Customer Churn Analysis

**Problem Statement**:

Customer churn is when a company’s customers stop doing business with that company. Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales budgets to gain additional customers. Existing customers will often have a higher volume of service consumption and can generate additional customer referrals.

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

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

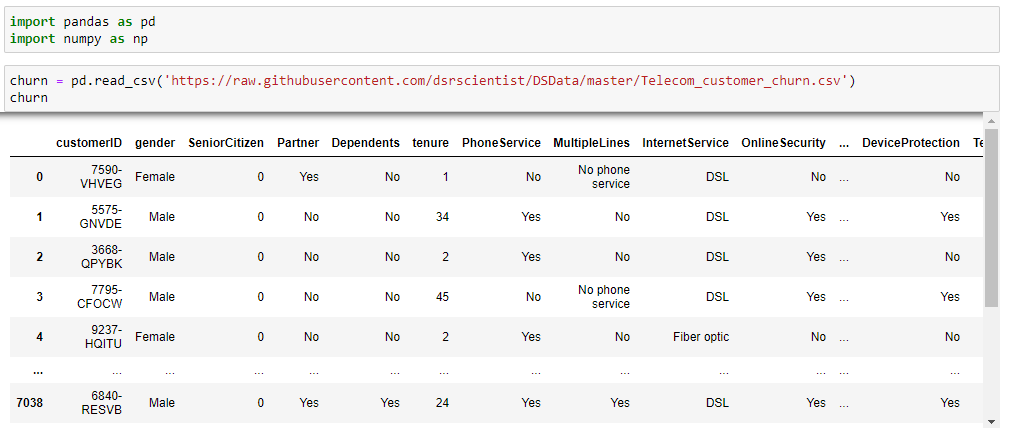
You will examine customer data from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models.

1. **Problem Definition**:

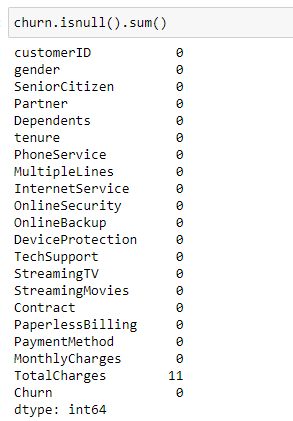
The churning phenomenon is quite common in the competitive markets as the customers naturally tend to choose the firms which offers a reasonable deal and ditch the ones which doesn’t. The companies do not want to lose customers as bringing the new lot would cost extra as it takes a lot for advertising. In order to keep the firm from losing money, they’ll have to retain customers as much as possible. The following problem is to check the factors that are responsible for customer churn and we’ll have to analyse the factors given to come up with a feasible solution possible to retain the customers.

1. **Data Analysis**:

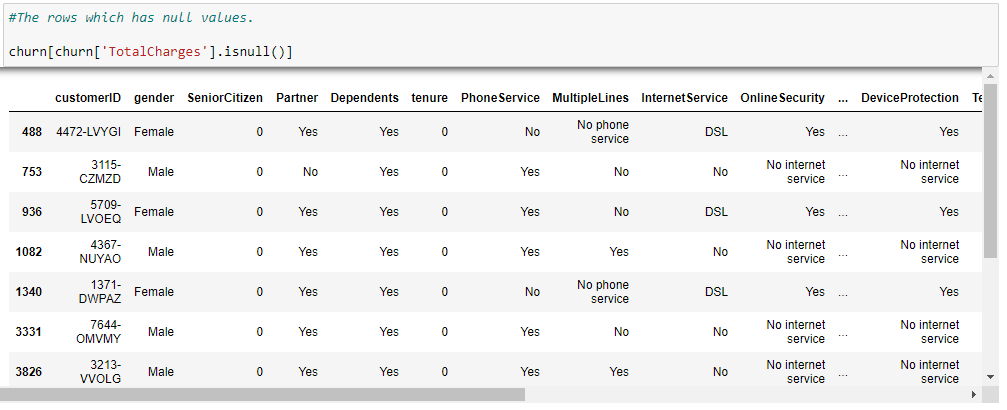
The Data is from a telecom company’s customer churn where there are dependent variables which are like factors influencing the customers to stay or leave.



**Null values check:** The null value presence is very likely in the heavy data like these. The human errors are common and the errors lead to many complications, one such is the presence of Null/Missing values. They can be treated in many ways; we shall first look at the count of null values present in the data.

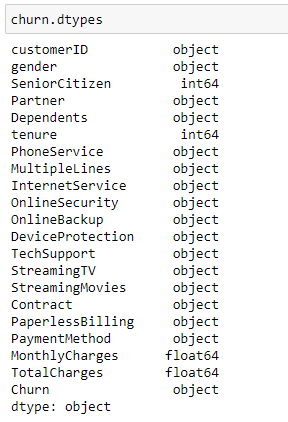


‘Total charges’ have null values in it and it shall be treated later in the pre-processing stage. There is a method to check the null values with respect to the rows.



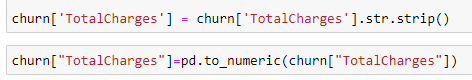
There are 11 rows and 21 columns of null values present in the ‘TotalCharges’.

**Datatypes check**: The data may contain the irregular datatype as there are varied columns with varied values associated with it. Let’s check the datatypes with the help of “.dtypes” method.

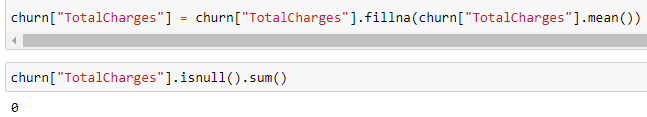


Clearly the numerical data are “Senior Citizen”, “Tenure”, “Monthly Charges”, “Total Charges”.

If the data is clearly seen, ‘Total Charges” column has object datatype in spite of having numerical data. Post data inspection. It is evident that the data has spaces in it. We’ll have to get rid of it to by using the strip() function and convert the datatype from object to float.



As we’ve seen there are null values in the data, since the data is linear type and the null count is only 11, we can go for mean of the data for replacing.



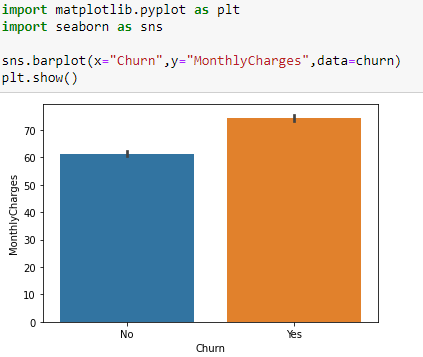
We’ve filled the missing values with the mean value.

1. **EDA**:

Exploratory data analysis is an approach of analysing data sets to using statistical graphics and other data visualization methods. This method is used by data scientists in order to investigate and analyse the data through a visual format and drawing the required conclusions out of it.

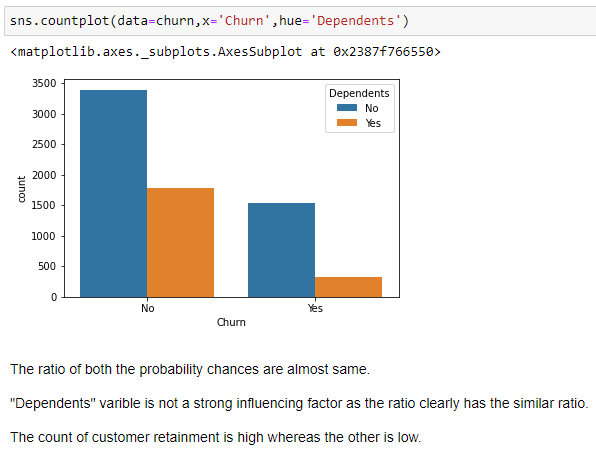
The libraries like matplotlib.pyplot and seaborn are imported with alias like plt and sns respectively. These libraries help to visualise the data and draw the conclusions easily.

* **‘MonthlyCharges’ influence on ‘Churn’**



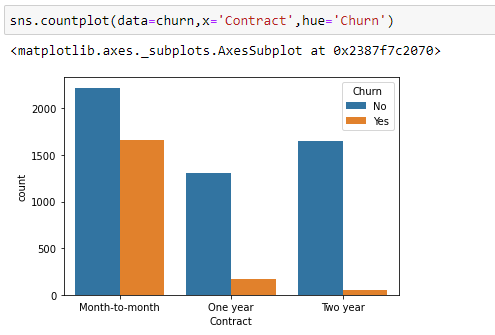
It's quite evident from the above barplot, more the monthly charges, more the customer churn. In order to reduce the customer churn, the firm has to reduce the monthly charges.

* **'Dependents' variable effect on 'Churn’**

****

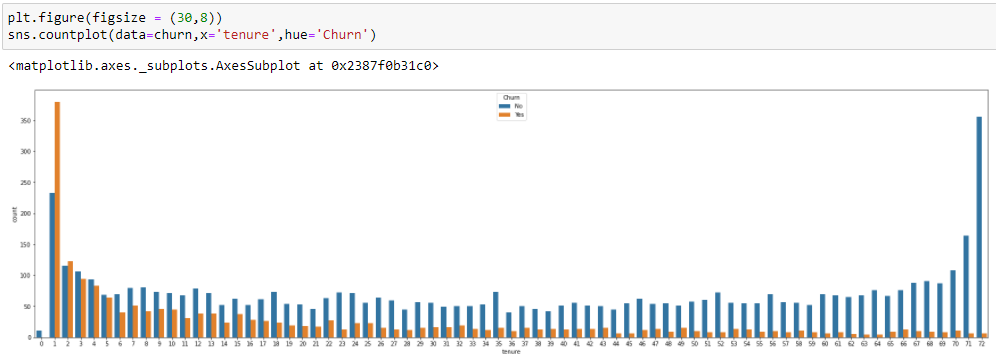
The ratio of both the probability chances are almost same. "Dependents" variable is not a strong influencing factor as the ratio clearly has the similar ratio. The count of customer retainment is high whereas the other is low.

* **'Contract' influence on 'Churn'**



Month-to-month contract customers are highly to be churned compared to other contracts. The company will have to focus more on Month-to-month contract customers and provide them with whatever that is required in order to retain them.

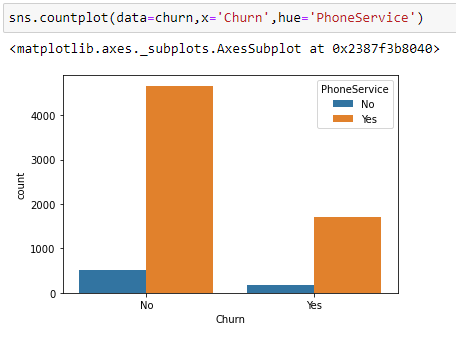
* **'tenure' effect on 'Churn'**



The customers with tenure of one year seems to leave. The trend of customer churn slowly seen to reduce post that first one year.

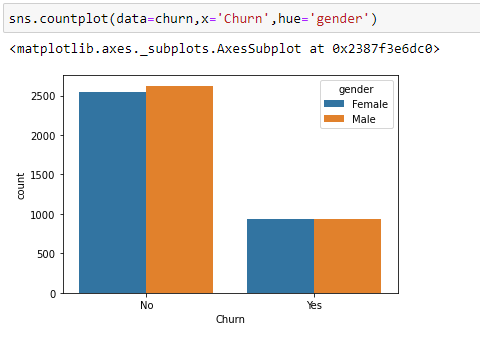
The firm needs to put their maximum efforts to retain the customer during the first one year as much as possible.

* **'PhoneService' variable effect on 'Churn’**

****

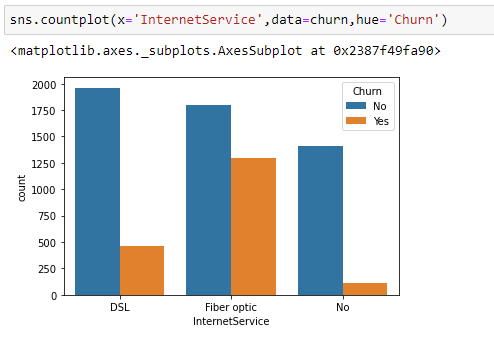
The customers with phone service seem to stay as the count is high in above plot. The firm needs to deploy more Phone Service to customers in order to retain them for longer.

* **'gender' variable effect on 'Churn’**



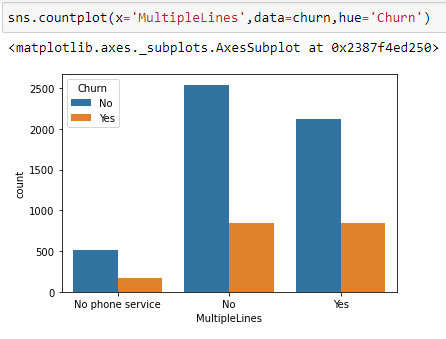
'Gender' variable is ineffective. There is no gender preference or gender-backed attrition in the firm.

* **'InternetService' influence on 'Churn'**

****

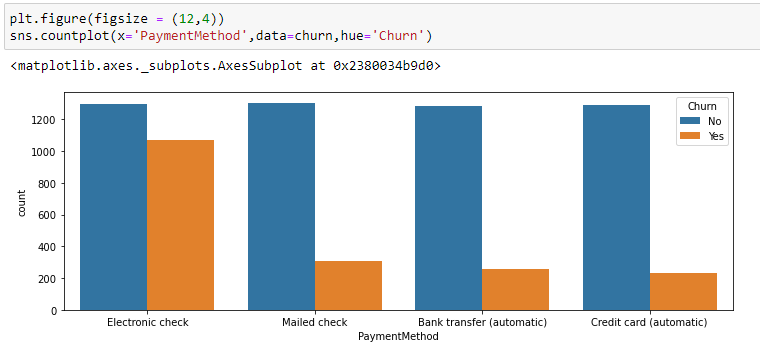
More number of churned customers seem to have Fiber Optic internet connection which looks bad. The firm needs to look it and change the Internet service to DSL.

* **'MultipleLines' influence on 'Churn'**

****

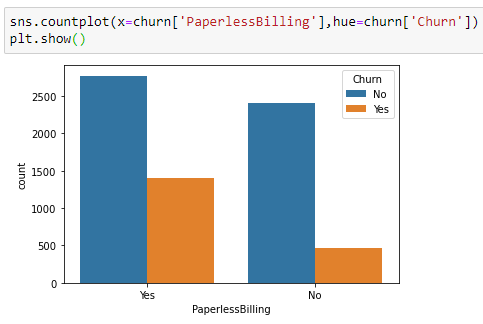
The customers churning count seems to be same irrespective of multiple lines factor.

* **'PaymentMethod' effect on 'Churn'**

****

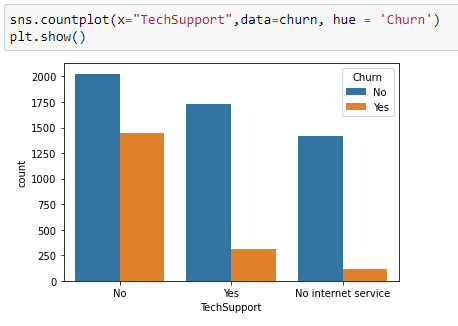
From the countplot, it is quite evident that customer churns are more with the case of Electronic check as payment method. The company needs to look into this and change the payment method to retain the customers.

* **'PaperlessBilling' influence on 'Churn'**

****

The countplot clearly conveys that paperless billing impacts the customers to leave the firm’s association with the company. It's unfortunate that paperless billing concept doesn't seem to work for the firm, instead they'll have to use papers for billing in order to see less churn count.

* **'TechSupport' influence on 'Churn'**

****

The firm needs to give full support on 24X7 Tech Support in order to reduce the customer churn.

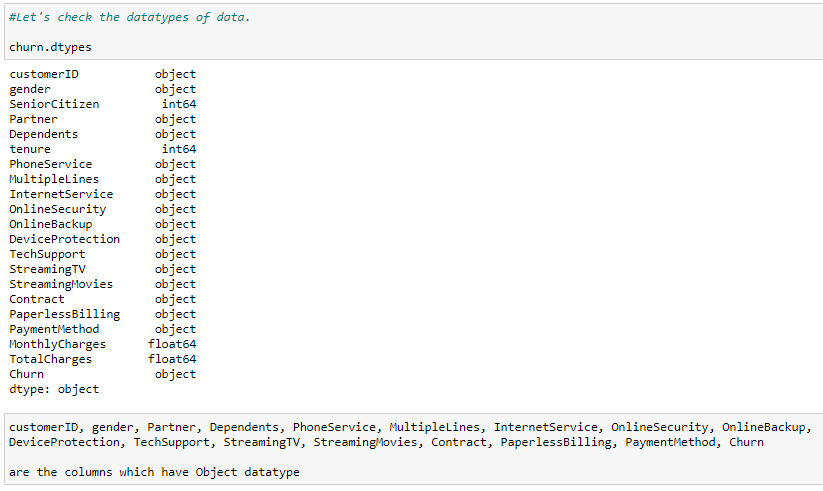
The EDA for the data is mostly seen in the visual format (countplots and barplots). With the help of EDA, we’ve given the necessary suggestions to the company on how to deal with the customer churning based on above variables and its respective effect on the target (‘Churn’ column).

1. **Pre-Processing Pipeline**:

This is another crucial step where it deals with the data mining/data modification to make the data ready for splitting it for training and testing.

In our problem, we’ve come to the stage where there are no more null values/missing values, duplicates. The data is only left to be encoded and scaled according to model readability.

We’ll check the datatypes of the data and then the columns which are object type are taken into account for converting them into numerical.

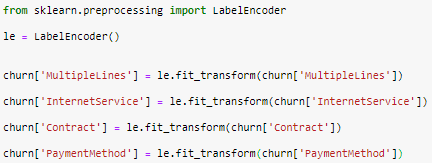


We’ll write a dictionary code where their categorical values are assigned to 0 and 1 respectively. Then, we’ll map the columns to that dictionary code.



customerID, MultipleLines, InternetService, Contract, PaymentMethod are the columns which have either unique values or multi-categorical values. Thus, these columns need special attention for the conversion. Label Encoding will do the job.

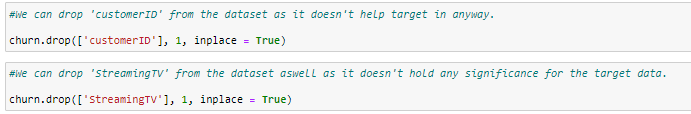
* **Label Encoder**

****

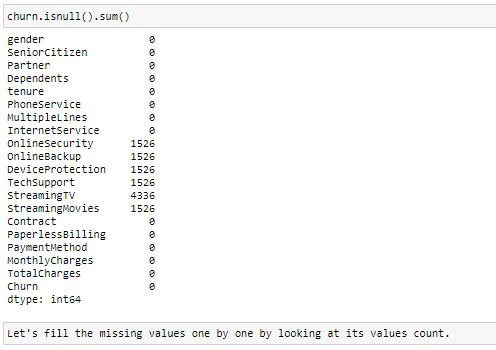
The Encoding is done and the columns now will have their categorical values replaced with the numbers respectively. The order of the numbers being assigned will be in alphabetical order.

* **Data Drop**

The data sometimes needs to be omitted according to their effectiveness and impact over the target variable. If it is of no use, we can get rid of it with the help of drop() function.

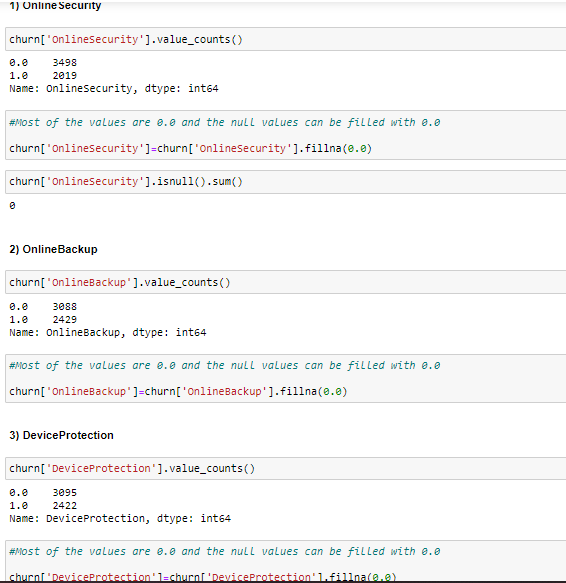


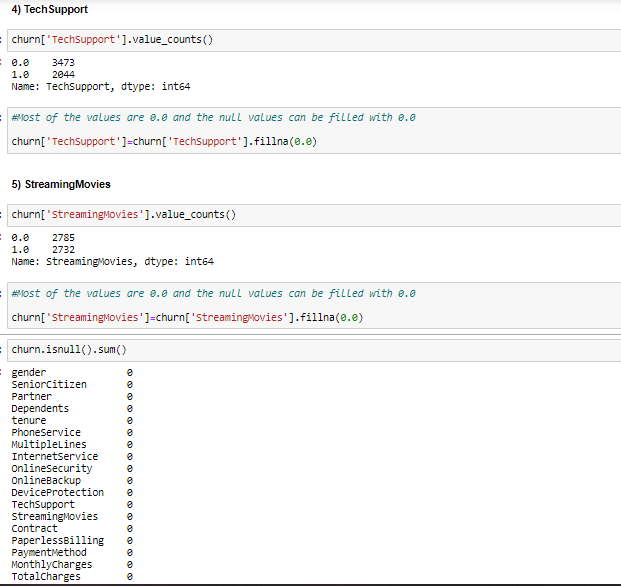
* **Filling Missing/Null values**



Post Label Encoding, some of the missing values might’ve got highlighted since the missing ones are foreseen by the encoder.

We’ll have to work on this one by one by going through its unique values. If a certain value’s count is high, we can replace the missing ones with the most occurring value.

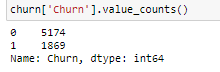




We’ve successfully removed the null data by replacing every missing data with the most occurring value in the respective columns.

* **Over-Sampling**

The target variable is imbalanced as we can see the distribution of the data is uneven.



The imbalance is large enough to give us the biased scores while testing. We shall need to balance the data for model training.

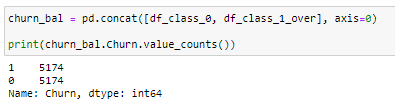






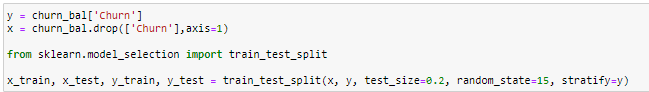
We’ve created a data where the ‘Churn’ column is only 1 as the duplication came into the picture in order to compensate with the count of 0s in it.

Finally, we’ve sampled the final data where the value count is equal and balanced.



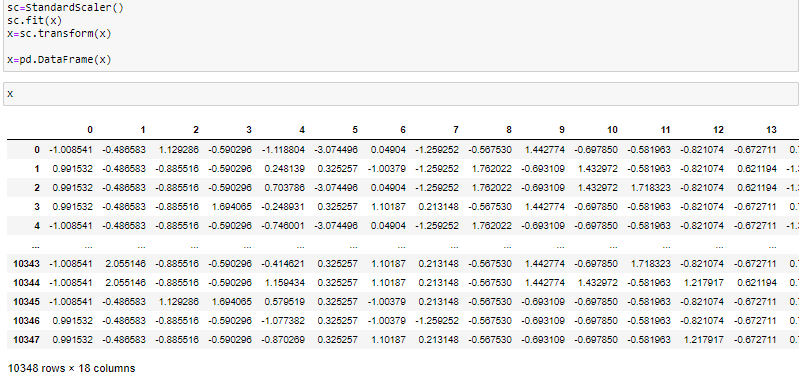
* **Data Split (train\_test\_Split)**

The data is ready with no presence of null/missing values. Whatever the modification required, has been done till here. Now, the data will be split into ‘x’ and ‘y’ where the ‘y’ will be assigned with the target variable (‘Churn’) and ‘x’ with the rest of data.



* **Standard Scaler**

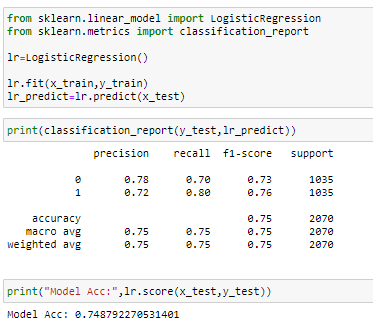
The purpose of Standard Scaling is to scale the data into a common scale in order to avoid the model from falling prey to variances/bias.



1. **Building Machine Learning Models**

The target variable (y) is a liner data thus regression models are used one by one.

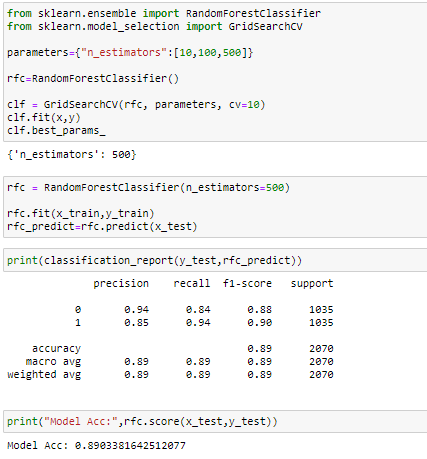
* **Logistic Regression**

****

The Logistic Regression is imported from ‘linear\_model’ library of ‘sklearn’ package.

The model accuracy aka model’s performance in prediction is 74.8%.

* **Random Forest Classifier**

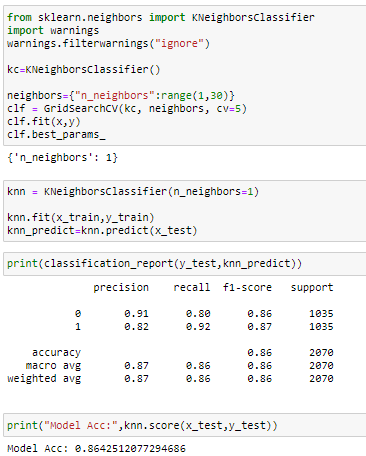
****

The model is imported from ‘ensemble’ library of ‘sklearn’ package.

Using GridSearchCV, we can get the best parameters for using in rfr in order to get the best score. The final score after using the best estimators and apt parameters is 0.89.

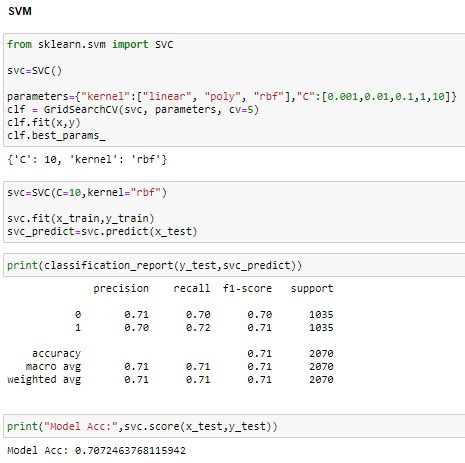
The model accuracy aka model’s performance in prediction is 89%.

* **KNN**

****

The model is imported from ‘neighbors’ library of ‘sklearn’ package. The model with apt parameters from GridSearchCV, selective neighbors will make the model’s performance in prediction to be 86%.

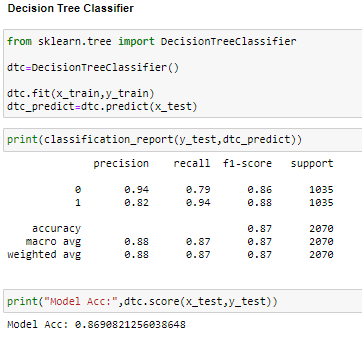
* **SVM**



The model is imported from ‘svm’ library of ‘sklearn’ package.

The model accuracy aka model’s performance in prediction is 70.7%.

* **Decision Tree Classifier**

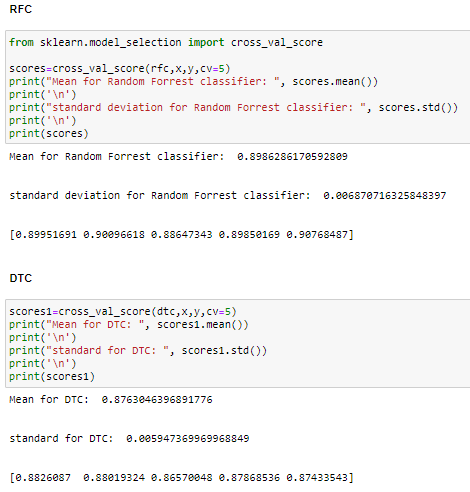


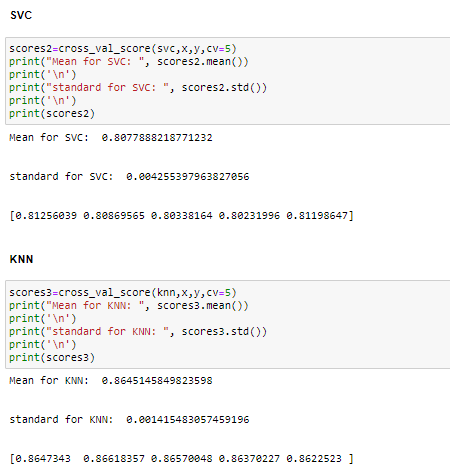
The model is imported from ‘tree’ library of ‘sklearn’ package.

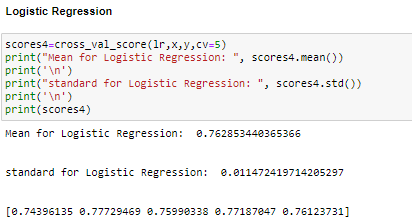
The model accuracy aka model’s performance in prediction is 86.9%.

* **Cross-Validation Score**

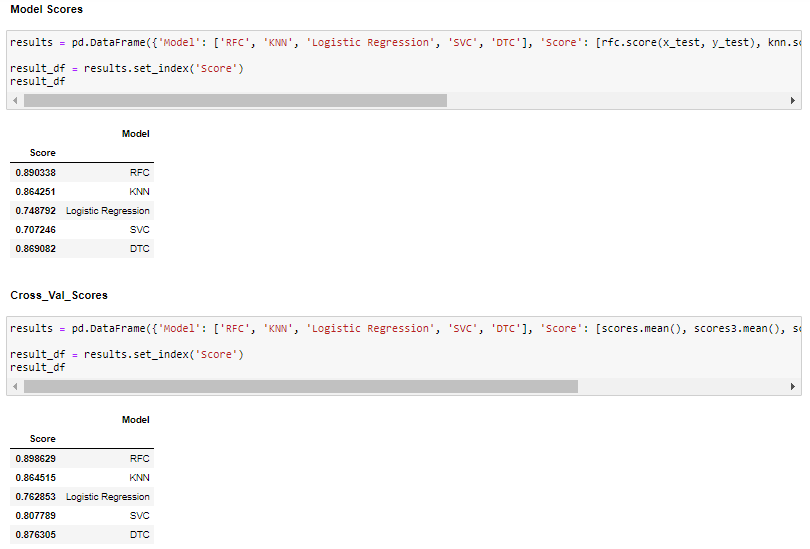
It is a technique and not a model which majorly helps the models from falling prey to variance and bias. The models tend to have sensitive stimuli to the errors like bias/variance, which will lead the model to read the biased data or may ignore the minor set of the data. Thus, with the help of cross validation, the K-fold technique will help in splitting the data into 5 or 7 or any number of folds (cv = any number) to improve the genuine scoring of models.

****

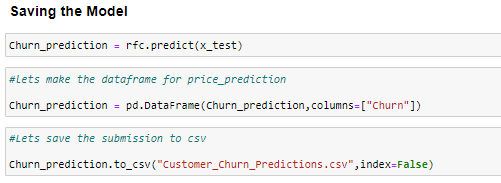




1. **Concluding Marks**

****

Random Forest Classifier has the best score after the Cross Validation with 0.89. We can zero-in this model and save the model to predict the score for further use.



We’ve given the company with a model which can predict the customer churn probability to 89%. The company can follow the model and take the churning factors into account so that they can improve their support for retaining the customers as much as possible.