**ML – ASSIGNMENT**

**CAP – 1**

**Submitted To: Dr. PP Rahul Krishnan**

**Submitted By: Shashwat Srivastava**

**Colab Notebook Link :** [**https://colab.research.google.com/drive/16hfL-\_-JhX6\_IGt3Kw30M2o4cmm-GXJJ?usp=sharing**](https://colab.research.google.com/drive/16hfL-_-JhX6_IGt3Kw30M2o4cmm-GXJJ?usp=sharing)

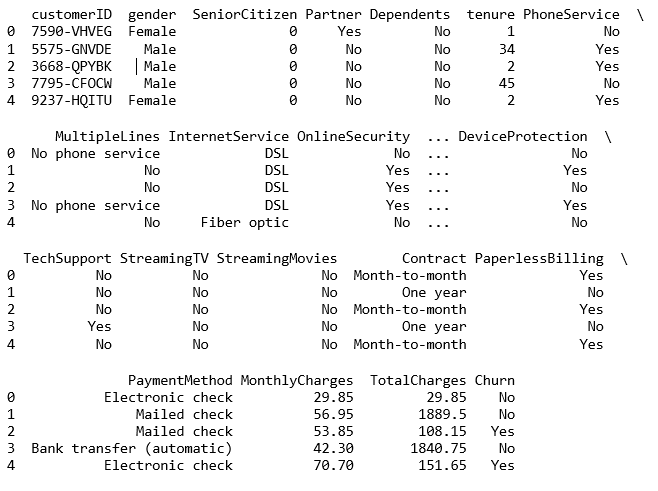
**Dataset used for the assignment :**

Telco-Customer Dataset ( From Kaggle)

This dataset is available on Kaggle. It contains 19 columns in total. The column considered as response variable in the dataset is the ‘Churn’ column. The churn column defines if the customer has churned out of the company or not.

1. **Data Reading :**

Successfully read the data and described the heads of the dataset.

****

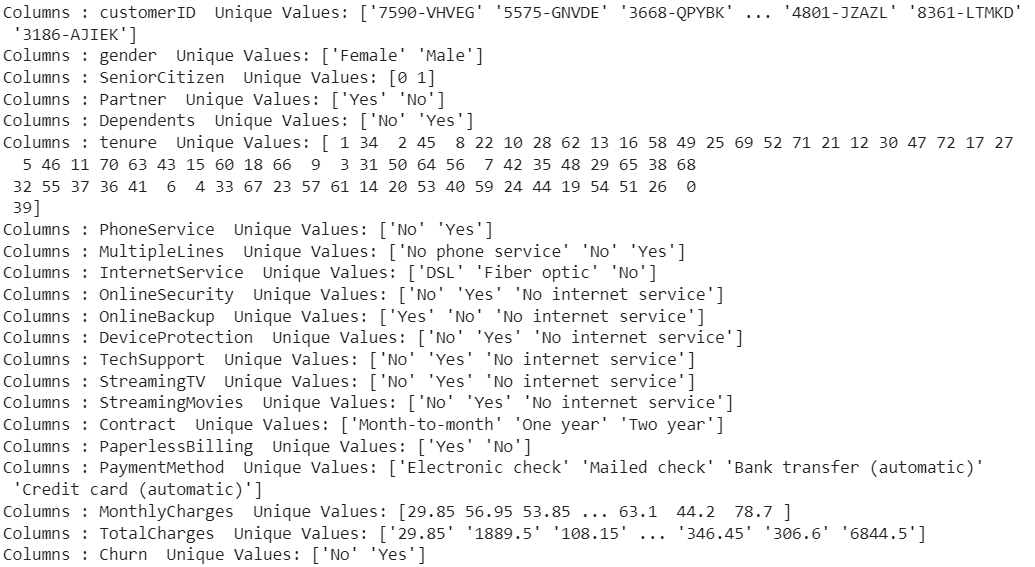
1. **Exploratory Data Analysis :**

The first step towards EDA, is to analyze the type of variables are present in the column. To do that, we need to find the unique values of all the columns present in the dataset.

We find out that, there are **13 categorical variables** and **6 continuous variables.**

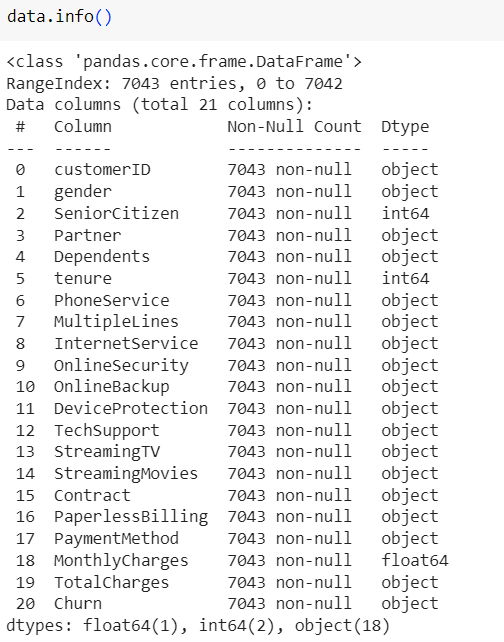
**Categorical Variables :** gender, SeniorCitizen, Partner, Dependents, PhoneService, MultipleLines, InternetServices, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies.

**Continuous Variables :** tenure, contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges



We can further divide the categorical variables into Demographic and Non-Demographic Variables. So now, there are **4 Demographic Variables** (gender, SeniorCitizen, Partner, Dependents) and **9 other are Non-Demographic.**

**Now, we try to find the missing values in the dataset.**

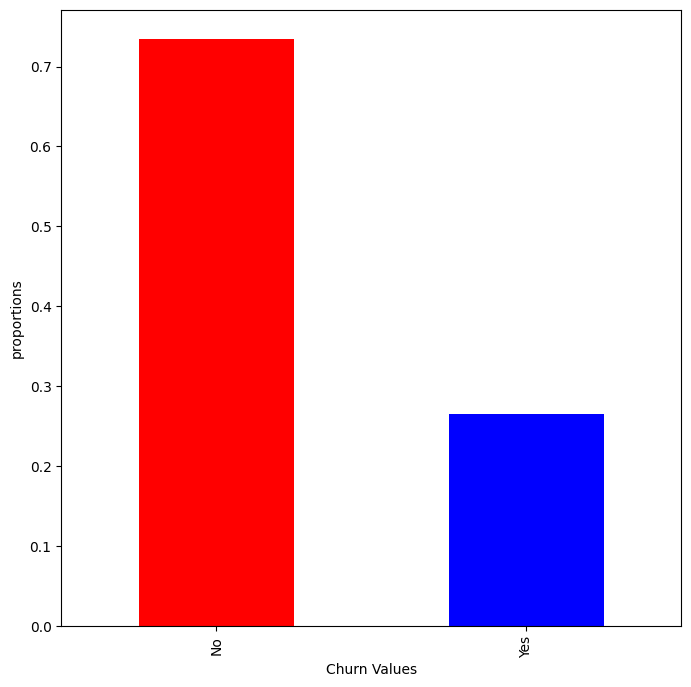
****

* We conclude that there are 7043 observations and 21 columns in the dataset.
* There are no NULL values in the dataset.
* We find that TotalCharges has been misclassified as an object when it is numeric variable.
* We tried to convert the TotalCharges column to a numeric type but it was affecting the data and making it dirty, so we decide to drop it.
* Also we drop the CustomerID since, it is of no use in the analysis.

Now we move towards Data Visualisation.

1. **Data Visualisation :**

We try to analyse the Response Variable at first i.e. ‘Churn’ column and see its proportions

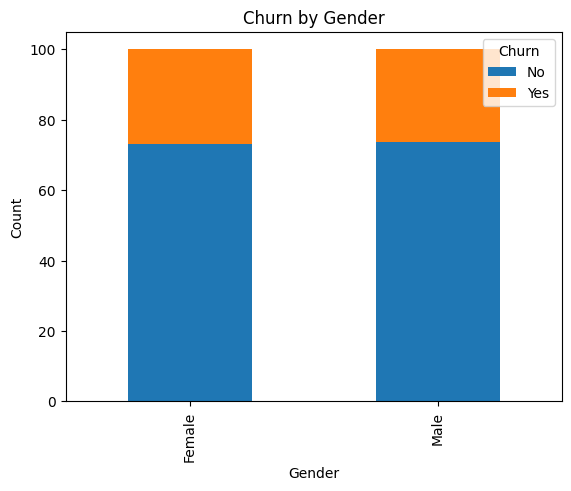


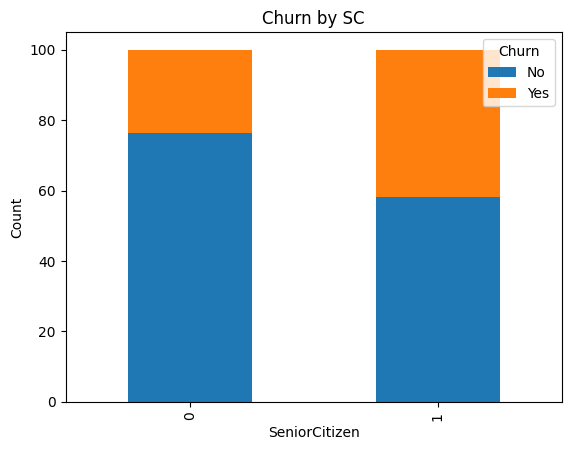
We can see that the around 75% of the data is not churned and the remaining is churned. This also shows that the data is imbalanced and both the classes i.e. yes and no are not distributed equally in the dataset.

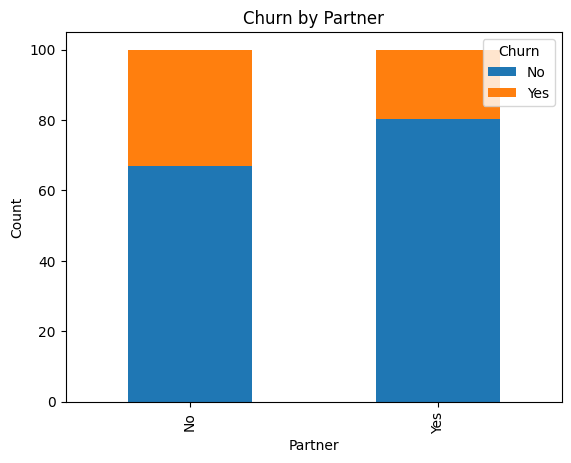
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

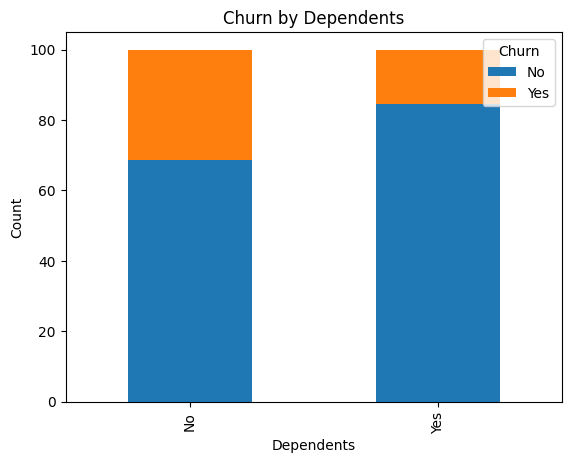
We now try to analyze the categorical variables, most precisely the demographic ones first.

To analyse the categorical variables vs Churn Variable, we can use the Stacked Bar Chart Plot.









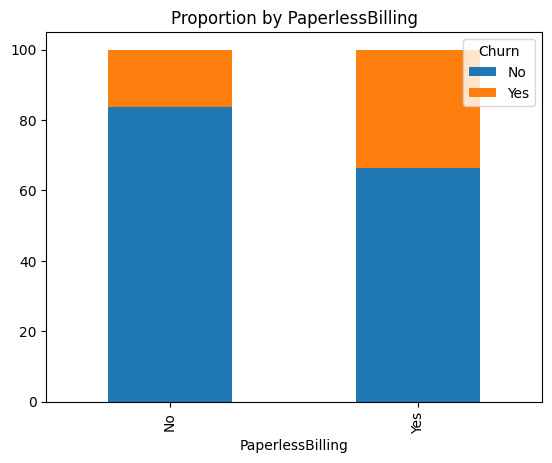
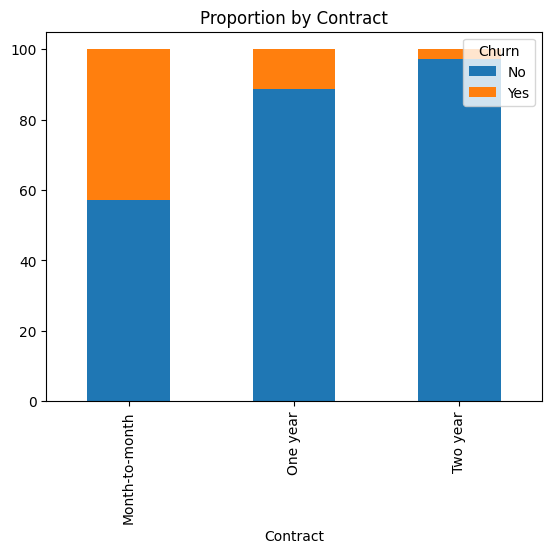
Analyzing the plots, we can conclude that :

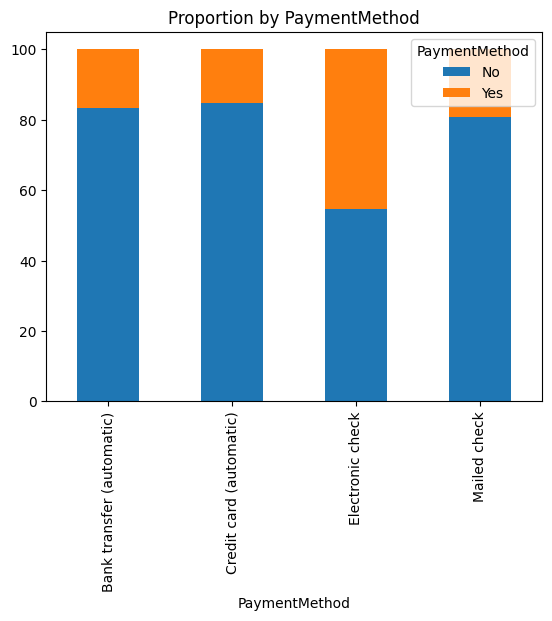
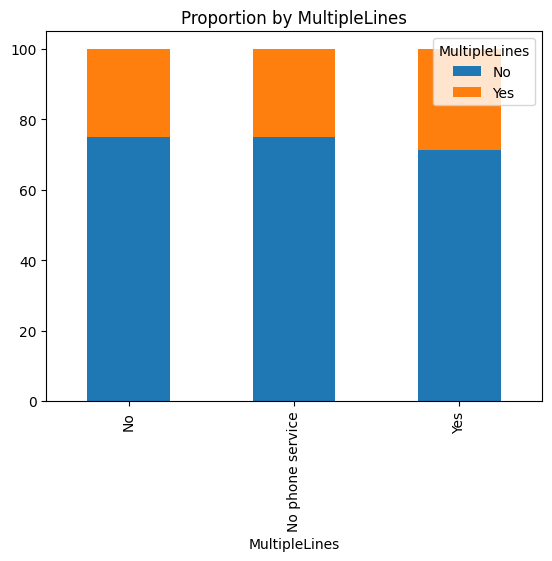
* We should not use gender column, as it does not predict anything. It shows almost the same level of Churns for both the genders.
* For the SeniorCitizen, the churn rate is almost double as compared to the other population.
* Churn Rate for Customers with partners is less than those with no partners
* Customers with Dependents have a lower Churn Rate.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Now , we analyze the dataset for the Categorical variables :

We take into consideration Contract, PaperlessBilling, MultipleLines and PaymentMethod. Since, all the other categorical variables seem to be more of a service, so we’ll analyze it separately.



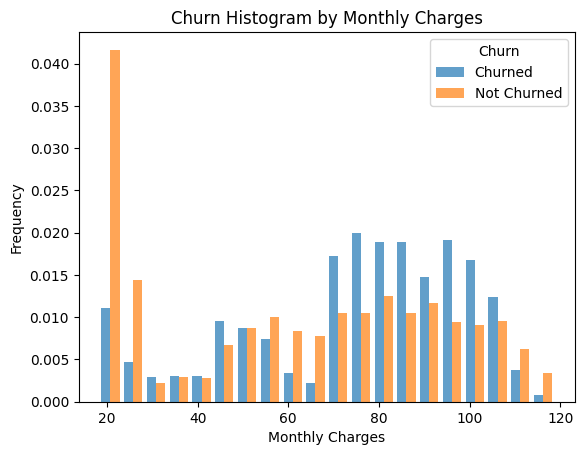
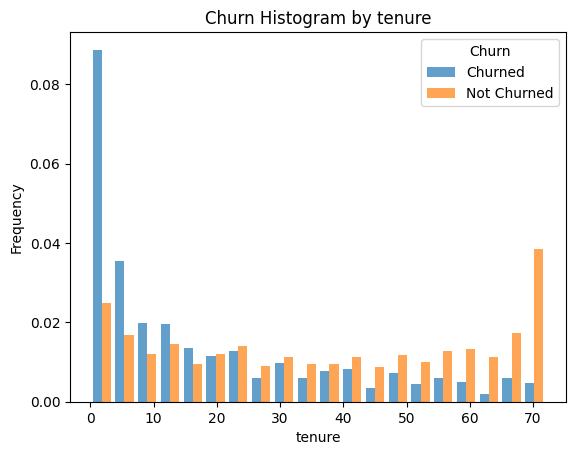
****

Analyzing the plots we can conclude that :

* MultipleLines cannot be used as a parameter to analyze any output.
* Longer the period of Contract, Lower will be the Churn Rate.
* Customers with PaperlessBillings have a higher Churn Rate.
* Customer who have ElectronicCheck PaymentMethod churn more than than any other. All the other have comparatively equal Churn Rates.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Now we try to analyze the Continuous variables. We usually use Histograms to present Numerical vs Categorical EDA. We analyze tenure, MonthlyCharges and Churn variables and plot its distribution.



From the above plots, we can conclude that :

* Customers with a lower tenure value, have a higher Churn Rate.
* Customers with a lower MonthlyCharges values, have a lower Churn Rate value. Customers with MonthlyCharges of 70-105, have a higher Churn Rate value.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Encoding :**

Encoding is an important part of Feature Engineering where we transform the data that can be fit into a model.

1. **Label Encoding:**

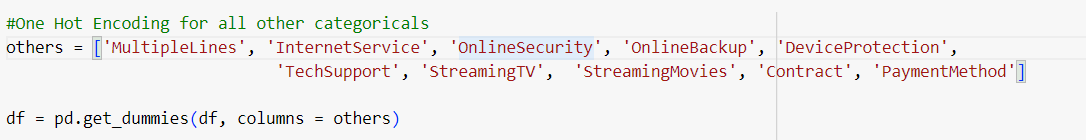
Used to replace categorical values with numerical values. Here we replace the values of binary variables into numerical values. The variables, [ gender, Partner, Dependents, PaperlessBilling, PhoneService, Churn ] can be converted to 0s and 1s since they are binary.

We create a copy of the original dataset and we alter the modifications in that. All the encoding is done on the new copy of the dataset and the same will be analyzed further for model fitting.



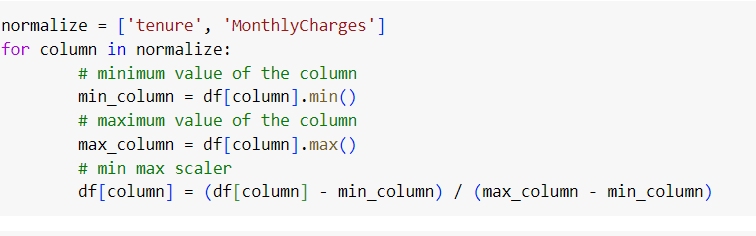
1. **One Hot Encoding :**

Used to convert categorical variables to binary vectors. Now here when the category is present, it is given as value = 1 and any other scenario would be given the value = 0.



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

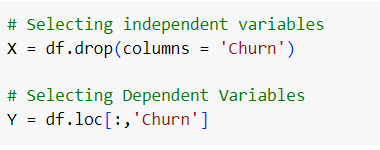
1. **Normalization:**

We use normalization to bring all the numeric values to a common level. We use the min-max method of Normalization here to rescale the numeric columns [tenure, MonthlyCharges] to a common scale. 

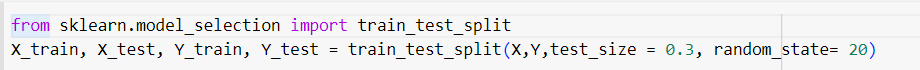
1. **Split the data into Train and Test Set :**

We split the data into test and train data, where 70% of the dataset is training data, while the other 30% is the test data.

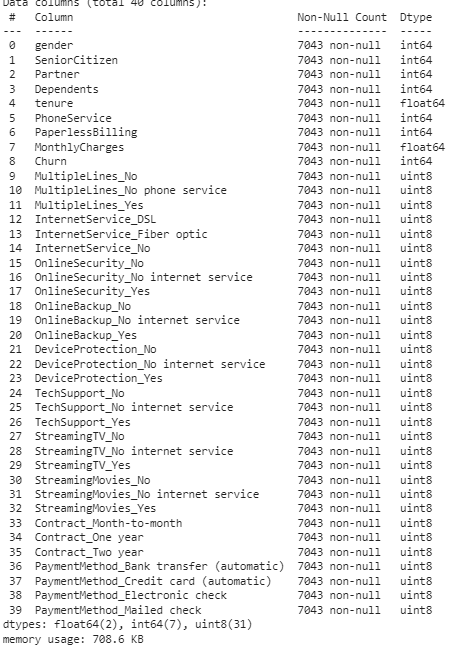
We need to store the independent variables and the dependent or the target variable seperately. So we store the independent variables in X and Churn variable in Y.



Using the train\_test\_split function we split the data into Test and Train Dataset.



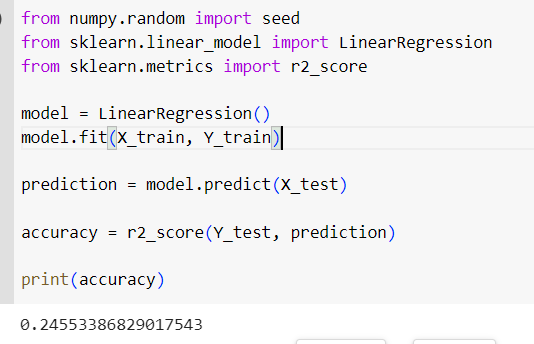
We again check for the information summary of the dataset to see if there is any other issue in it.



And we find out that the dataset is completely correct.

1. **Algorithm Selection**
2. **Linear Regression :**

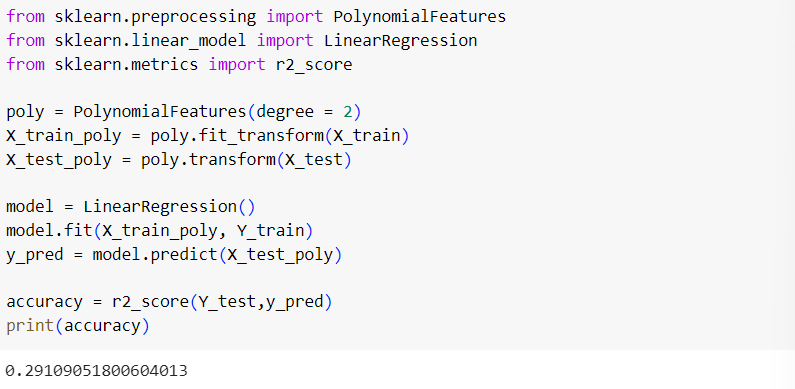
We fit the data into the Linear Regression Model imported from the Scikit Learn Library and it gives the **accuracy of 24.55%**. Linear Regression works on the Mean Squared Error Model and hence accuracy\_score metric is not used here. Instead we use r2\_score metric to predict the accuracy of the model.



1. **Polynomial Regression:**

In the model developed for the Linear Regression, we try to fit the same model with Polynomial Features and degree of 2. We aren’t using Hyperparameter tuning here, so we can only do hit and trial.

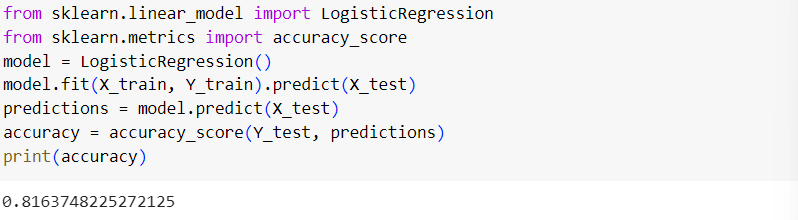
**Using degree = 2, it shows the accuracy of 29.19%**

****

1. **Logistic Regression :**

We try to fit the data into LogisticRegression classifier imported from Scikit Library, We fit the train data and try to predict using the test data. The accuracy\_score metric is used here to calculate the accuracy of predictions when test data and predicted data are compared.

**Using Logistic Regression, the accuracy score came out to be 81.65%**

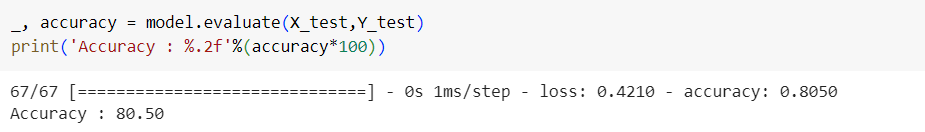
****

1. **ANN models:**

* **3 – Layered Model :**

Firstly we consider the 3 layered model of standard units 12, 8 and 1 respectively. We use the adam optimizer and a loss Binary Cross Entropy loss function. We fit the data into this 3-layered model and gain an **accuracy of 80.50%**





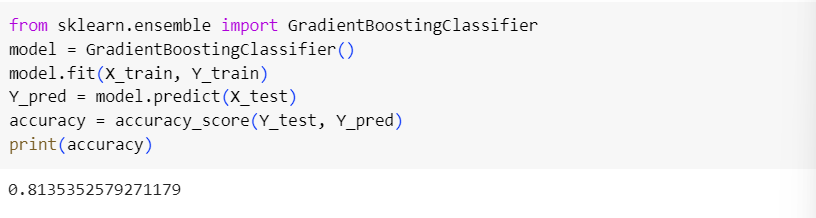
* **4 – Layered Model : (Better Model Architecture)**

Secondly, we consider the 4 layered model of standard units 19, 15, 10 and 1 respectively. **We use this architecture because there are 19 columns in the dataset after dropping the originals of the dummies made. So there are 19 units in the input layer and 2 hidden layers.** We use the adam optimizer and a loss Binary Cross Entropy loss function. We fit the data into this 4-layered model and gain an **accuracy of 79.79%**

1. **Gradient Boosting :**

From the Scikit Learn Library, we use GradientBoostingClassifier and fit our data on to the model and predict the scores of the test data. We check for the **accuracy and it comes out to be 81.35%.**



**Accuracy Scores of different Algorithms:**

1. *Linear Regression Model:24.55%*
2. *Polynomial Regression Model:29.10%*
3. *Logistic Regression Model:81.63%*
4. *3-layered ANN Model:80.50%*
5. *4-layered ANN Model:79.79%*
6. *Gradient Boosting Model: 81.35%*

**Note:**

**We should not forget that we have trained all these data using the default hyperparameter and that this training has not been done with Hyperparameter Tuning.**

**Therefore, the best model that fits the data is the Logistic Regression Model.**