**Blog-1**

**Customer Churn Analysis**

**Author**

**Abhishek Kumar**

**Customer Churn Analysis**

**Problem Definition:**

About Problem Defination of Customer Churn Analysis- Customer churn prediction is used in many businesses to evaluate a company’s loss rate. Customer churn occurs when a company’s customers stop doing business with that company. Businesses are very keen on measuring churn. 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. Predicting the churn rates accurately is important as it helps the business in better understanding future expected revenue. It can also help in identifying mistakes and improve in areas where there is a lack of customer satisfaction.

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 prioritize focused marketing efforts on that subset of their customer base.

Customer churn rate can be calculated by dividing no of customers lost in a given time interval by the total no of customers multiplied by 100. For example: If there are 100 customers in a company and it has lost 5 customers in a month then that month’s churn rate will be (5/200)\*100 = 1.5%. So that month’s churn rate is 1.5% of that company.

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

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

Link to find the .csv file is on Github is : <https://github.com/dsrscientist/DSData/blob/master/Telecom_customer_churn.csv>

**Predicting whether a customer will churn by learning models on telecom industry dataset provided by IBM data Company-**

This notebook covers following contents –

Reading the data

Overview of data's structure - how various features and their respective values look like?

Finding and handling missing values

Dealing with categorical attributes

Identifying higher correlation features (with the target)

Generating relevant insights about values of these high correlation features for churned customers (

Preparing data for models

Model generation and performance evaluation

**Importing the data:**

We need to import all the relevant libraries:

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.preprocessing import LabelEncoder

from sklearn.preprocessing import PowerTransformer

from sklearn.preprocessing import power\_transform

from sklearn.preprocessing import StandardScaler

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LogisticRegression

from sklearn.naive\_bayes import GaussianNB

from sklearn.tree import DecisionTreeClassifier

from sklearn.svm import SVC

from sklearn.neighbors import KNeighborsClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.ensemble import AdaBoostClassifier

from sklearn.ensemble import GradientBoostingClassifier

from sklearn.metrics import accuracy\_score,confusion\_matrix,classification\_report

from sklearn.model\_selection import GridSearchCV

from sklearn.model\_selection import cross\_val\_score

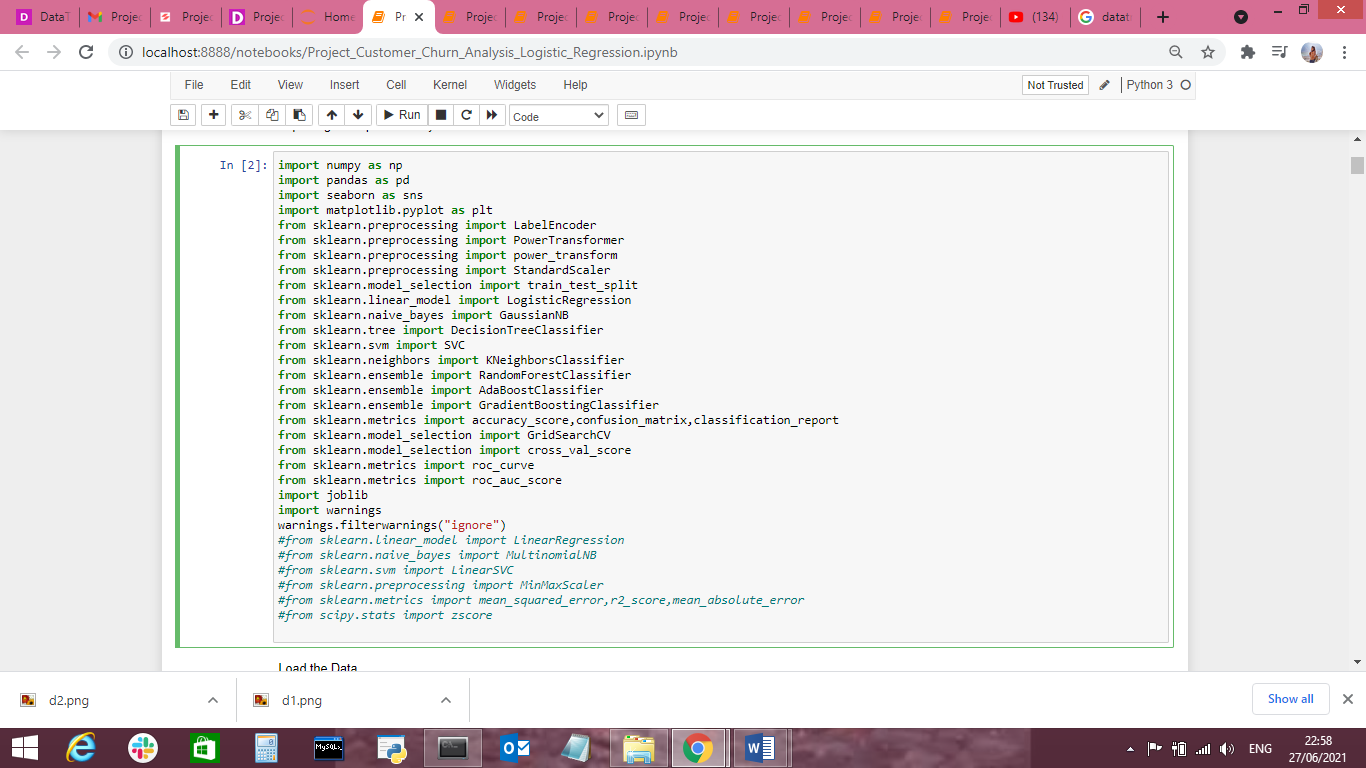
from sklearn.metrics import roc\_curve

from sklearn.metrics import roc\_auc\_score

import joblib

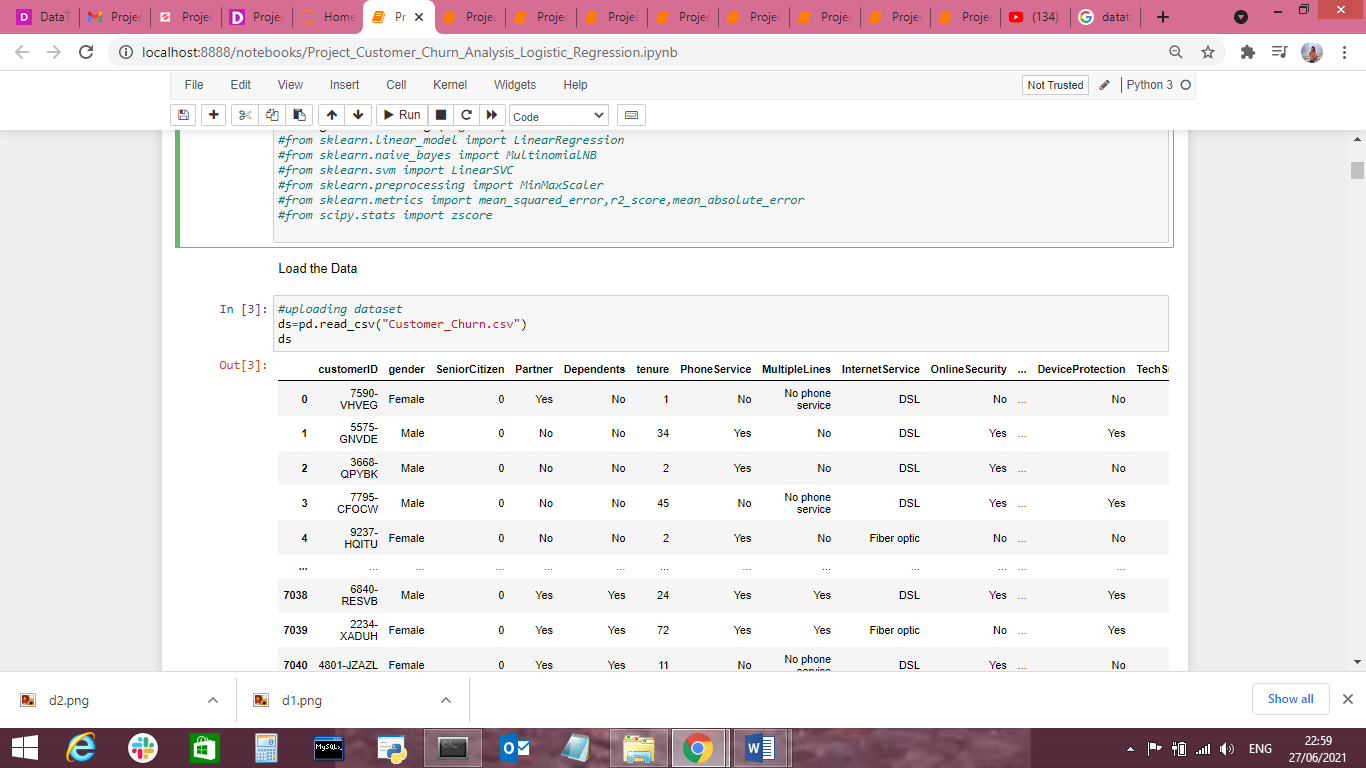
import warnings

warnings.filterwarnings("ignore")



We have now imported all the important libraries that we will be needing during analysis.

We need to import the .csv file into the Jupyter notebook as shown below.



This data set contains both Independent and Dependent (target) variables.

Independent variable: They are also known as Input variables. These are the input for a process that is being analyzed.

Dependent variable: They are also known as Output or Target variables. They are dependent on Independent variables for their outcome

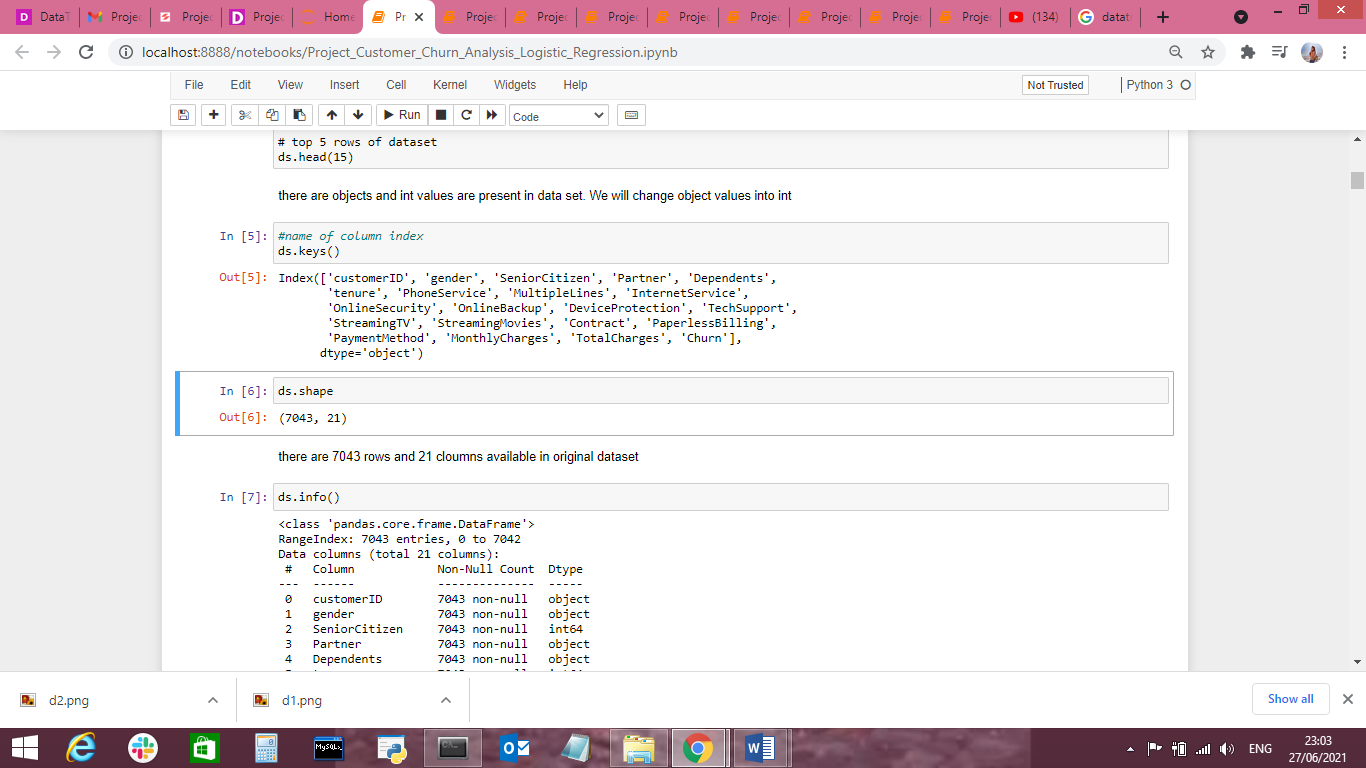
After importing the dataset, display a sample of data. The variables in the dataset are as follows:

* customerID
* gender
* SeniorCitizen
* Partner
* Dependents
* Tenure
* PhoneService
* MultipleLines
* InternetService
* OnlineSecurity
* OnlineBackup
* DeviceProtection
* TechSupport
* StreamingTV
* StreamingMovies
* Contract
* PaperlessBilling
* PaymentMethod
* MonthlyCharges
* TotalCharges
* Churn

**Dataset Analysis (EDA):**

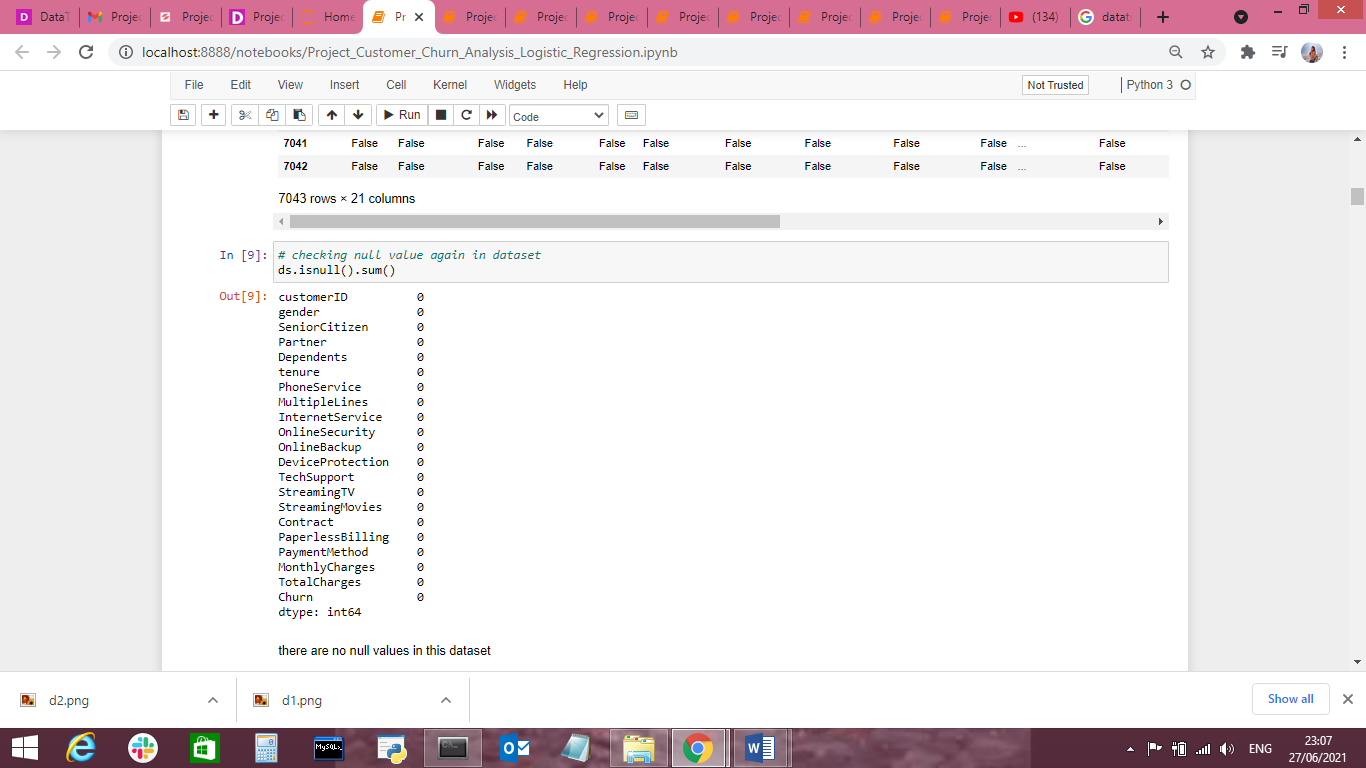
**Now we need to understand the dataset by performing Exploratory Data Analysis. We will check every detail of dataset like shape,info,keys, values, etc…**

Let’s check the shape and keys of the data set:



We can see that there are 7043 rows and 21 columns in the dataset.

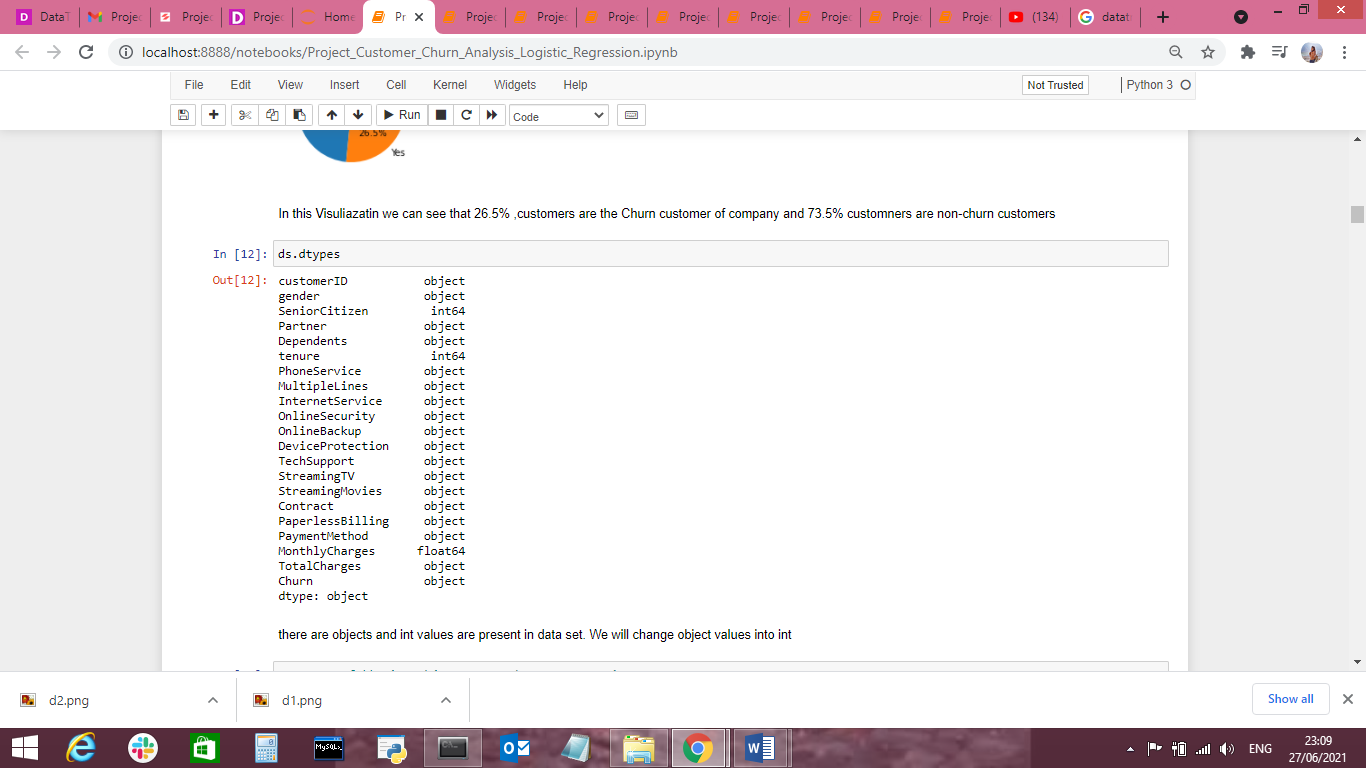
**Checking for null values in the data set:**



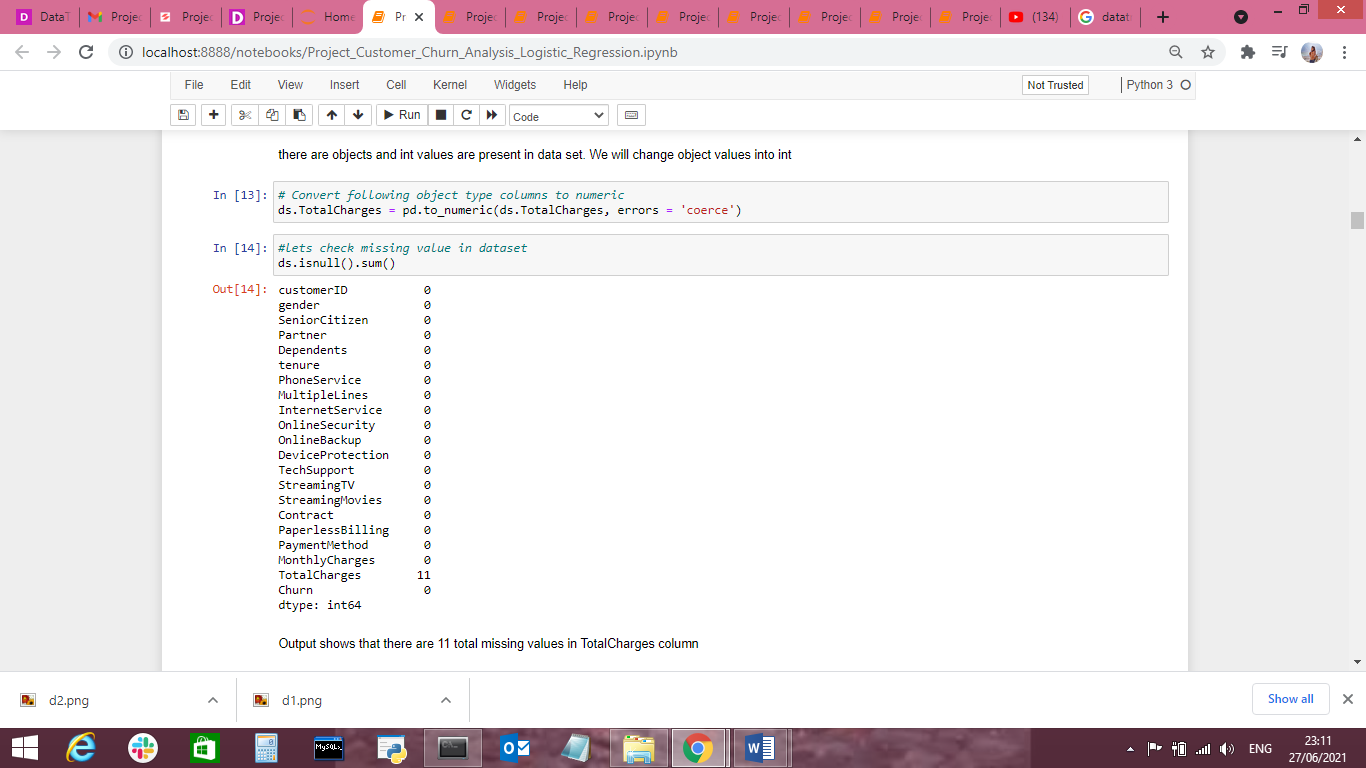
We cannot have null values in the data as this will affect the data and eventually, the predicted result will not be correct. Therefore we must check for any null values in the dataset.

And We can see that there are no null values in the dataset and we are good to go now.

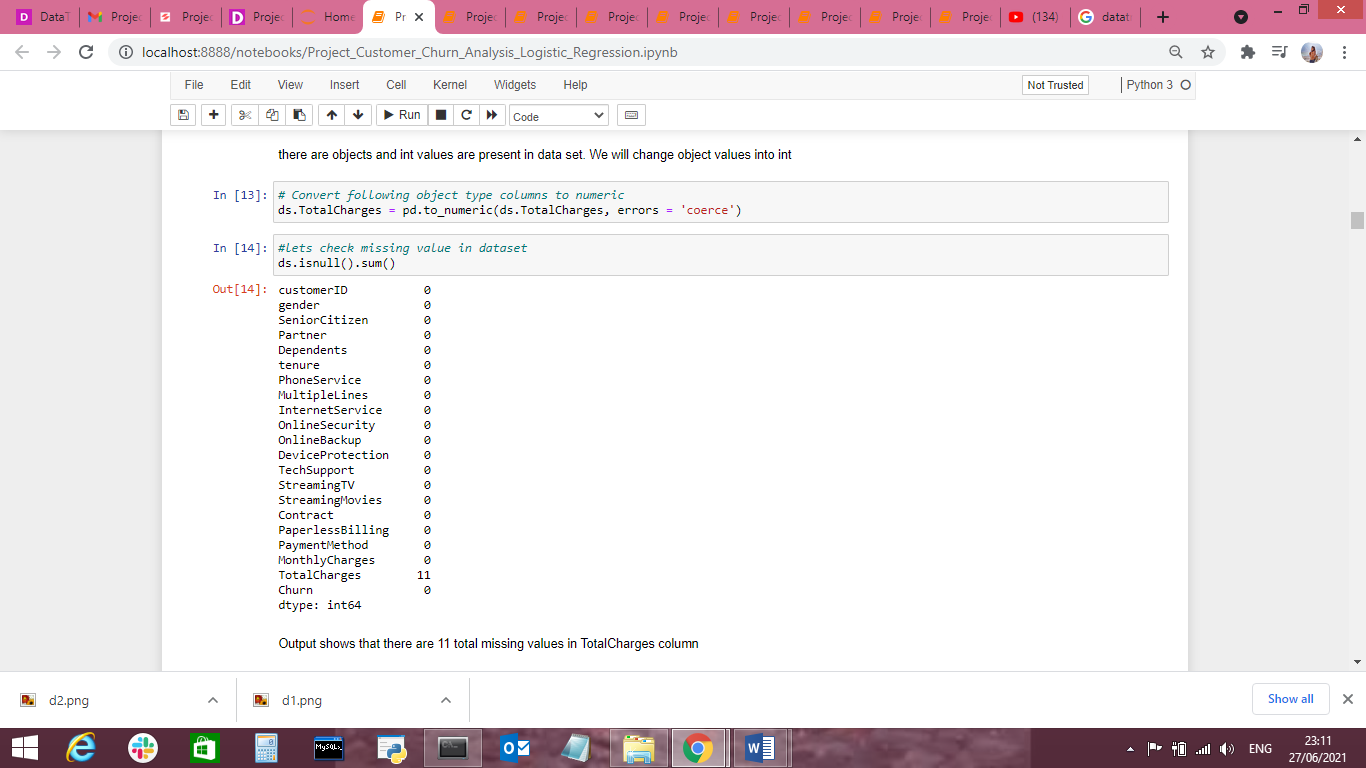
**Checking data type:**



IN this dataset TotalCharges column type is object so we have to change it into Numeric values



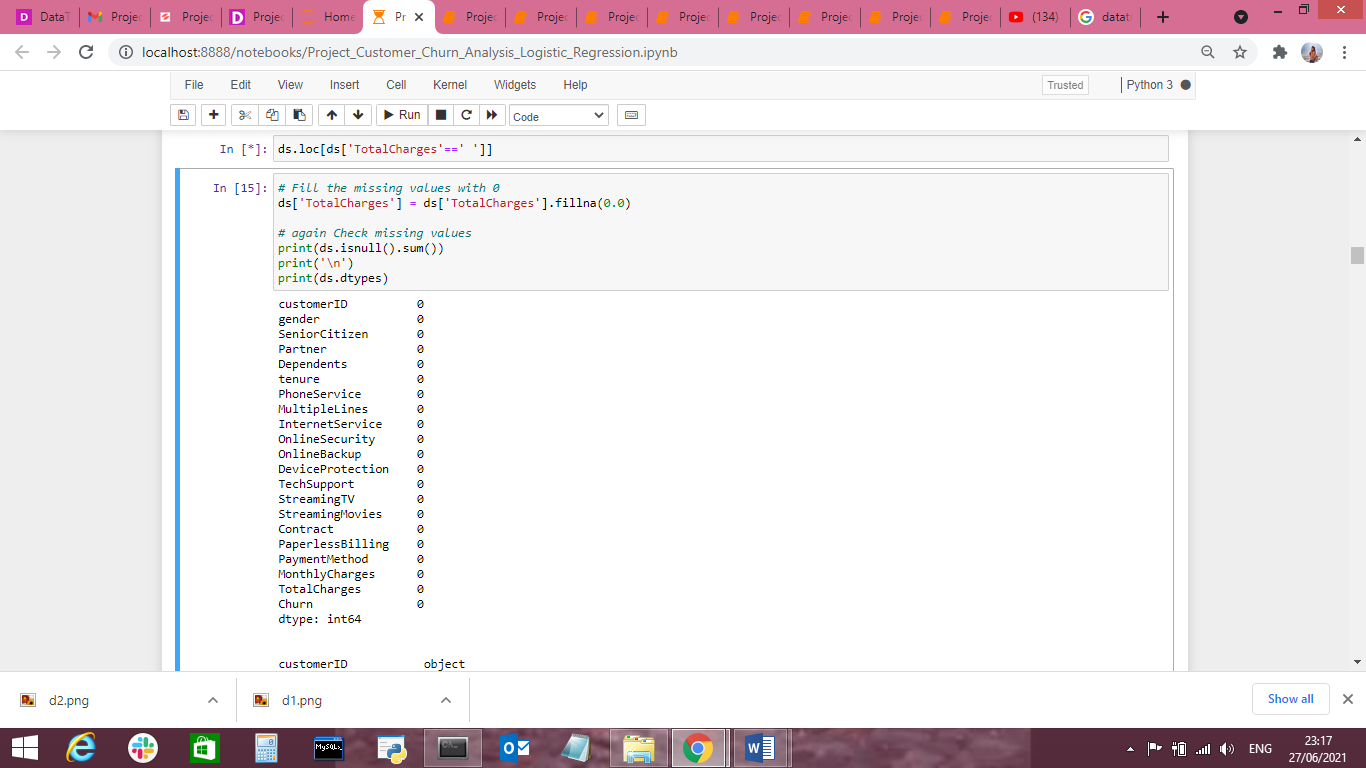
After changed column “TotalCharges” into Numeric ,lets check again the null values



**And now we can see that Totalcharges column has 11 missing value so we have to deal with it.**

**Handling Missing data:**

From the above figure, we can see that there are 11 missing values in the ‘TotalCharges’ column as highlighted. As the dataset has 7043 rows we can drop these missing values as it will have a negligible effect on the dataset.

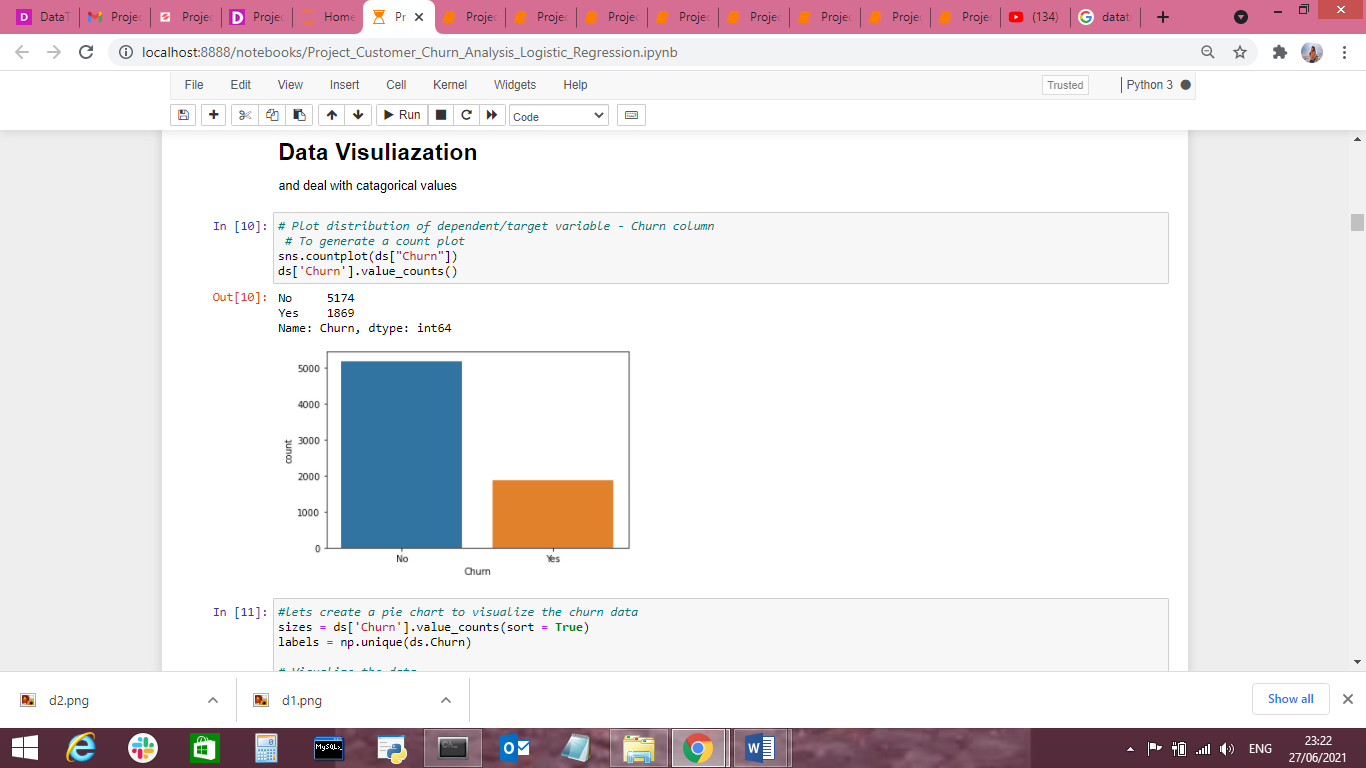


Now theres no missing value into column Total Charges.

**Data Visualization and EDA Concluding Remarks:**

In the given data, ‘churn’ feature is the Target feature or variable. The unique values of this feature are only 2 i.e Yes and No, which means it has only two classes. So, as there are only two unique values this is a ‘Classification Problem.’, Logistic regression

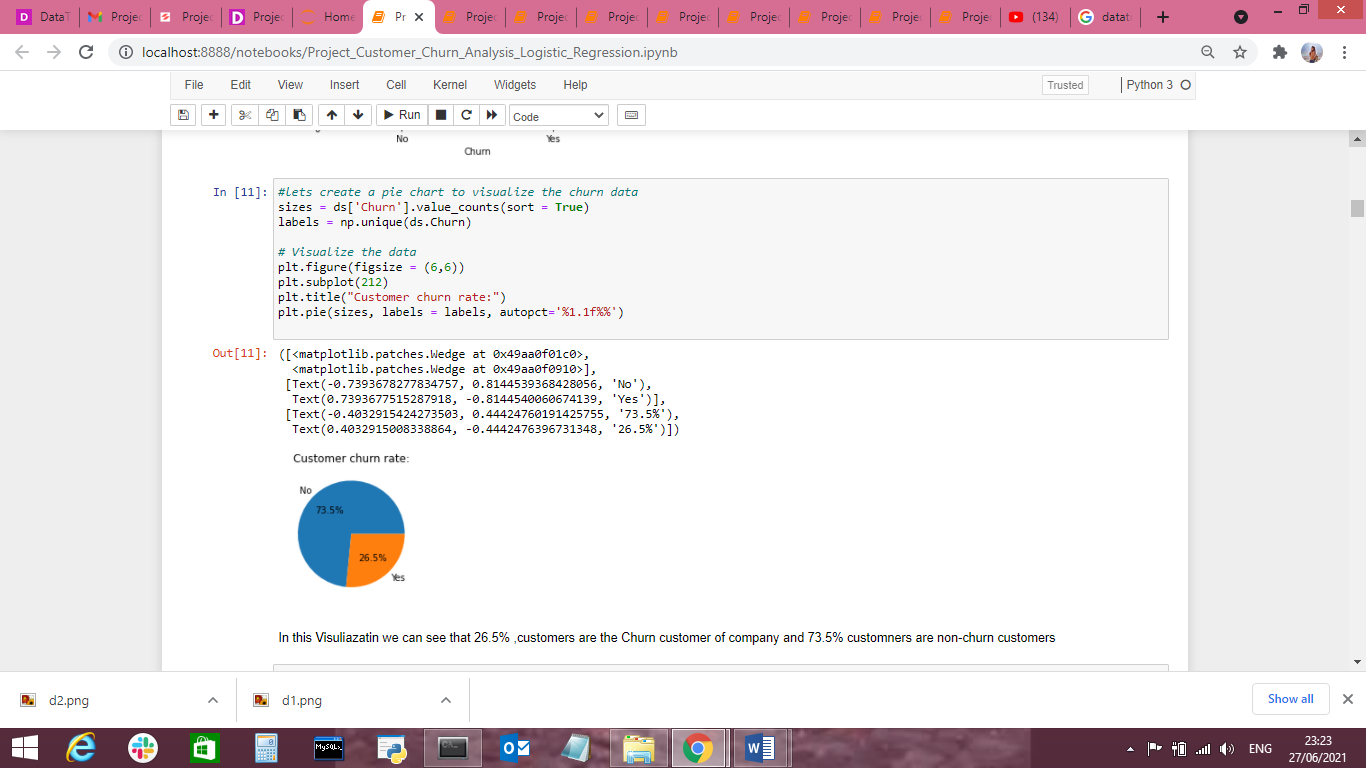
The dependent or target variable has 5174 (73.5% of data ) No’s and 1869 (26.5% of data) yes cases which can be seen below in the form of value counts and bar chart.



This means that 5174 customers were not churned (retained) and 1869 customers were churned.

We can even get the Churn rate from this by dividing no of customers lost in a given time interval by the total no of customers multiplied by 100.

In this case, about 73.46% are retained and the churn rate is 26.53%.



Comparing the Independent variables with the dependent variable (churn):

**Pre-processing Pipeline:**

The data set has variables in both object type and numerical type (int and float)

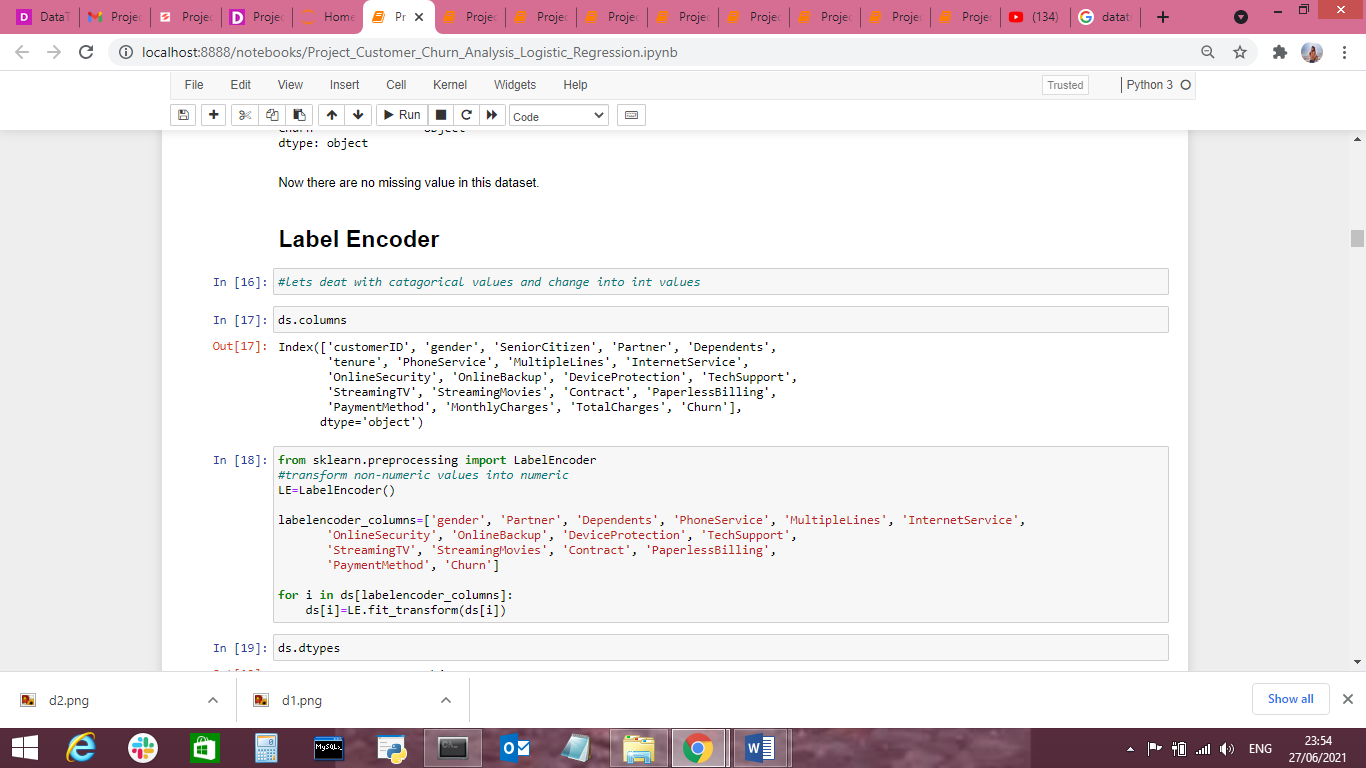
Therefore we have to pre-process the data to move forward.

All the float type or int type variables should be converted into the same scale since the range of values of raw data varies widely, in some machine learning algorithms, objective functions do not work correctly without normalization.

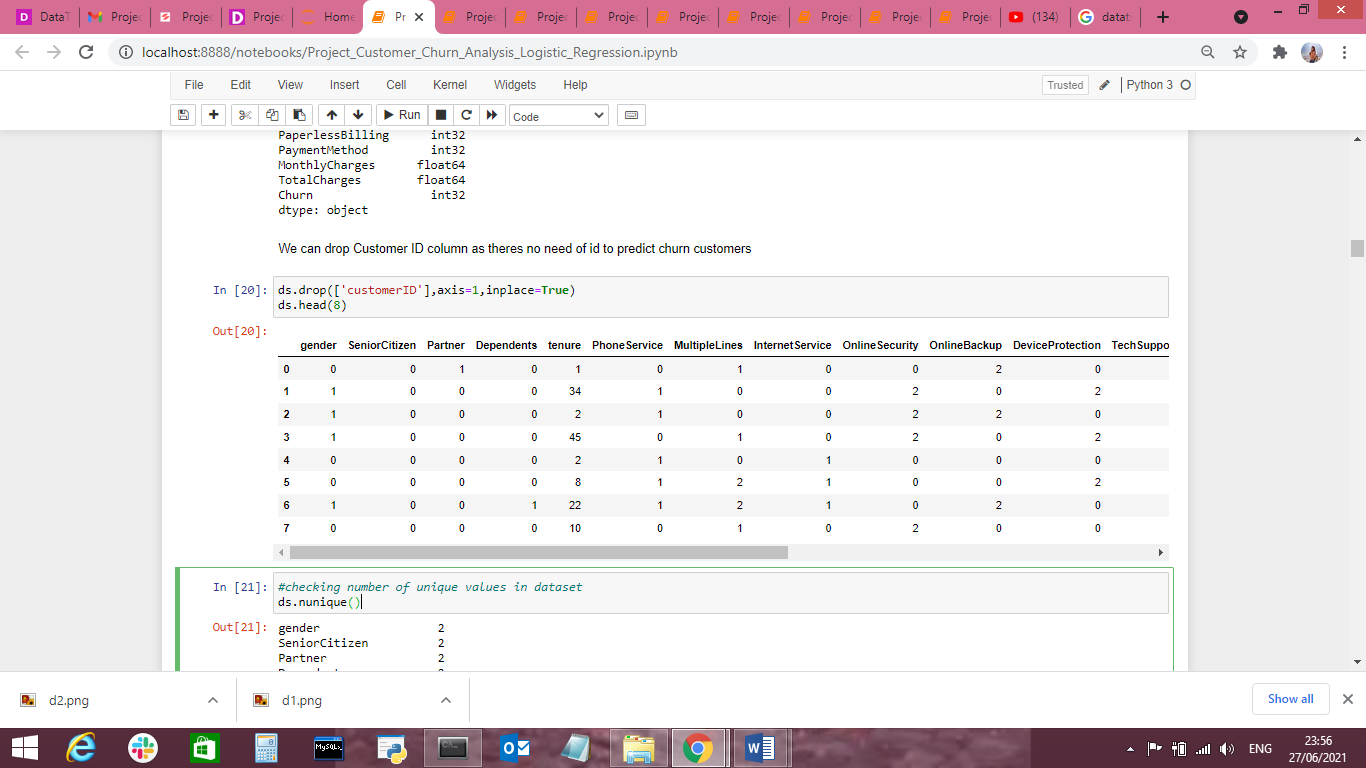
Therefore normalization is to be performed only on the numerical type (int and float type) variables.

We can see that majority of variables are object-type. These variables contain string data that cannot be passed into the machine learning model as it won’t be able to recognize string data type. It only recognizes numerical data.

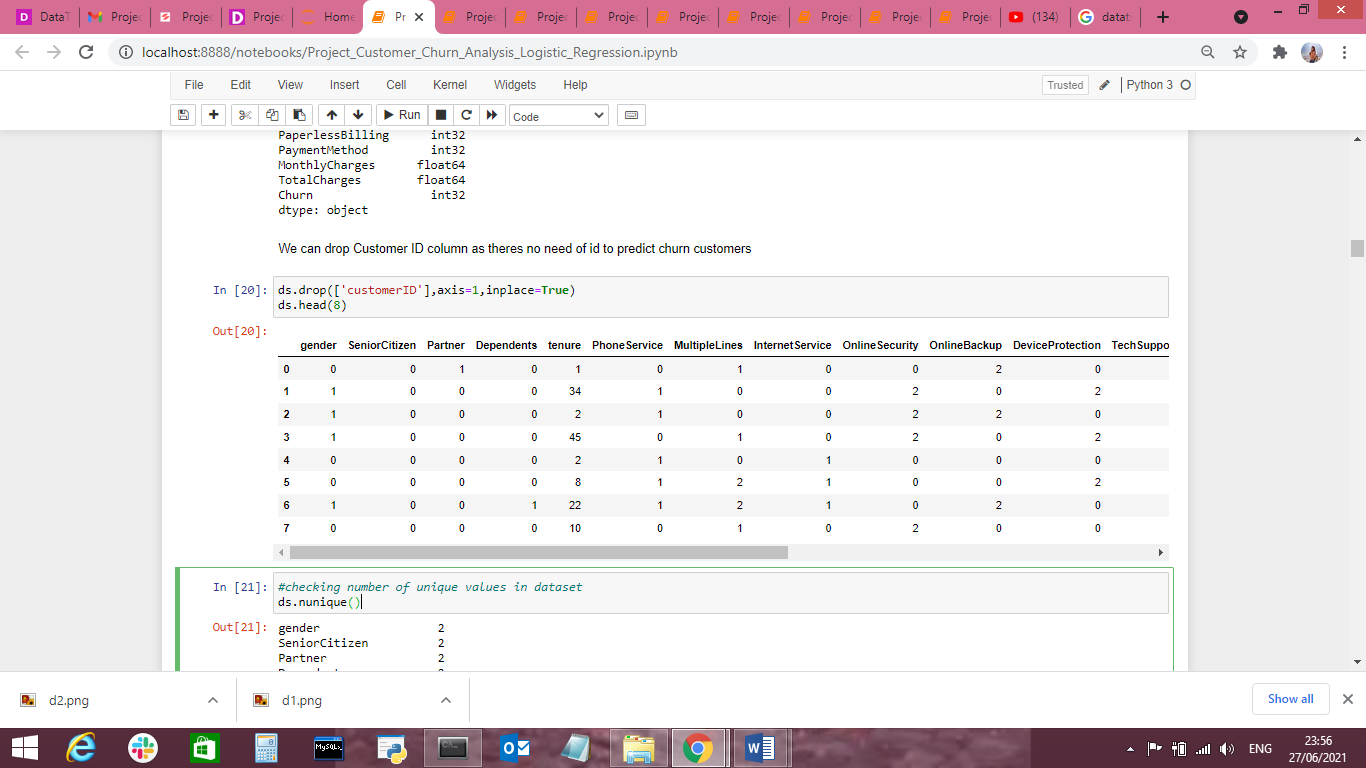
Therefore we need to convert the string data into numerical data. This can be done by manually encoding or by using an encoder such Label Encoder, one-hot encoder etc. For example: The target variable churn consists of only two unique values, Yes & No. after encoding this will get converted to 0 and 1. Similarly, if there are three unique values then it will be converted to 0,1, and 2.



We can drop Customer ID column as theres no need of id to predict churn customers



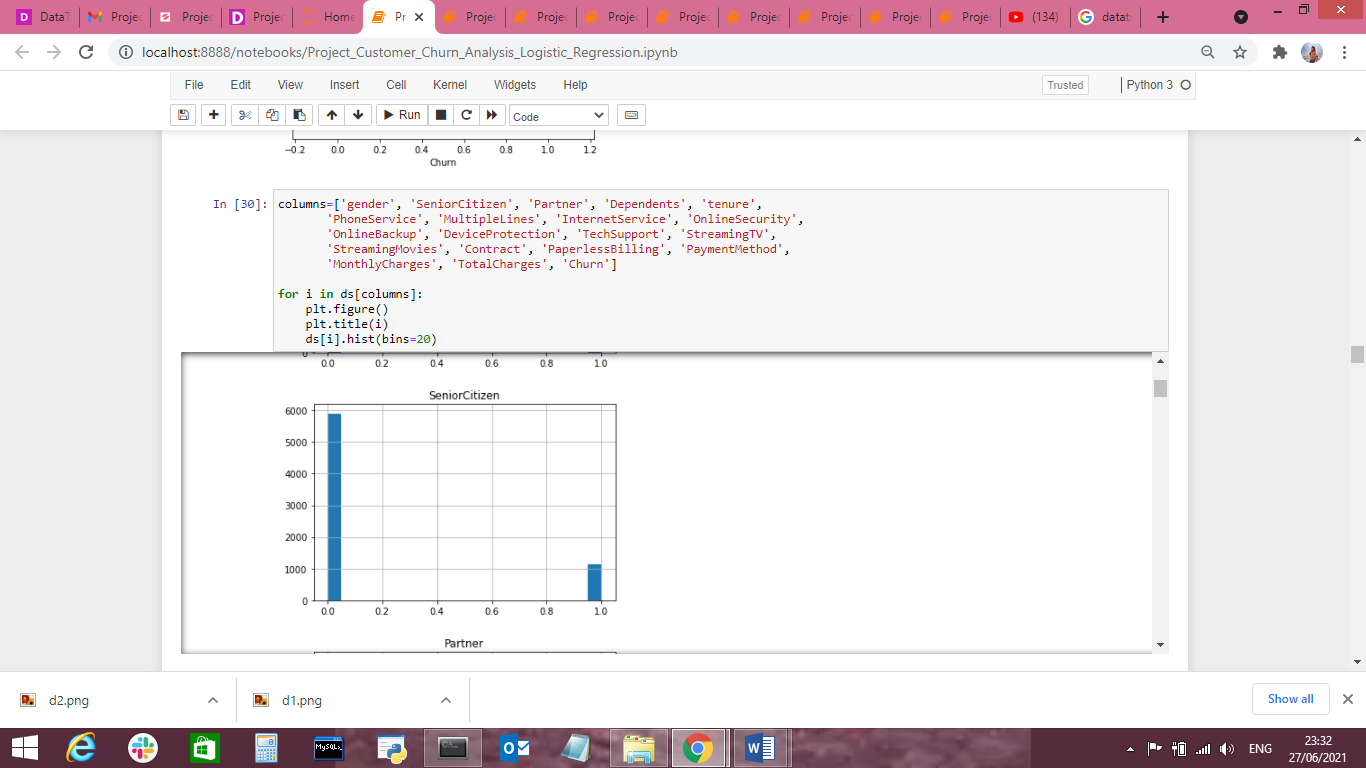
Sample data after preprocessing:



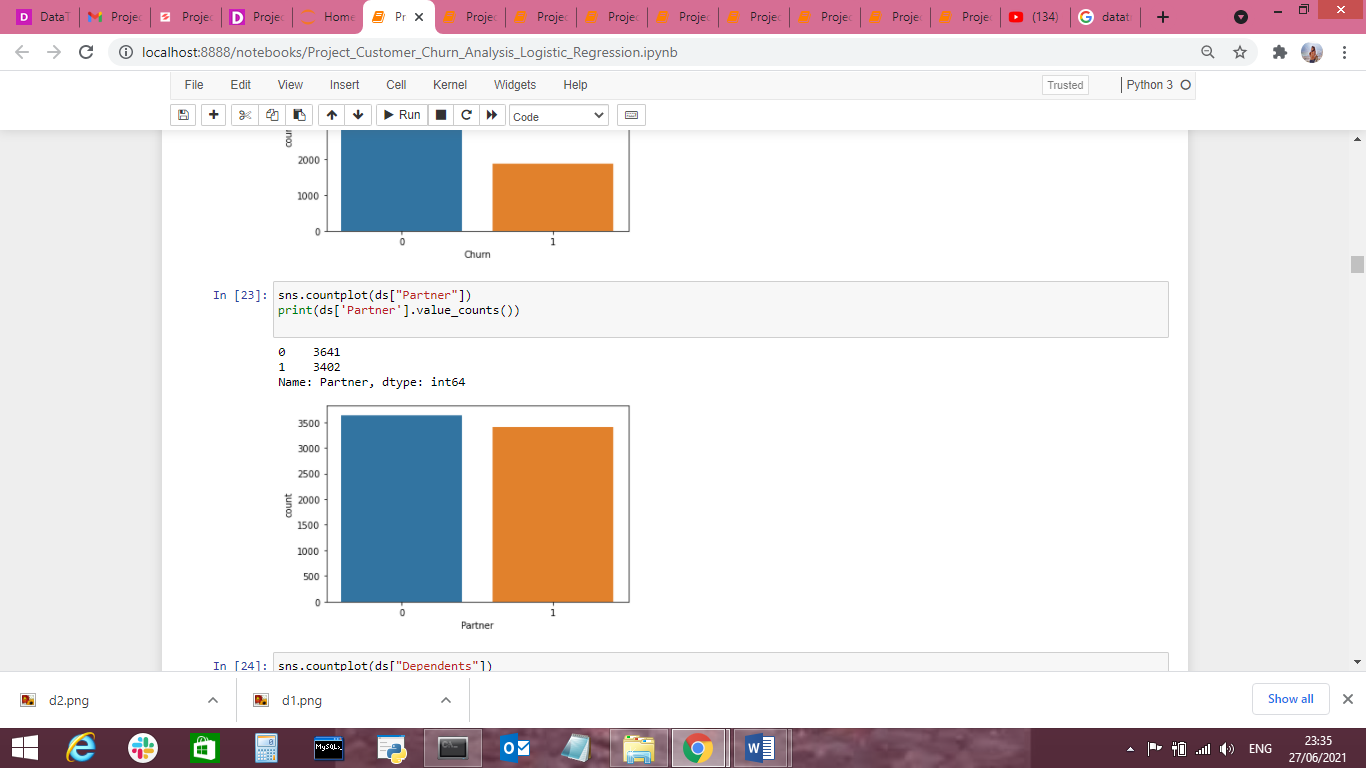
**EDA**



As shown above, we can see that there’s a very slight difference between the two genders, which says that gender doesn’t play a role in customer churn.

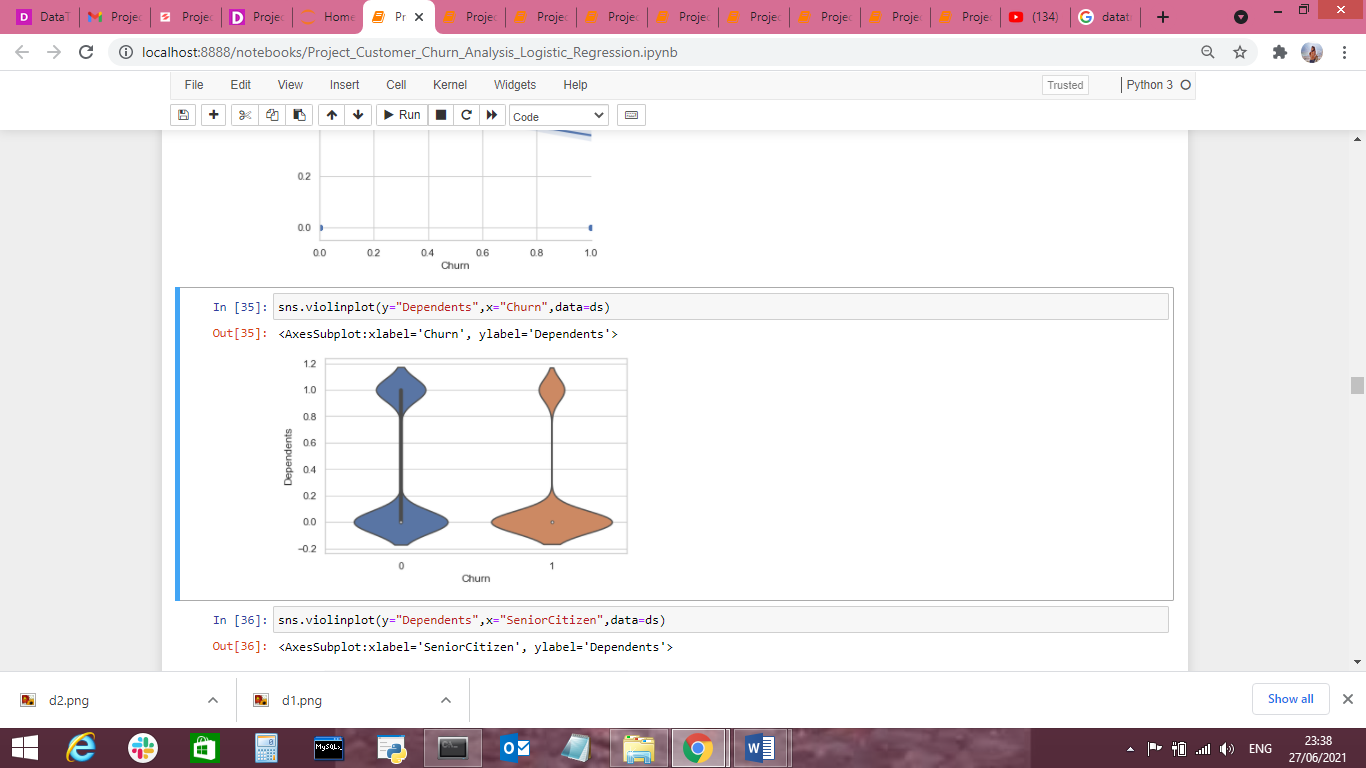


We can see that the churn is lesser if the customer is a senior citizen as they might have to start searching for a new company which would again take a lot of time and thus think it would be better to stick to the present company.

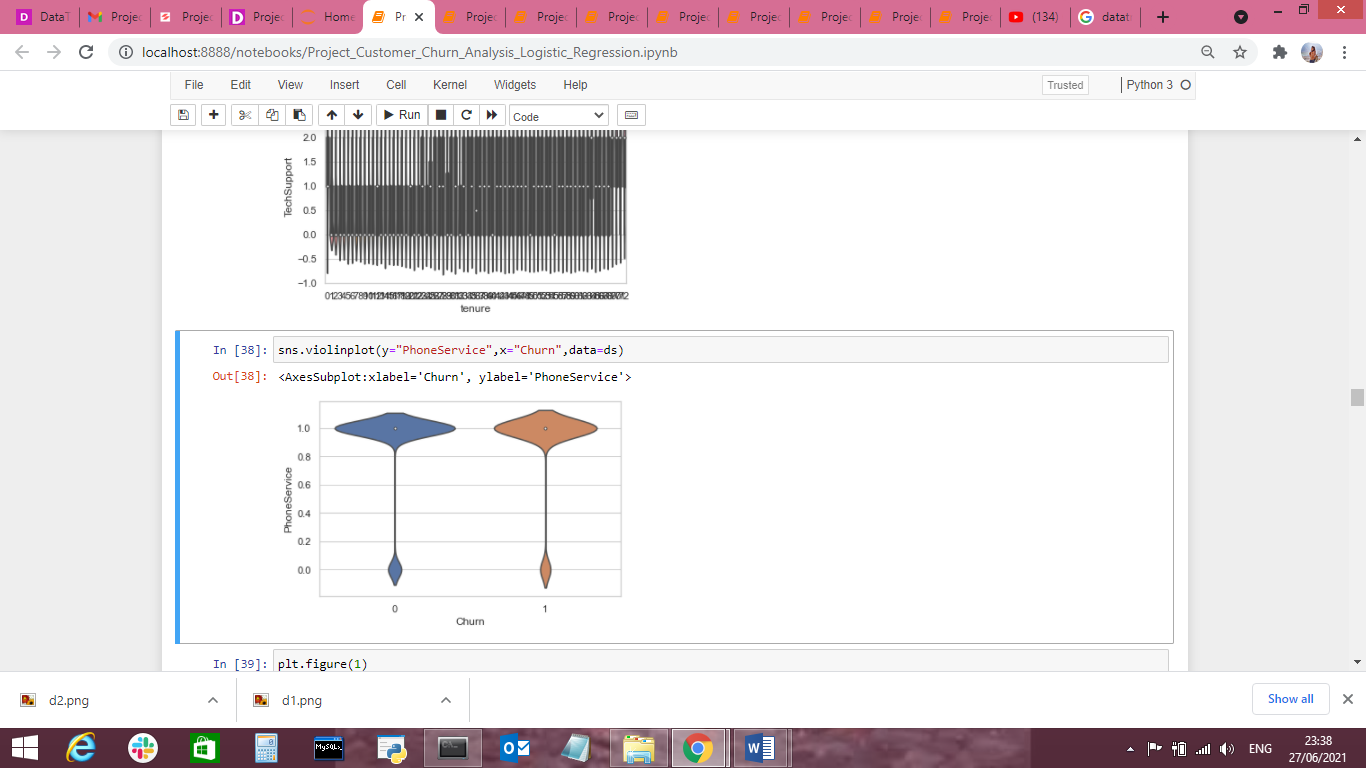


Churn Partner- No=3641, yes-3402

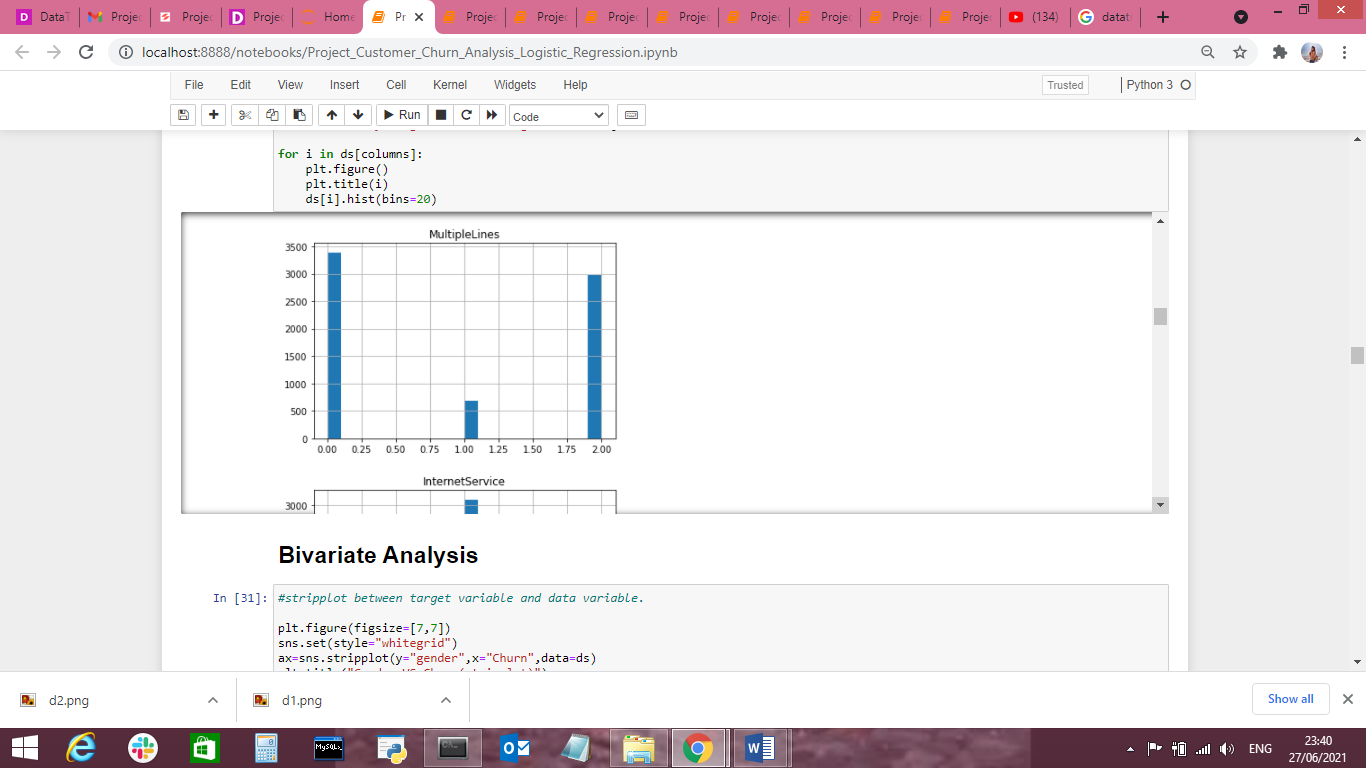
We can see that churn is less if there’s a partner. There might be instances where there would be one or more partners who would not agree to leave the company because of differences in opinion. That’s why maybe the churn is less where there’s a partner.



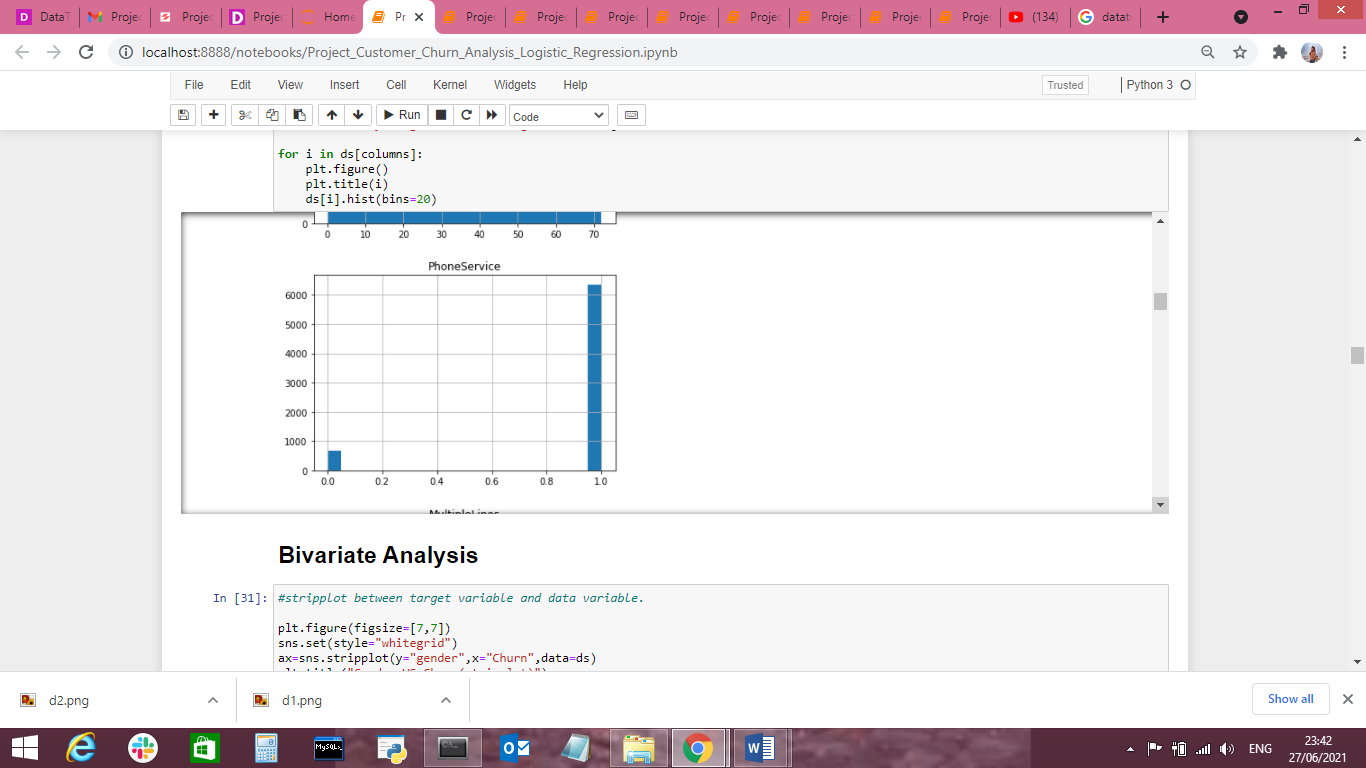
Here we can see that if there are dependents then the churn is less when compared to if there are no dependent. The reason might be similar to partner one.



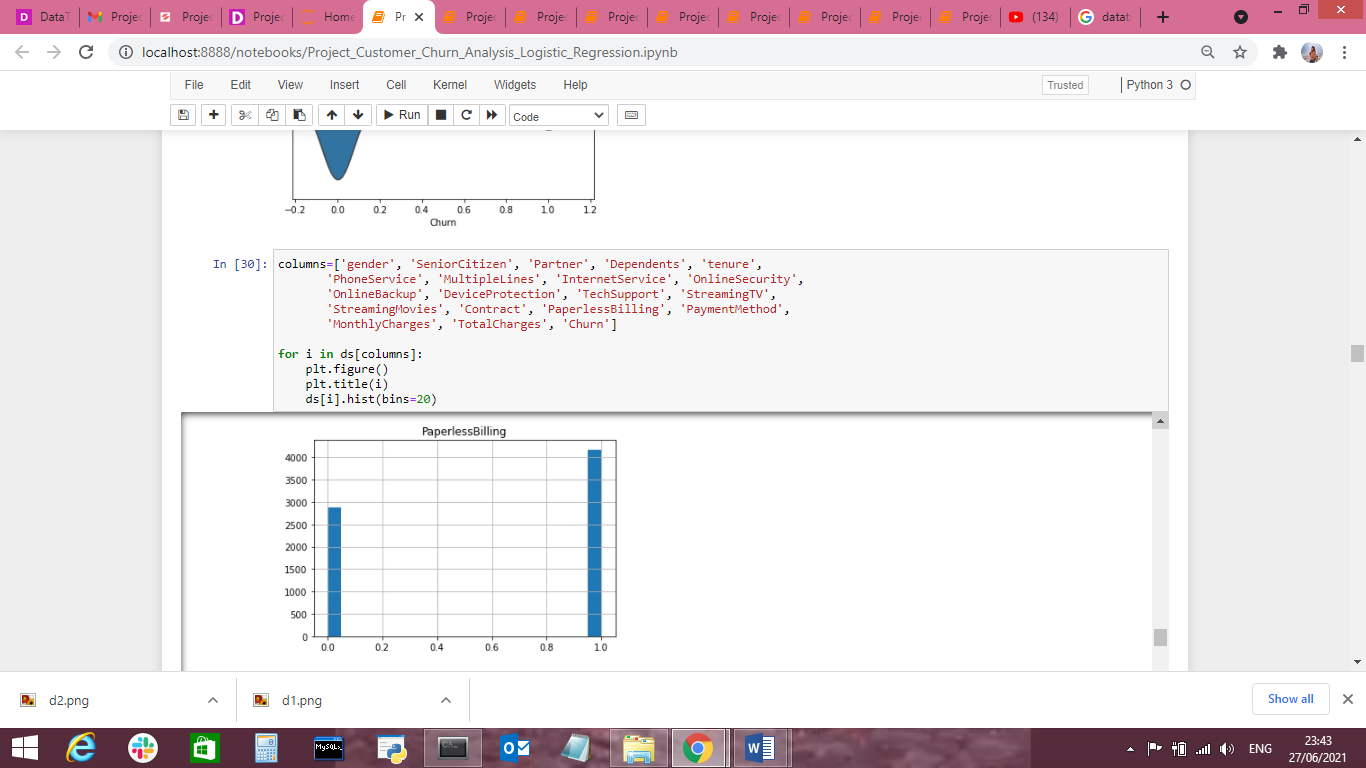
We can see that the churn is more if there’s a phone service.



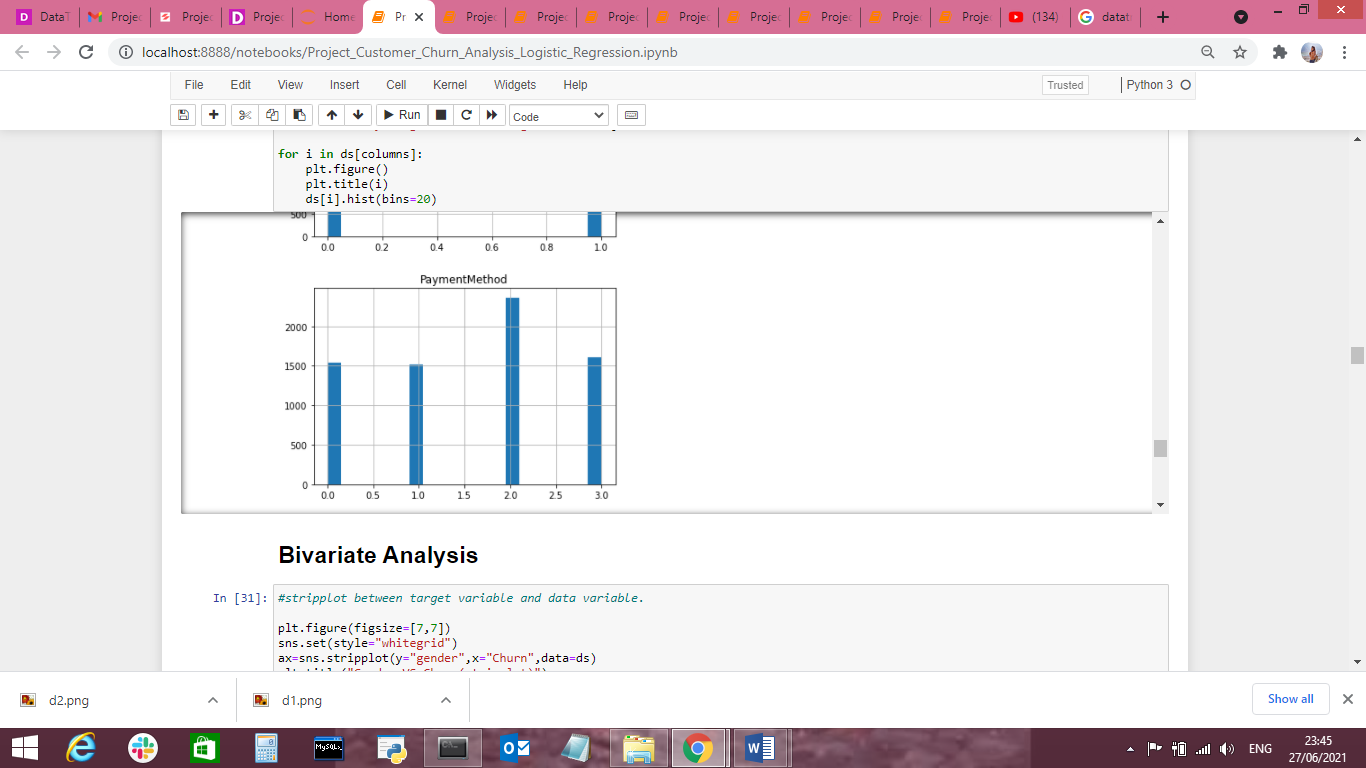
Here also we can see that the churn is less where there is no phone service or 1.



Here we can see that most customers that churned had the Fiberoptic internet service or 0. Maybe the company Should have only DSL internet service or 1.



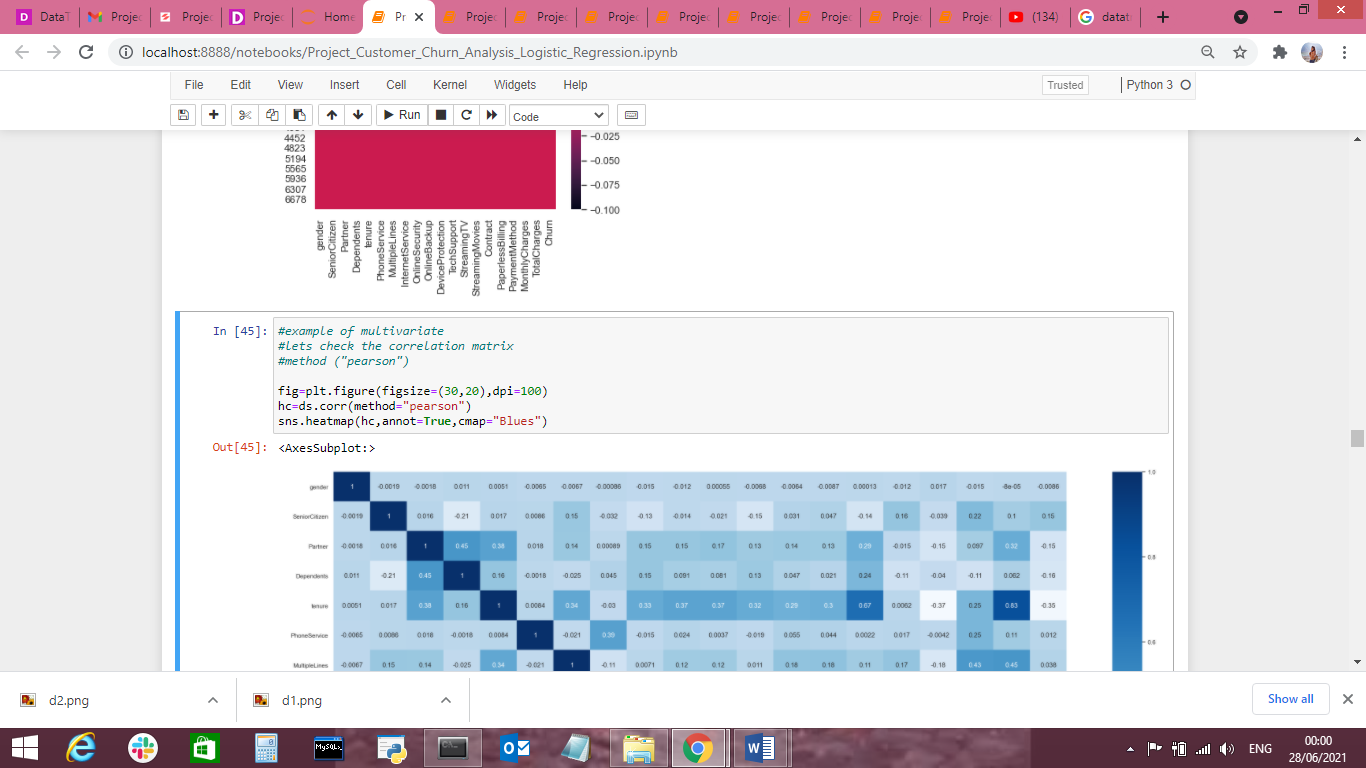
Here we can see that most customers that churned had the Paperlessbilling.

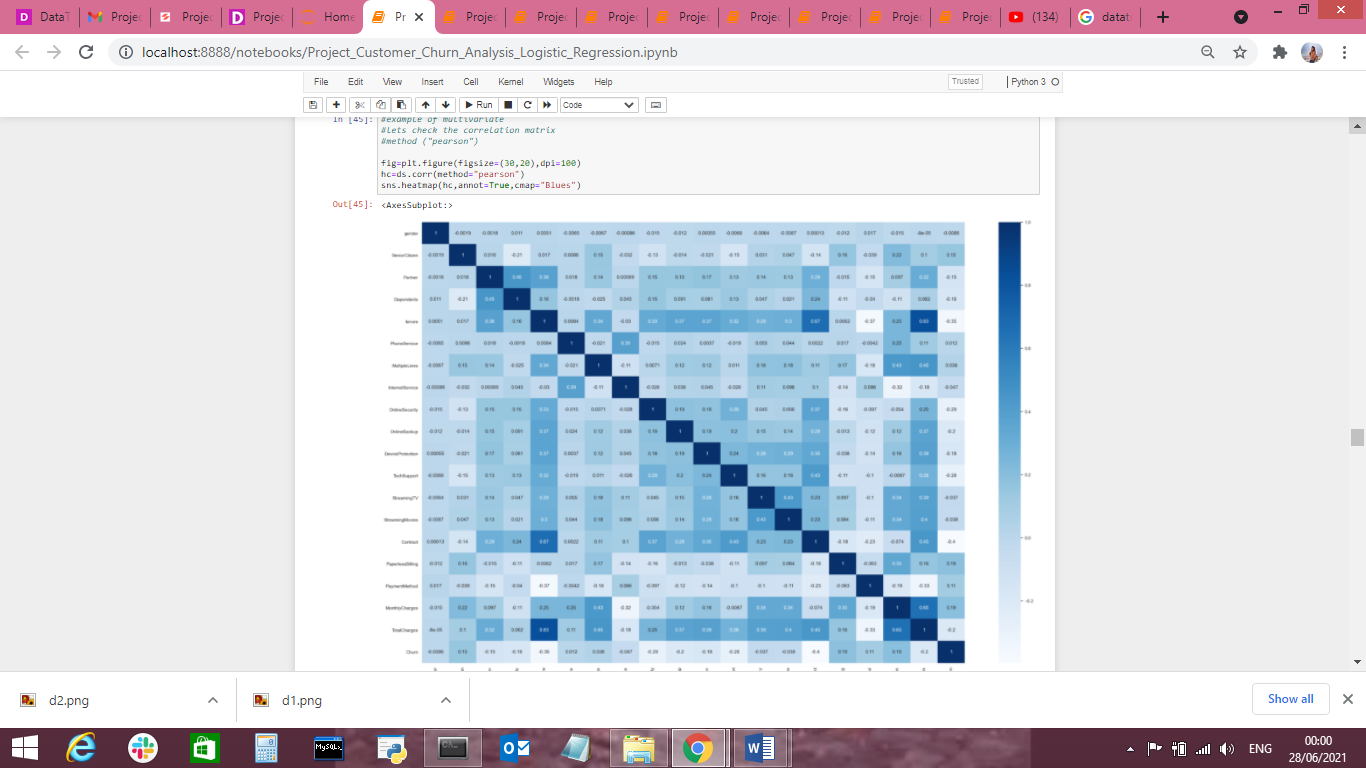


Here we can see that most customers that churned are having electronic check payment method or 2.

**Correlation between ‘Churn' and 'Independent features'**

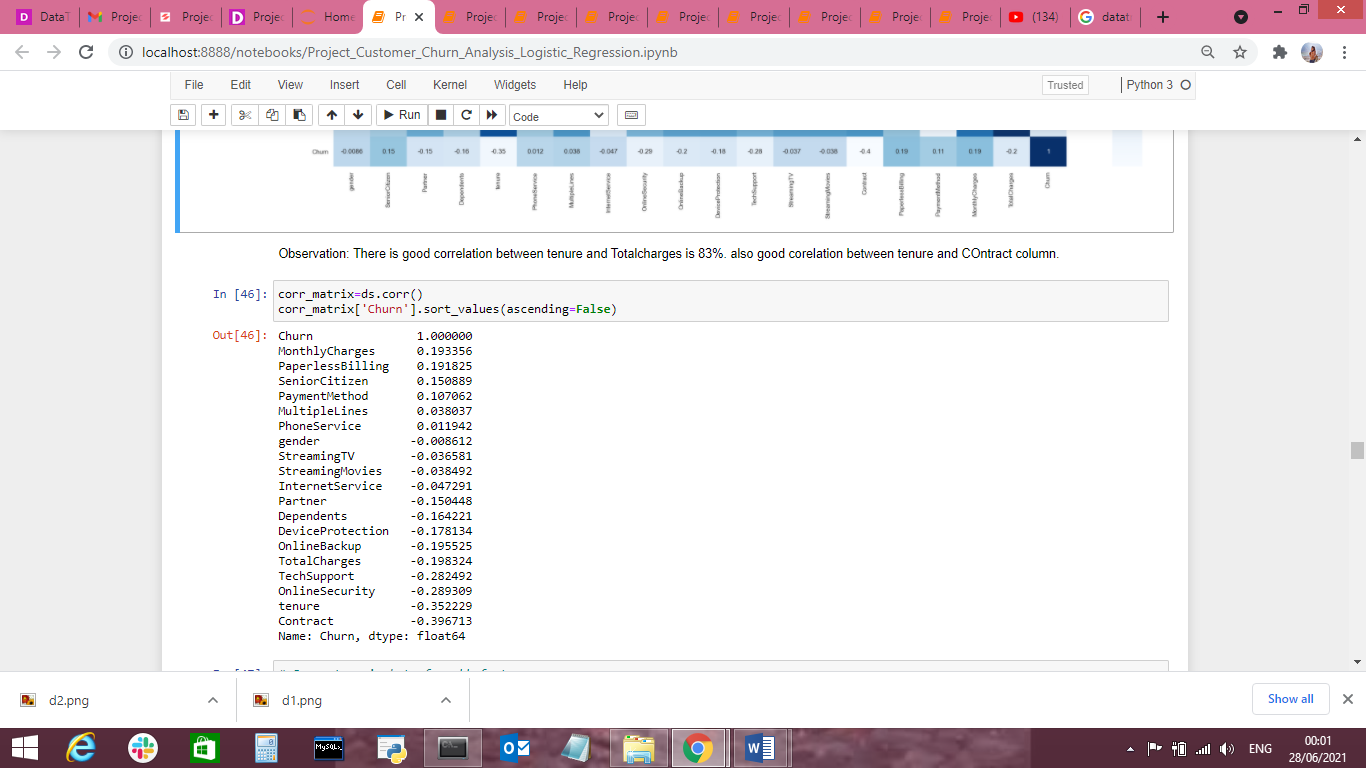
Now we can check the correlation between all the variables. (Note: correlation of all independent variables can be only done after encoding as correlation does not consider string values)





We can see that correlation between independent variables is low(i.e. <0.7). We are good to go.

From the figure below we can see the important features in descending order from top to bottom.

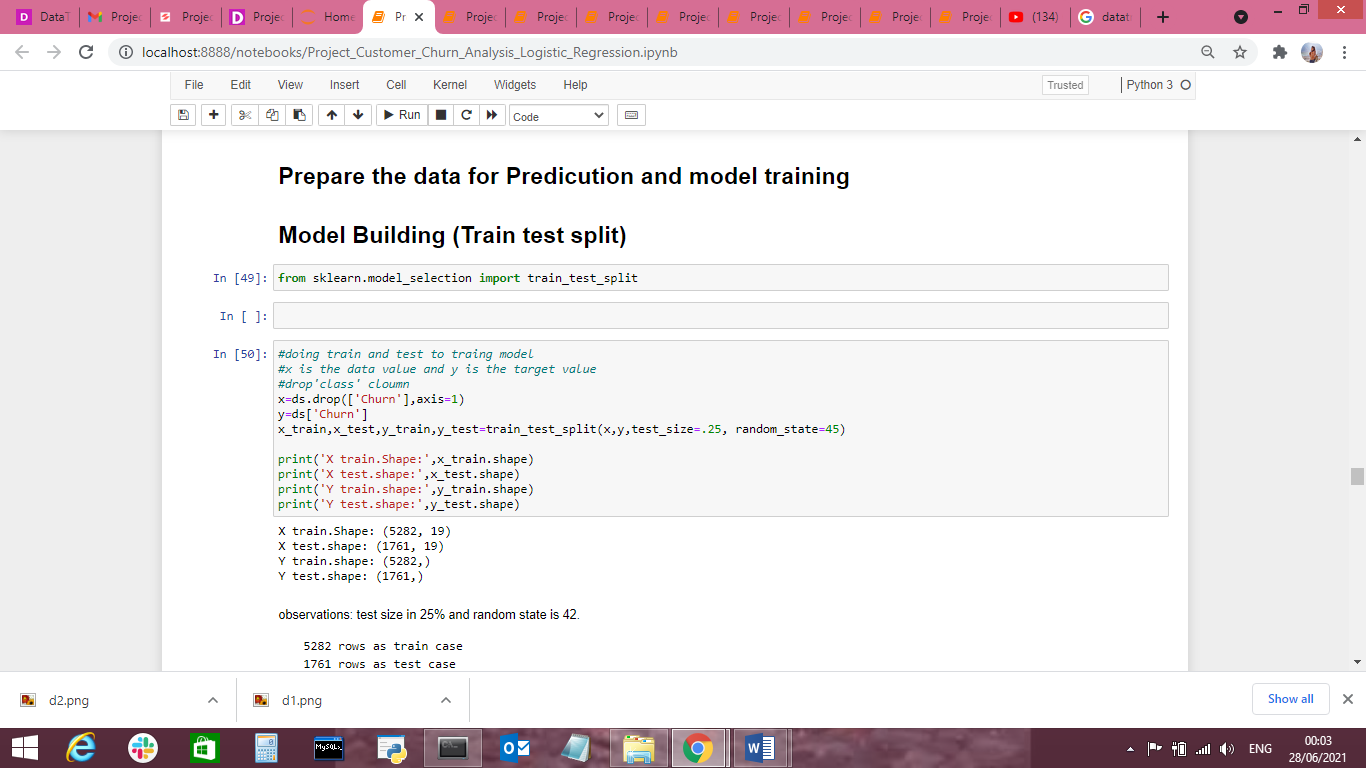


We can see that MonthlyCharger and PaperlessBilling are the top two important features for the taget variable ‘churn’

Now the preprocessing is completed. We now have to move to data modeling and prediction

**Building Machine Learning Models:**

We have to now split the data into independent and target variables.



Here the target variable is ‘Churn’ and the rest of them are independent variables.

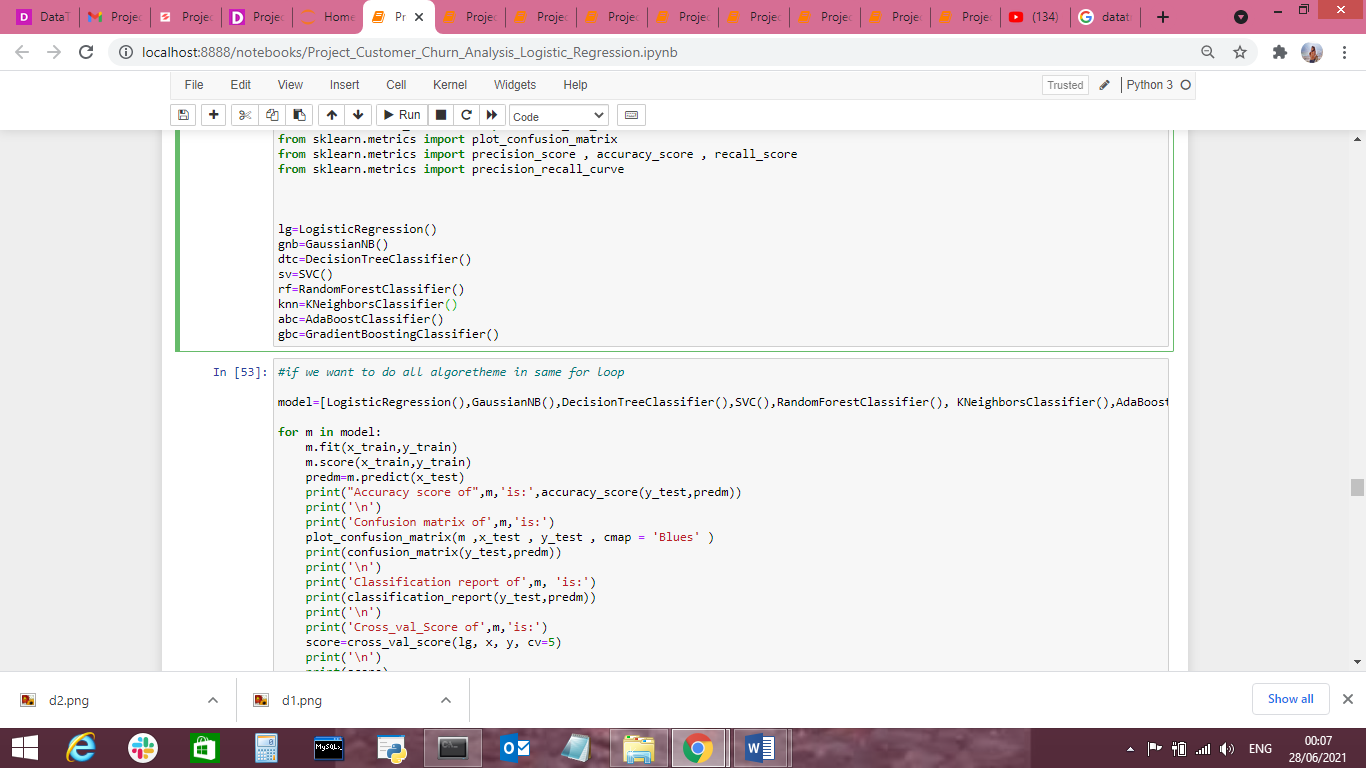
We have to now split the independent and target variables into training and testing datasets as shown below.

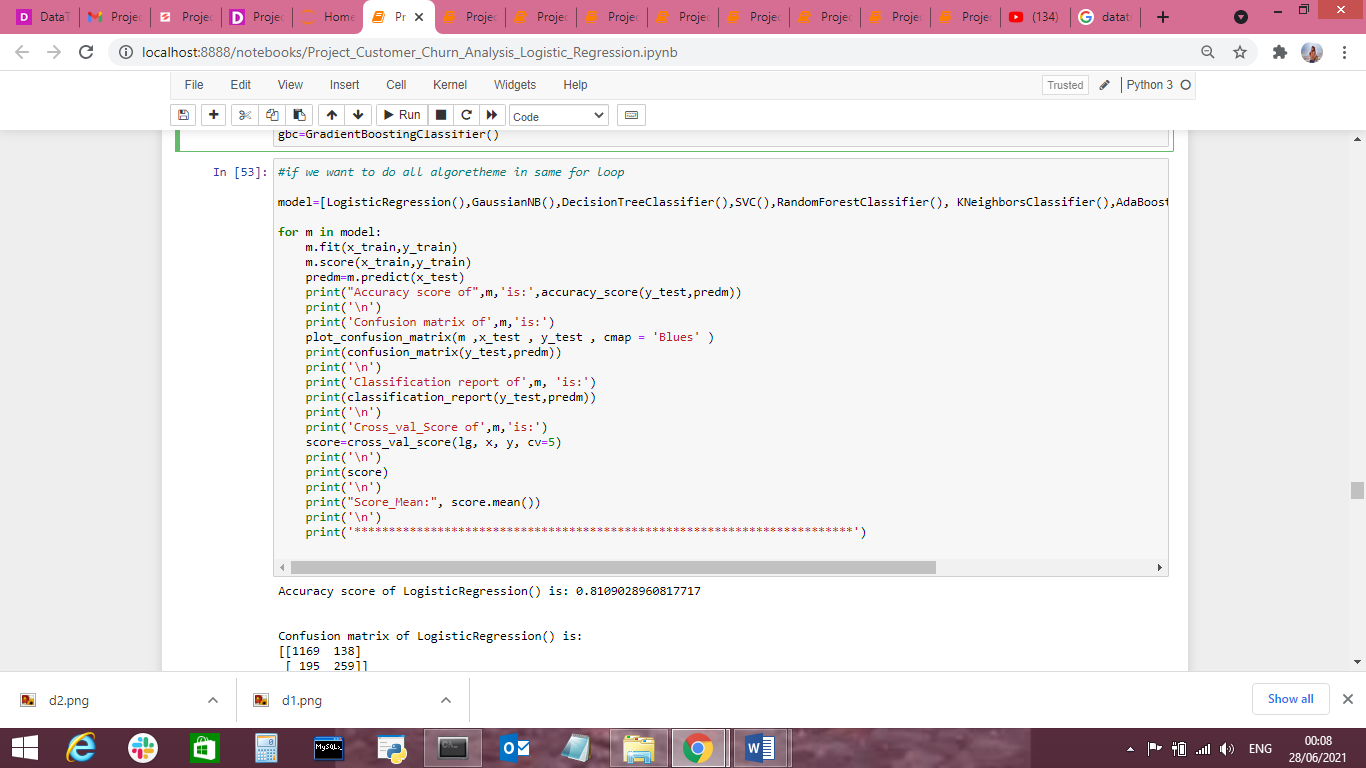
We will use a machine-learning algorithm to learn from the training set and use the model to predict the testing set and compare it with the predicted data with the target testing set to know how close the values. If the error between the predicted and target testing data is less that means the accuracy of the model is high and we can use this model to predict the result of similar datasets.

In this, we have used 8 Machine learning Algorithms

* LogisticRegression
* GaussianNB
* DecisionTreeClassifier
* Supportvectorclassifier
* RandomForestClassifier
* KNeighborsClassifier
* AdaBoostClassifier
* GradientBoostingClassifier

We can train and predict the data using the above these 8 ML algorithms and save the model which has the highest frequency.





Accuracy score of LogisticRegression() is: 0.8109028960817717

Confusion matrix of LogisticRegression() is:

[[1169 138]

[ 195 259]]

Classification report of LogisticRegression() is:

precision recall f1-score support

0 0.86 0.89 0.88 1307

1 0.65 0.57 0.61 454

accuracy 0.81 1761

macro avg 0.75 0.73 0.74 1761

weighted avg 0.80 0.81 0.81 1761

Cross\_val\_Score of LogisticRegression() is:

[0.80837473 0.79985806 0.79276082 0.81107955 0.79971591]

Score\_Mean: 0.8023578134073166

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Accuracy score of GaussianNB() is: 0.7569562748438388

Confusion matrix of GaussianNB() is:

[[983 324]

[104 350]]

Classification report of GaussianNB() is:

precision recall f1-score support

0 0.90 0.75 0.82 1307

1 0.52 0.77 0.62 454

accuracy 0.76 1761

macro avg 0.71 0.76 0.72 1761

weighted avg 0.81 0.76 0.77 1761

Cross\_val\_Score of GaussianNB() is:

[0.80837473 0.79985806 0.79276082 0.81107955 0.79971591]

Score\_Mean: 0.8023578134073166

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Accuracy score of DecisionTreeClassifier() is: 0.7336740488358887

Confusion matrix of DecisionTreeClassifier() is:

[[1074 233]

[ 236 218]]

Classification report of DecisionTreeClassifier() is:

precision recall f1-score support

0 0.82 0.82 0.82 1307

1 0.48 0.48 0.48 454

accuracy 0.73 1761

macro avg 0.65 0.65 0.65 1761

weighted avg 0.73 0.73 0.73 1761

Cross\_val\_Score of DecisionTreeClassifier() is:

[0.80837473 0.79985806 0.79276082 0.81107955 0.79971591]

Score\_Mean: 0.8023578134073166

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Accuracy score of SVC() is: 0.8006814310051107

Confusion matrix of SVC() is:

[[1189 118]

[ 233 221]]

Classification report of SVC() is:

precision recall f1-score support

0 0.84 0.91 0.87 1307

1 0.65 0.49 0.56 454

accuracy 0.80 1761

macro avg 0.74 0.70 0.71 1761

weighted avg 0.79 0.80 0.79 1761

Cross\_val\_Score of SVC() is:

[0.80837473 0.79985806 0.79276082 0.81107955 0.79971591]

Score\_Mean: 0.8023578134073166

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Accuracy score of RandomForestClassifier() is: 0.8018171493469619

Confusion matrix of RandomForestClassifier() is:

[[1184 123]

[ 226 228]]

Classification report of RandomForestClassifier() is:

precision recall f1-score support

0 0.84 0.91 0.87 1307

1 0.65 0.50 0.57 454

accuracy 0.80 1761

macro avg 0.74 0.70 0.72 1761

weighted avg 0.79 0.80 0.79 1761

Cross\_val\_Score of RandomForestClassifier() is:

[0.80837473 0.79985806 0.79276082 0.81107955 0.79971591]

Score\_Mean: 0.8023578134073166

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Accuracy score of KNeighborsClassifier() is: 0.7484383872799546

Confusion matrix of KNeighborsClassifier() is:

[[1100 207]

[ 236 218]]

Classification report of KNeighborsClassifier() is:

precision recall f1-score support

0 0.82 0.84 0.83 1307

1 0.51 0.48 0.50 454

accuracy 0.75 1761

macro avg 0.67 0.66 0.66 1761

weighted avg 0.74 0.75 0.75 1761

Cross\_val\_Score of KNeighborsClassifier() is:

[0.80837473 0.79985806 0.79276082 0.81107955 0.79971591]

Score\_Mean: 0.8023578134073166

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Accuracy score of AdaBoostClassifier() is: 0.8052243043725156

Confusion matrix of AdaBoostClassifier() is:

[[1182 125]

[ 218 236]]

Classification report of AdaBoostClassifier() is:

precision recall f1-score support

0 0.84 0.90 0.87 1307

1 0.65 0.52 0.58 454

accuracy 0.81 1761

macro avg 0.75 0.71 0.73 1761

weighted avg 0.80 0.81 0.80 1761

Cross\_val\_Score of AdaBoostClassifier() is:

[0.80837473 0.79985806 0.79276082 0.81107955 0.79971591]

Score\_Mean: 0.8023578134073166

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Accuracy score of GradientBoostingClassifier() is: 0.8086314593980692

Confusion matrix of GradientBoostingClassifier() is:

[[1187 120]

[ 217 237]]

Classification report of GradientBoostingClassifier() is:

precision recall f1-score support

0 0.85 0.91 0.88 1307

1 0.66 0.52 0.58 454

accuracy 0.81 1761

macro avg 0.75 0.72 0.73 1761

weighted avg 0.80 0.81 0.80 1761

Cross\_val\_Score of GradientBoostingClassifier() is:

[0.80837473 0.79985806 0.79276082 0.81107955 0.79971591]

Score\_Mean: 0.8023578134073166

**Cross Validation:**

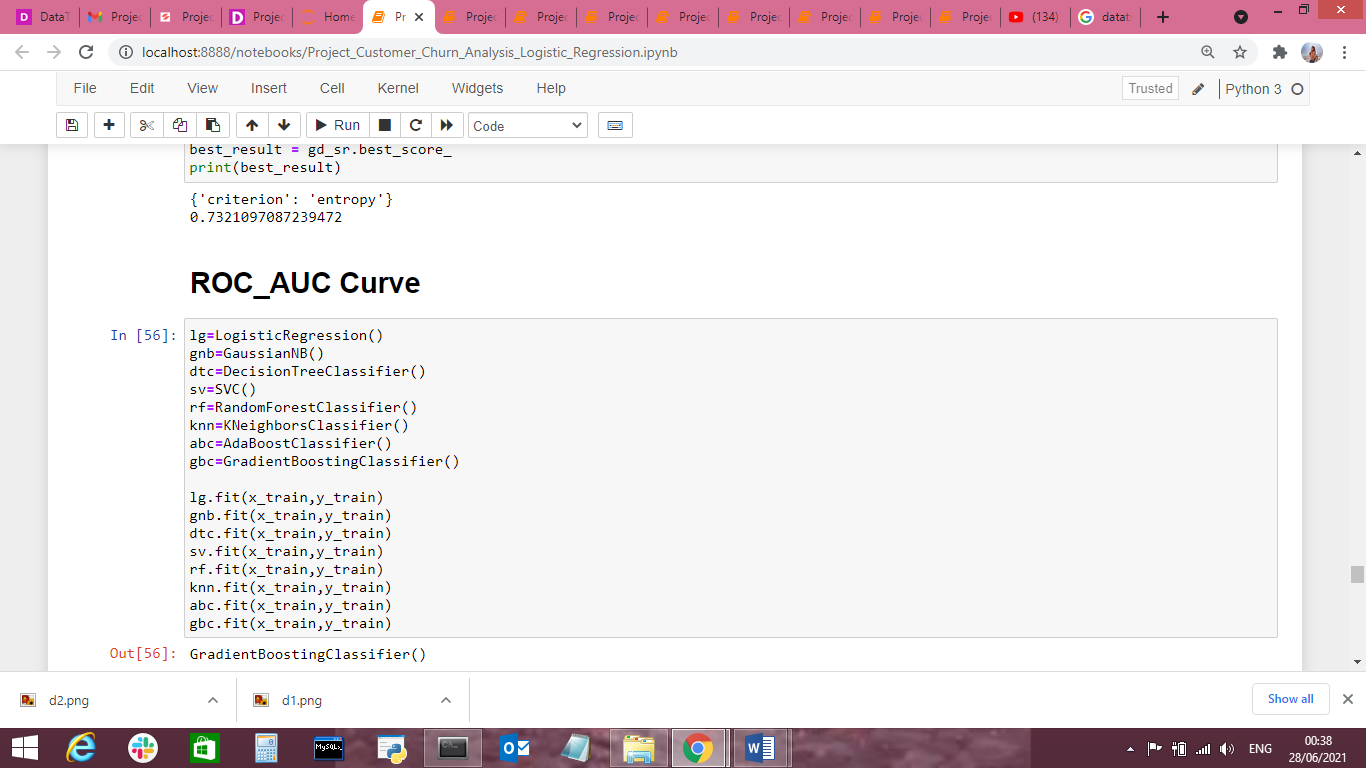
According to Cross val score and accuracy we can see that the Logistic Regression and GradientBoostingClassifier has the least difference between Accuracy and Cross val score, therefore we select Logistic Regression model.

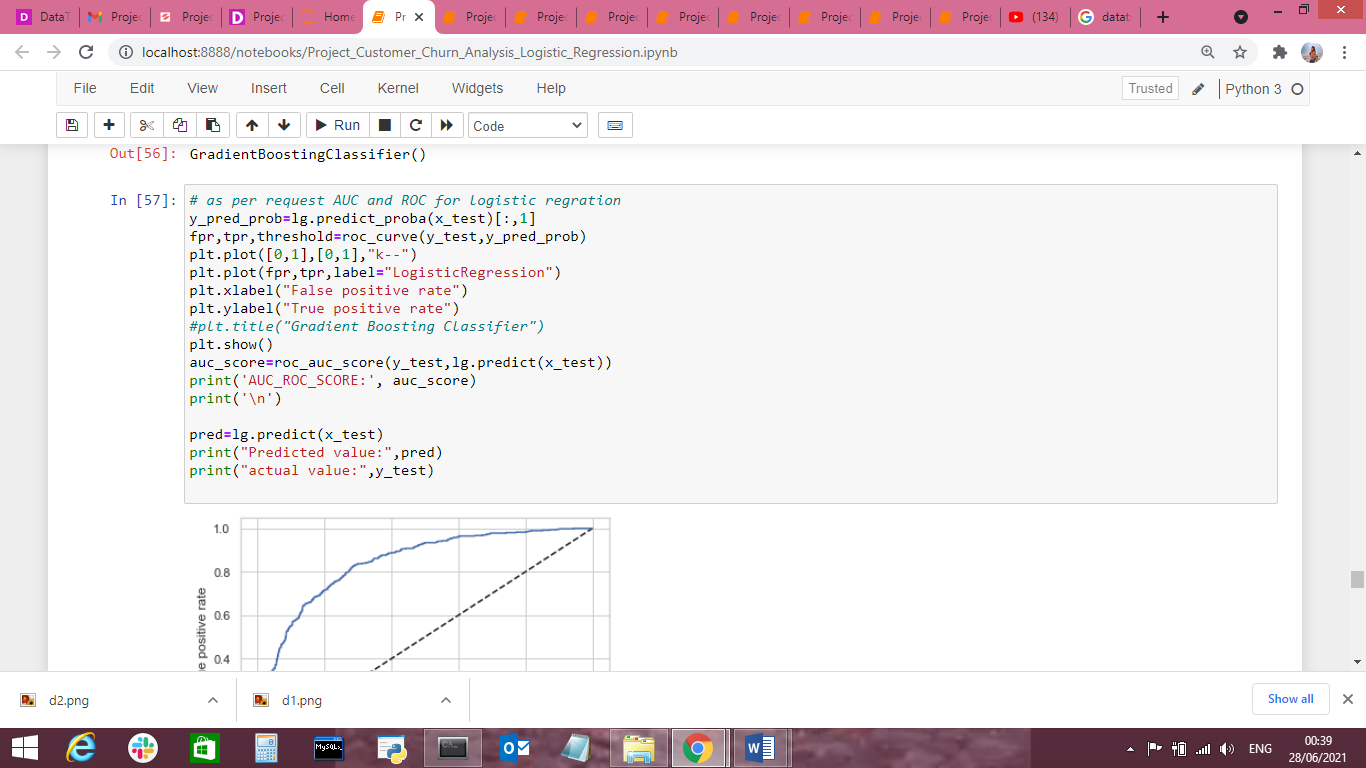
# AUC ROC CURVE:

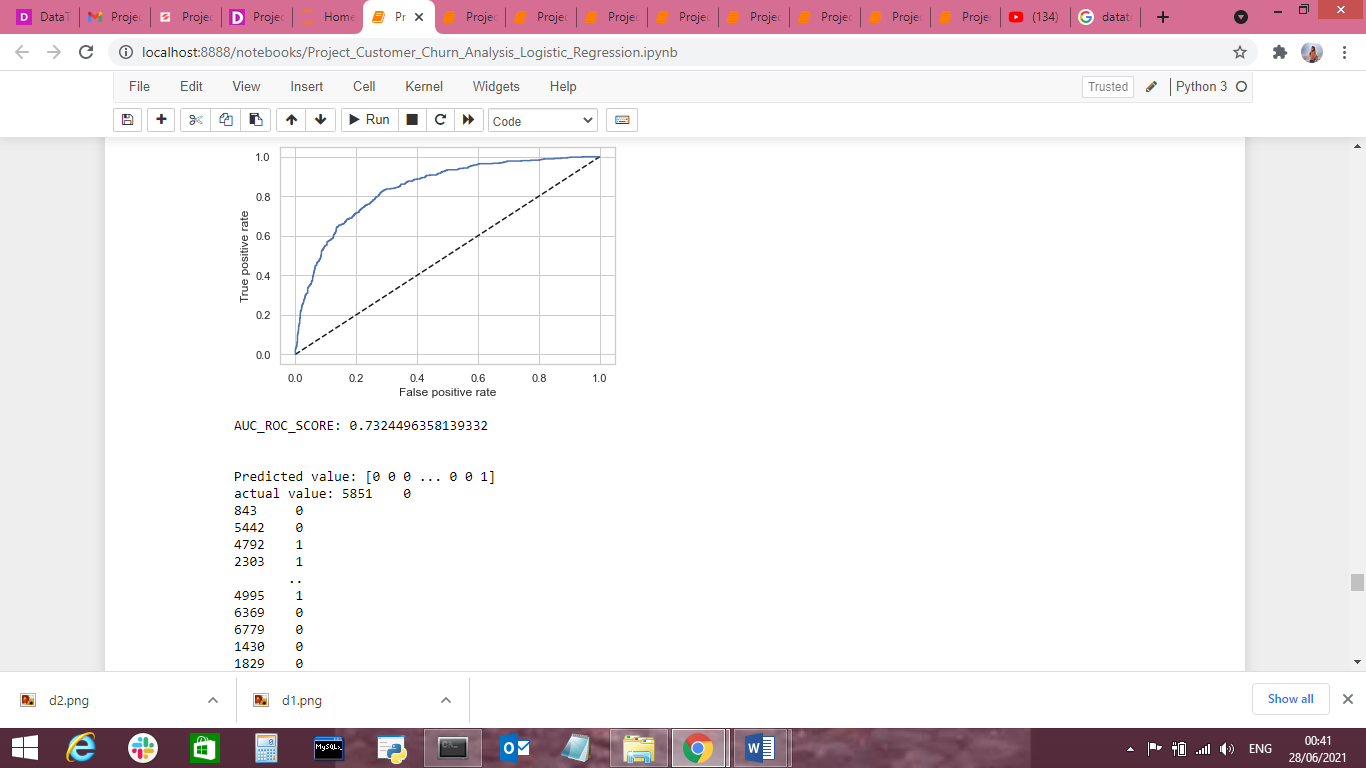
# AUC: Area Under the curve;   ROC: Receiver Operator Characteristic

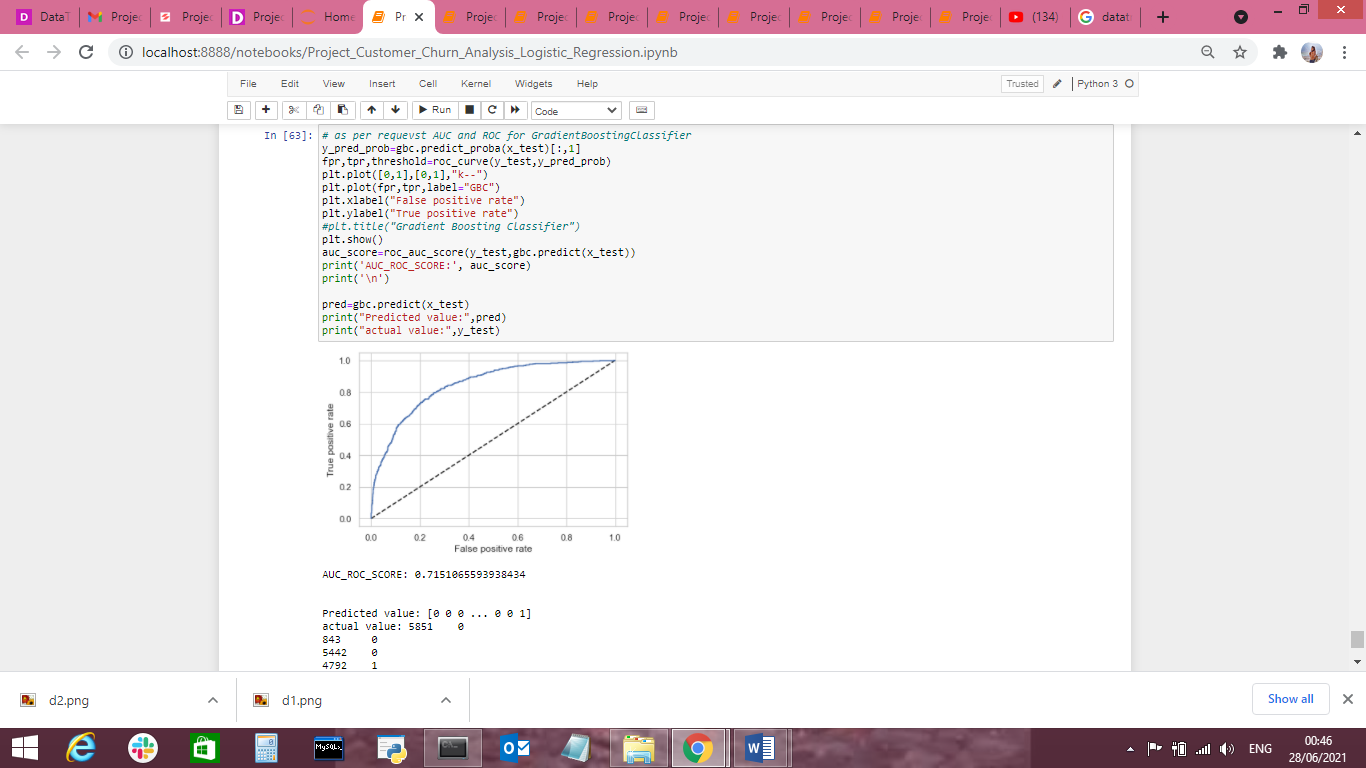
# The greater the ROC score the better is the model. If ROC=1, then it perfectly fits.

# If the maximum of the area falls under True positive then the model is doing good.



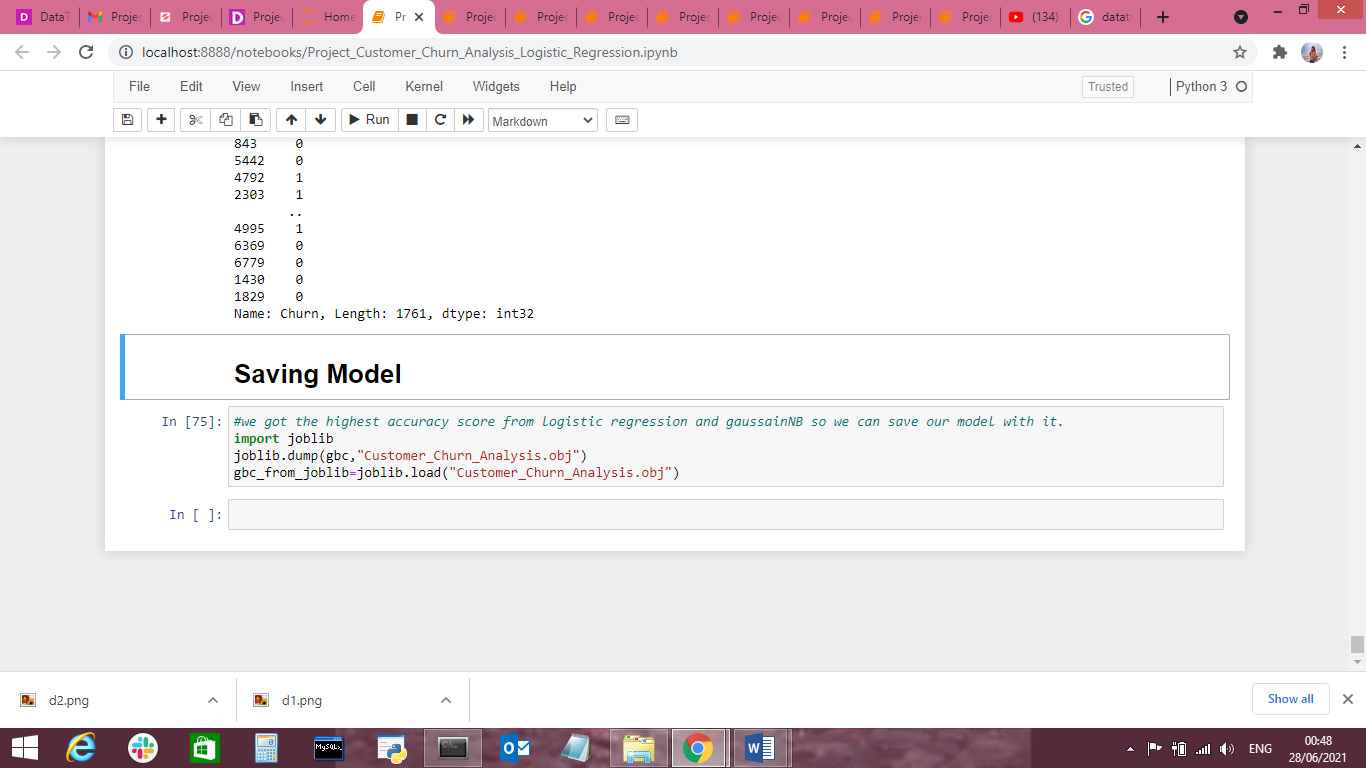






We can see that the ROC Score is lg=0.73 & gbc=0.71.5 and the area under the curve falls under True Positive Rate, Therefore, we can conclude that the model is performing well and we need to save the model in .obj file for future use.

By using ‘dump’ from joblib we can the best model (LogisticRegression in our case) in .obj as by doing this we can save the model and by using ‘load’ function we can load the model and predict the model with a different dataset.



**Concluding Remarks:**

From the above results of the data modeling and prediction we can see that the Decision Tree Model is performing well as the accuracy score, cross val score and Roc score are good also the maximum of the area under the curve fall under true positive rate. Therefore we can save the model as .obj file so that it can be used to predict the result of the different data sets.

In this kind of problems Pre-processing and data-cleaning is the most important thing. We need to handle both the categorical and numerical data properly and also need to check by building different ML model on the same dataset. We need to check the accuracy and cross val score of each model and chose the one which has the best of the same and also which has the least difference between them (i.e cross\_val\_score and accuracy).

We could see that there is no impact of gender on the churn rate. Also the company must avoid using phone service, Paperless billings, electronic check payment method.

By using this model many companies can find their mistakes and improve which will lead to financial gain.

**Thanks to Visit My Blog**