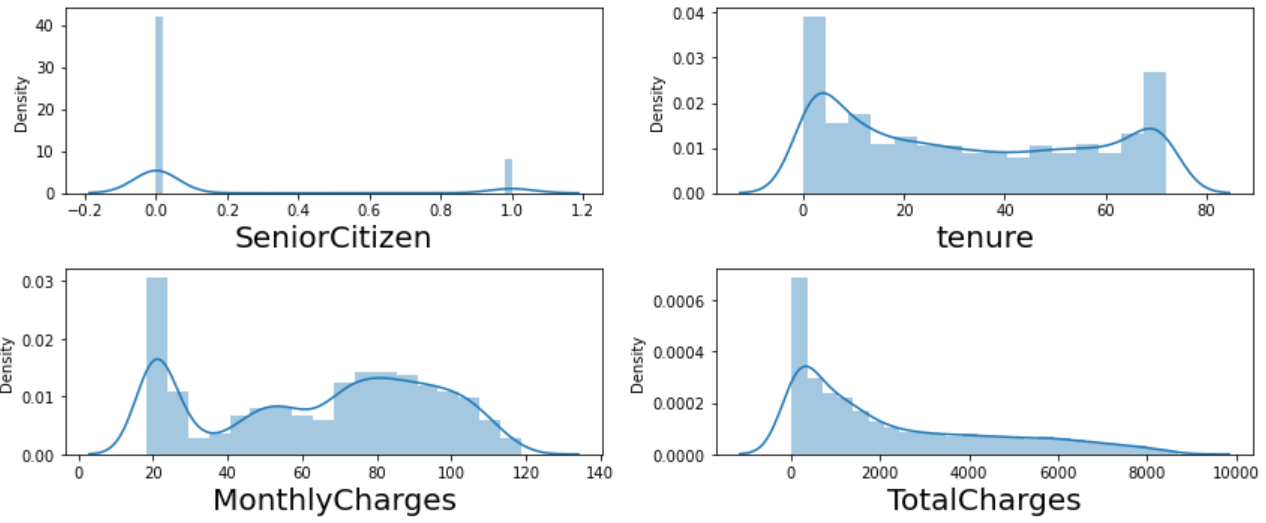
Electronic check - 33%

Mailed check - 23%

Bank transfer (automatic) & Credit card (automatic)- 22% (both)

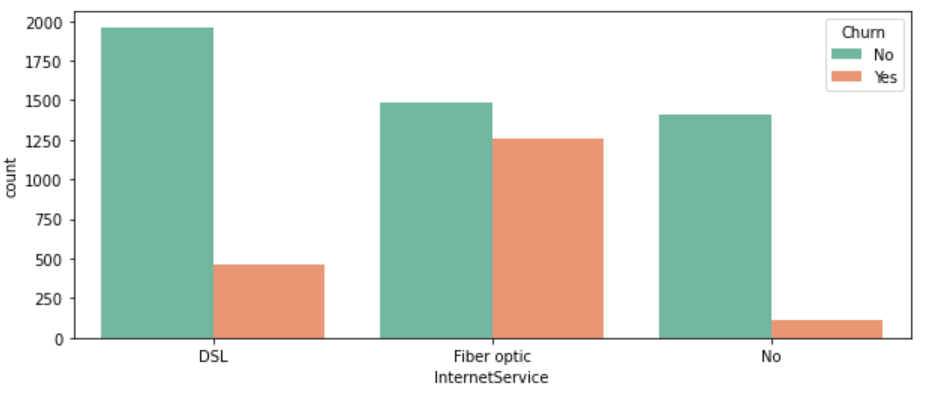
* Lets visualize how data is distributed in every continous column



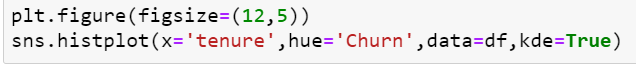


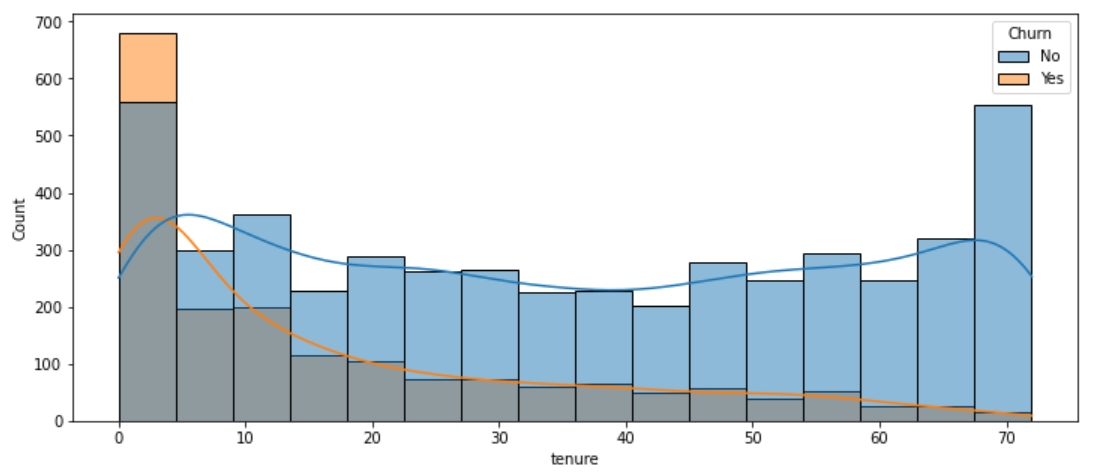
* Skewness can be seen mostly for TotalCharges
* To visualize Internet\_Service vs Churn

C:\Users\Nitish\OneDrive\Desktop\churn1\Screenshot 2021-09-02 110336.png



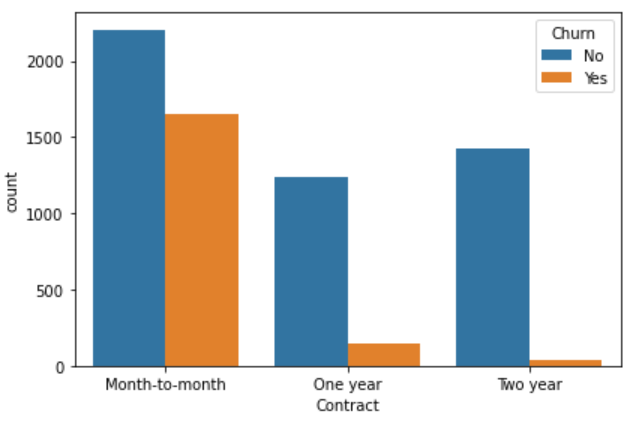
* Lots of customers using fibre optic for internet service has discontinued their service.



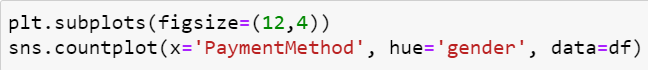


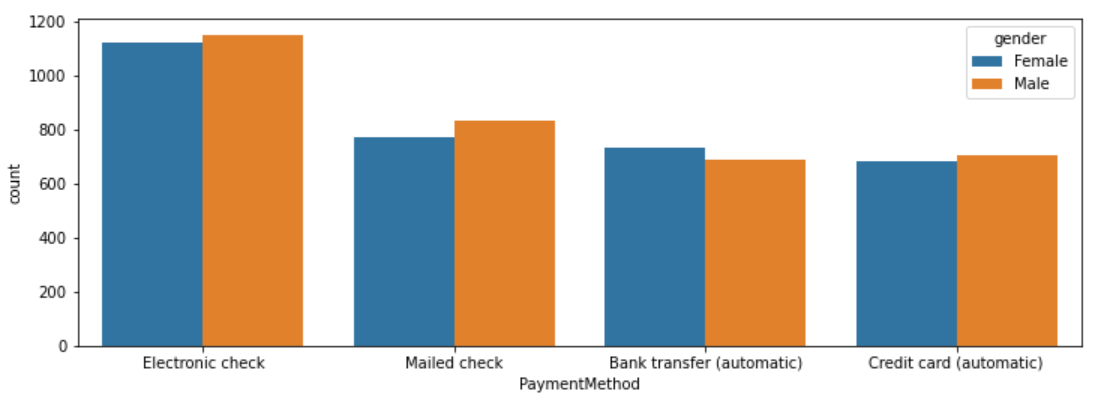
* If customer is using company service for a long period of time, then chances of churn is very less.
* Customer churn can be seen most at the starting month.



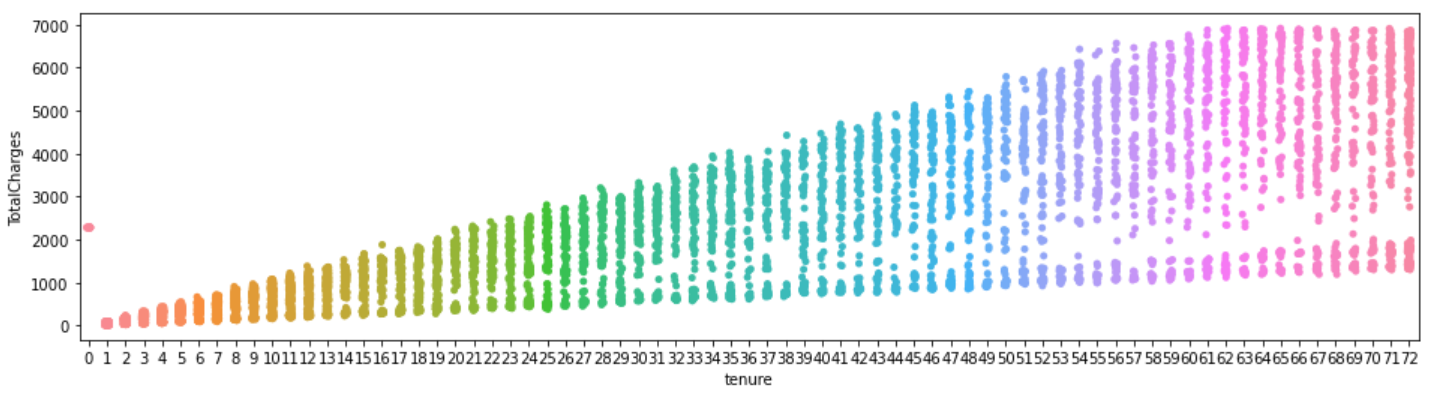


#### Customer churn is very less if subscription plan is opted for one and two years.



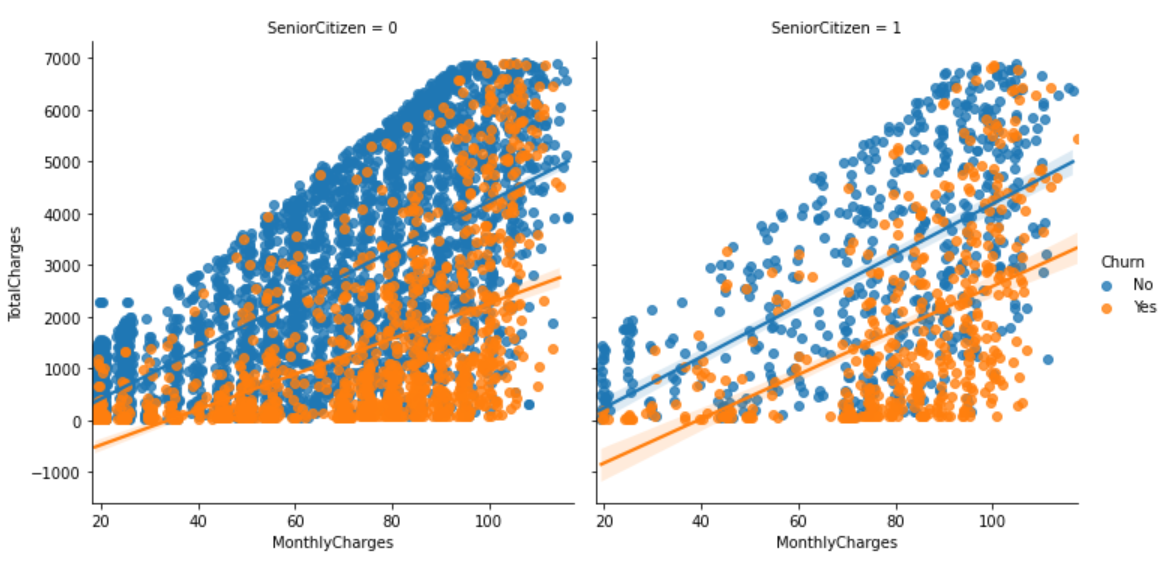


#### For payment, electronic check has been widely used by both gender.

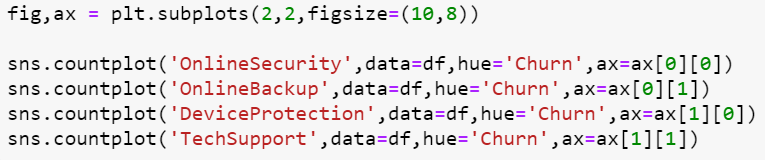


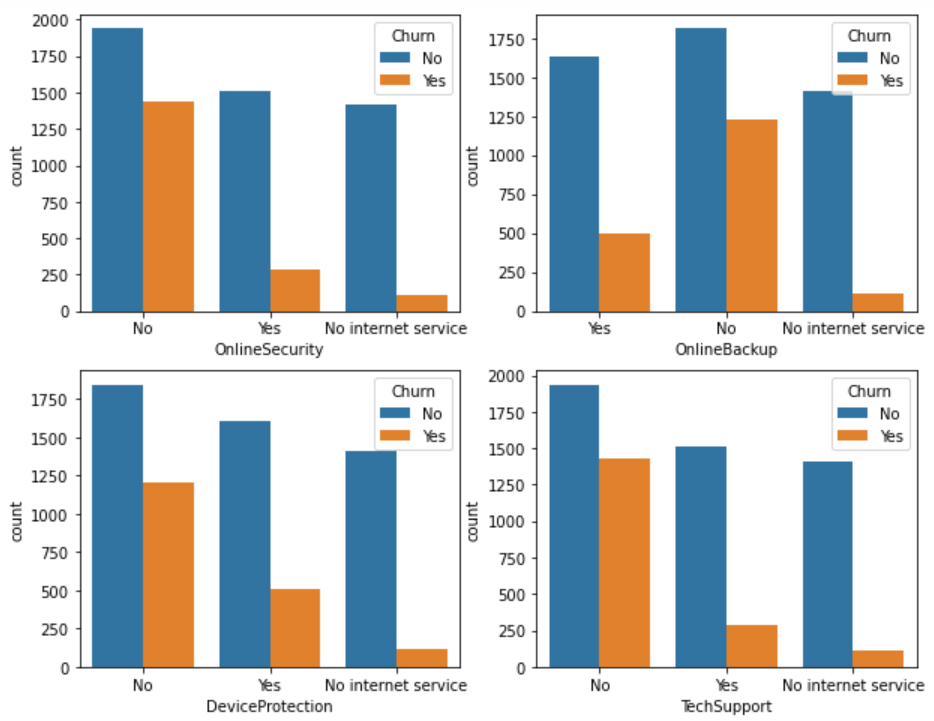
#### If using service for a long period of time, then total charges for services increases.

C:\Users\Nitish\OneDrive\Desktop\churn1\Screenshot 2021-09-02 110655.png



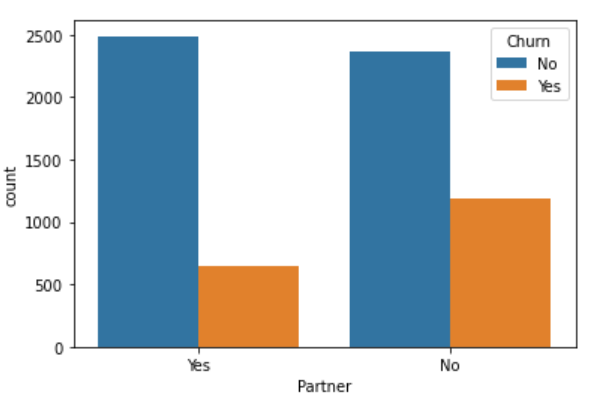
#### In case of senior citizen paying monthly charges for the services, the chances of churn is very less as compared to youth & adults.





* Customers not having online security,online backup,Device protection and Tech Support is most likely to churn.

C:\Users\Nitish\OneDrive\Desktop\churn1\Screenshot 2021-09-02 110742.png

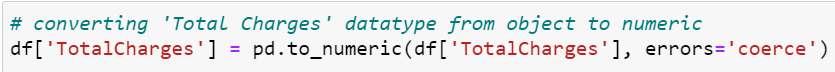


#### Customers having partners are less likely to churn.

**PRE PROCESSING PIPELINE**

* The datatype of Total Charges showing is object which is wrong as it has integer/float values in it, so converting it to numeric.

C:\Users\Nitish\OneDrive\Desktop\churn\Screenshot 2021-09-01 134243.png



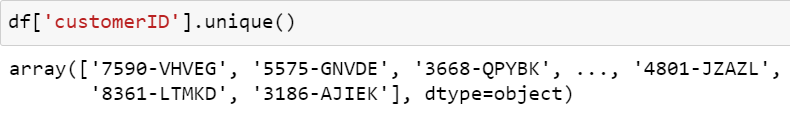
* To separate features into categorical and continuous features for visualization purposes.



* Null values present in the dataset. There are 11 null values present in ‘Total Charges’ column only. So treating the null values by applying mean.



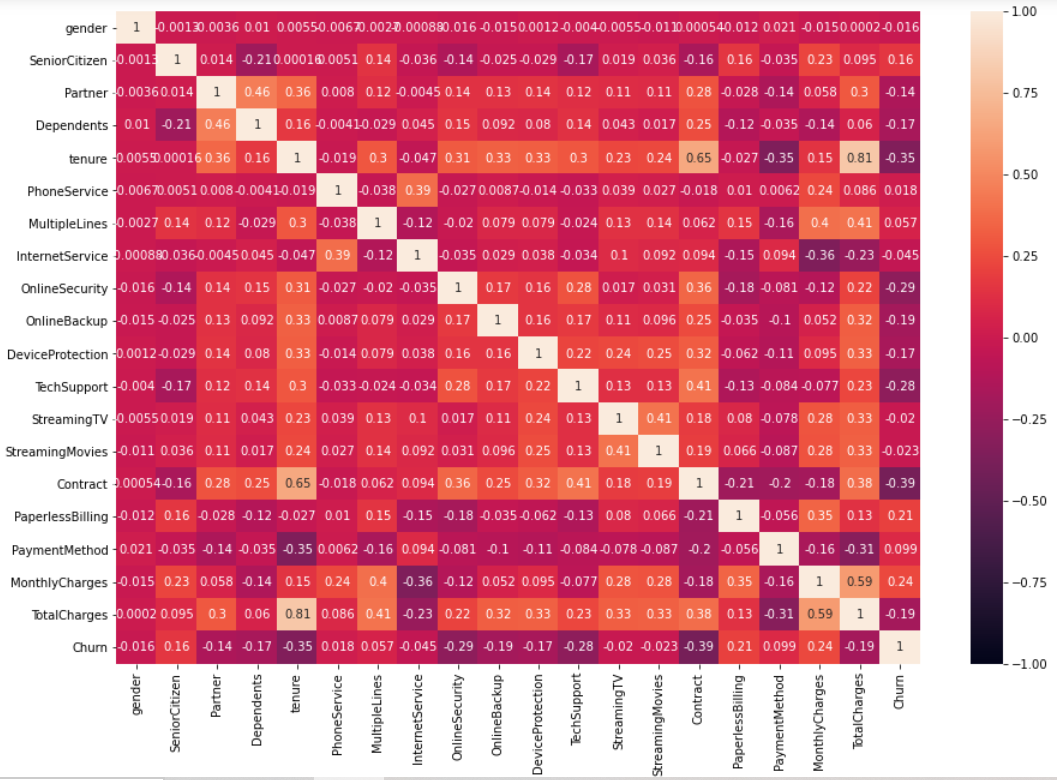
* Dropping ‘CustomerID’ column as it just containing identifier which is not useful for prediction.



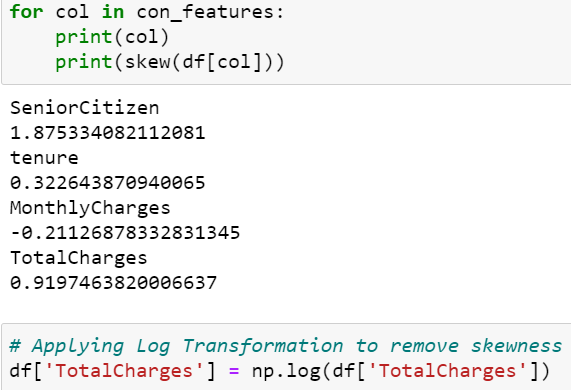
* Using Label Encoder to convert objects into integers/machine language



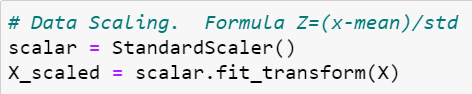
* Plotting heatmap to find correlation between features and target variable.



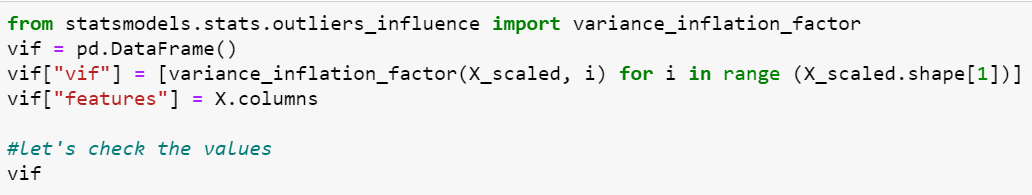
* Only MonthlyCharges and Paperless billing is showing positive relationship with target variable (Churn) rest all features are negatively correlated.
* Then checking for skewness for continous features and treating the skewness by applying log transformation.

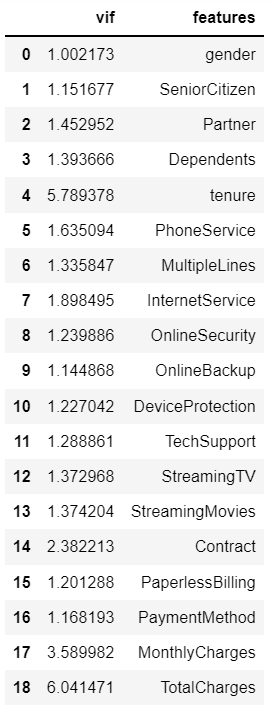


* Splitting target variable (Churn) to variable y and rest all features to variable x.
* Then doing data scaling to normalize the range of independent variables or features of data.

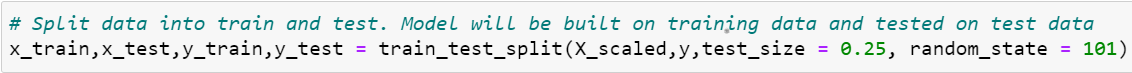


* Using variance inflation factor to test the effect of multiple variables/features on target variable.





* Multicollinearity problem exists here because Total Charges and tenure are inter-correlated with each other. So dropping Total Charges column to remove multicollinearity.
* At last splitting data into train and test where model will be built on training data and tested on test data. Sample Data was divided 25% for test and rest 75% for train for model building.

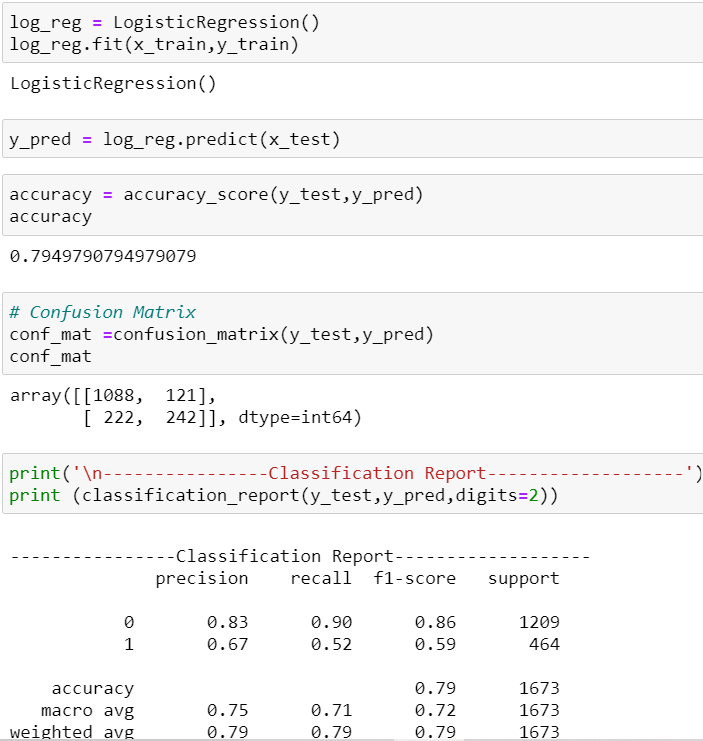


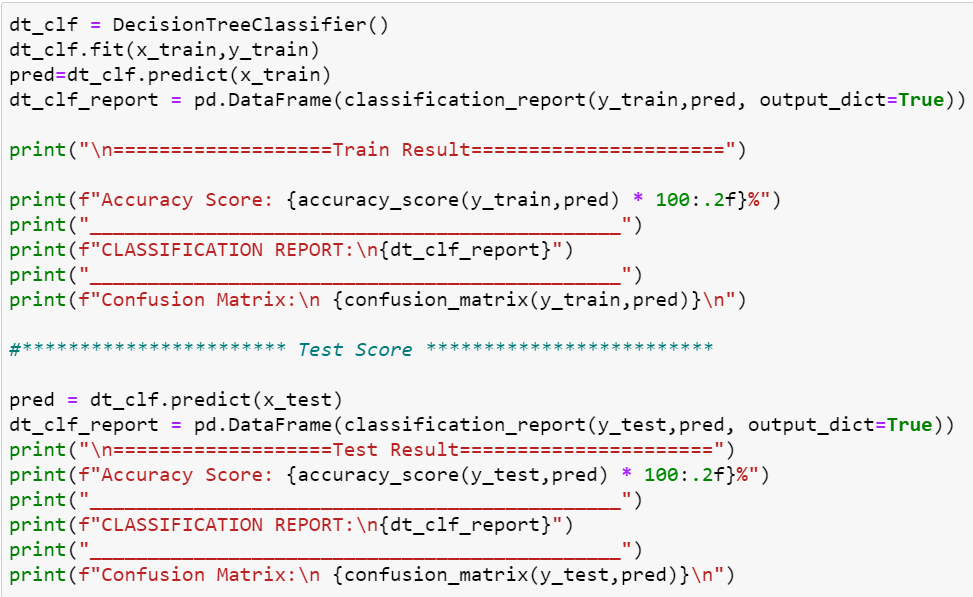
**BUILDING MACHINE LEARNING MODELS**

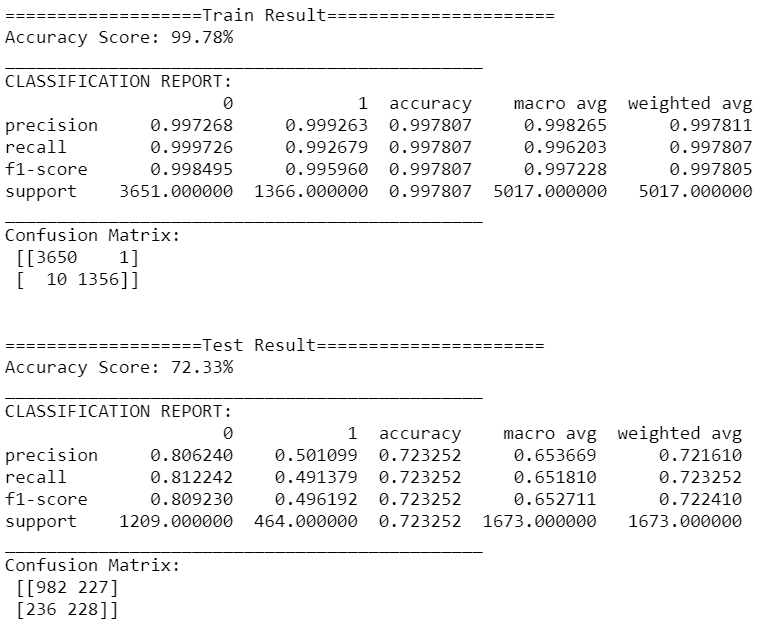
Different algorithms applied for prediction of churn of customers.

* Logistic Regression
* Decision Tree Classifier
* Random Forest Classifier
* K Nearest Neighbors (KNN) Classifier
* Support Vector Classifier
* Gradient Boosting Classifier
* Stochastic Gradient Descent Classifier

LOGISTIC REGRESSION

****

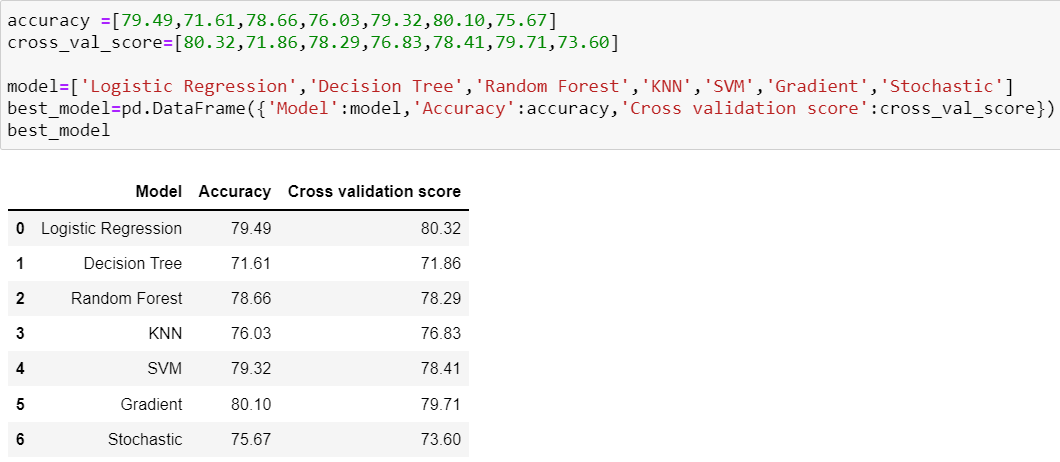
DECISION TREE CLASSIFIER

****

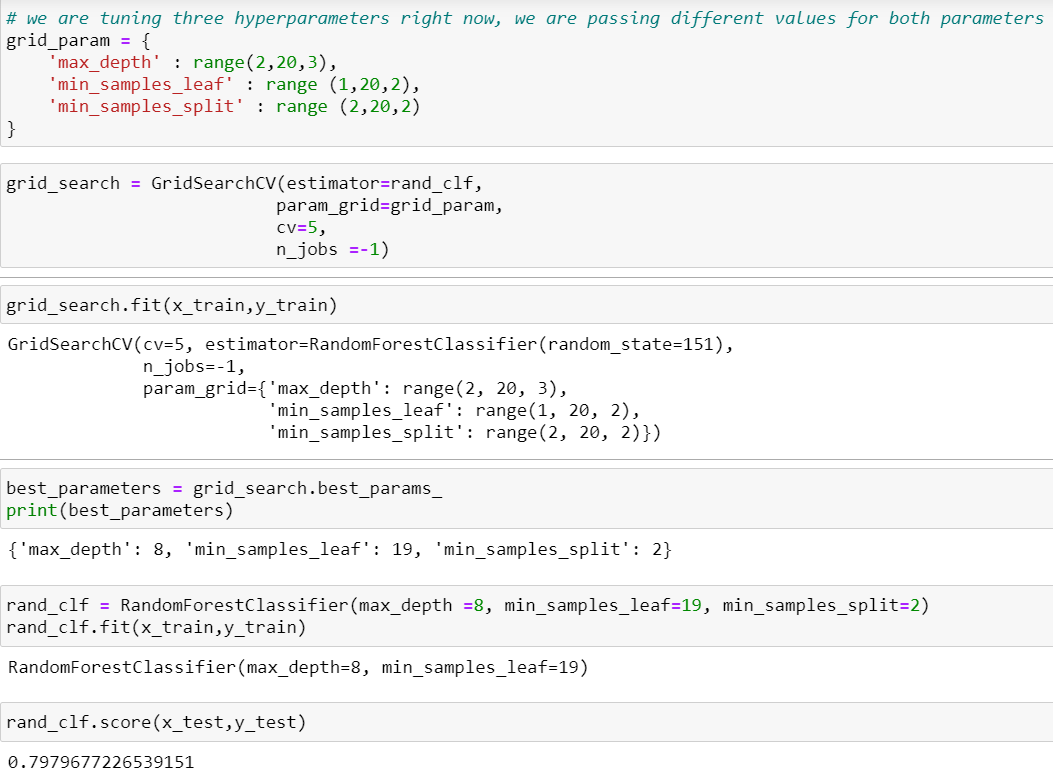
Similarly done for rest all models like Random Forest,

KNN, SVC, Gradient Boosting and Stochastic Gradient Descent.

* Created a dataframe to observe accuracy and CV score for all models in a proper frame.

****

* Random Forest Classifier is our best model because the difference between its accuracy and CV score is least among other models.
* Now we will apply hyper parameter tuning on our best model (Random Forest Classifier) using GridSearchCV.



* After doing hyper-parameter tuning accuracy increasd by 1.13%.

**CONCLUDING REMARKS**

* The dataset is imbalanced with majority of customers not changing their service provider or discontinuing the service.
* Total charges and tenure are inter-correlated with each other. So multicollinearity exists. Dropping total charges will resolve this problem.
* Senior citizen customers is less likely to churn.
* The dataset consists of young people mostly as customers.
* For payment customer prefer online mode most of the time than other methods.
* Customer churn is seen mostly if subscription plan is opted for monthly.
* New customers is more likely to churn who uses services at the starting month.
* Lots of features is negatively correlated with target variable (churn).