**Capstone Project**

**Telco Customer Churn Prediction Analysis**

A red figure with a black background

Description automatically generated with low confidence

**Background pattern

Description automatically generated**

PGA-01 Data Analytics and Machine Learning

**(2021-2022)**

**-Submitted by**

**Suganya Janardhanan**

**ABSTRACT:**

Customer churn is the loss of clients or customers for a company. Churn is an important business metric for subscription-based services such as telecommunication companies. For such companies to improve their revenue it is important that they predict the churn. Now all the churn data and various factors associated with churn when quantified in the form of electronic records, can be used to uncover trends and associations. Machine learning can use this data to predict customer churn of a company and identify the most important features among them.

So, in this project first we try to explore the dataset and then build model using various machine learning algorithms such as logistic regression, random forest classifier, adaboost, svc on the data set and try to figure out which algorithm, would turn to be a best fit for the model giving least errors while predicting. The coding for this project has been done using python.

**DATA DEFINITION:**

* **Customers who left** – the column is called Churn
* **Services that each customer has signed up for** – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
* **Customer account information** - how long they’ve been a customer, contract, payment method, paperless billing, monthly charges, and total charges
* **Demographic info about customers** – gender, and if they have partners and dependents

**TABLE OF CONTENTS:**

Graphical user interface, text, application

Description automatically generated

**IMPORTING LIBRARIES:**

Imported necessary libraries which are needed for the analysis and model building.

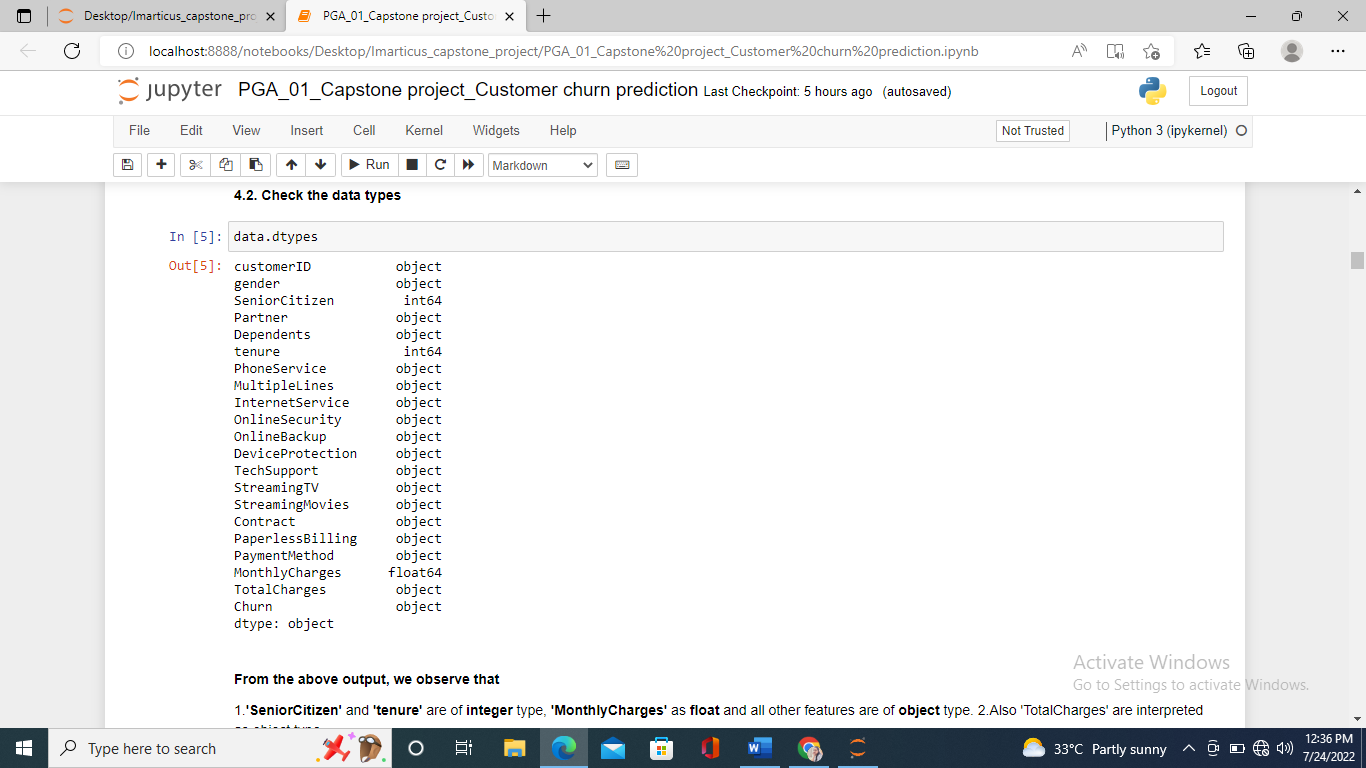
Graphical user interface, text, application

Description automatically generated

**READING DATA AND CHECKING INFO:**

**A screenshot of a computer

Description automatically generated**

****

Graphical user interface, text, application, email

Description automatically generated

**TREATING MISSING VALUES:**

Graphical user interface, text, application

Description automatically generated**Graphical user interface, text, application

Description automatically generated**

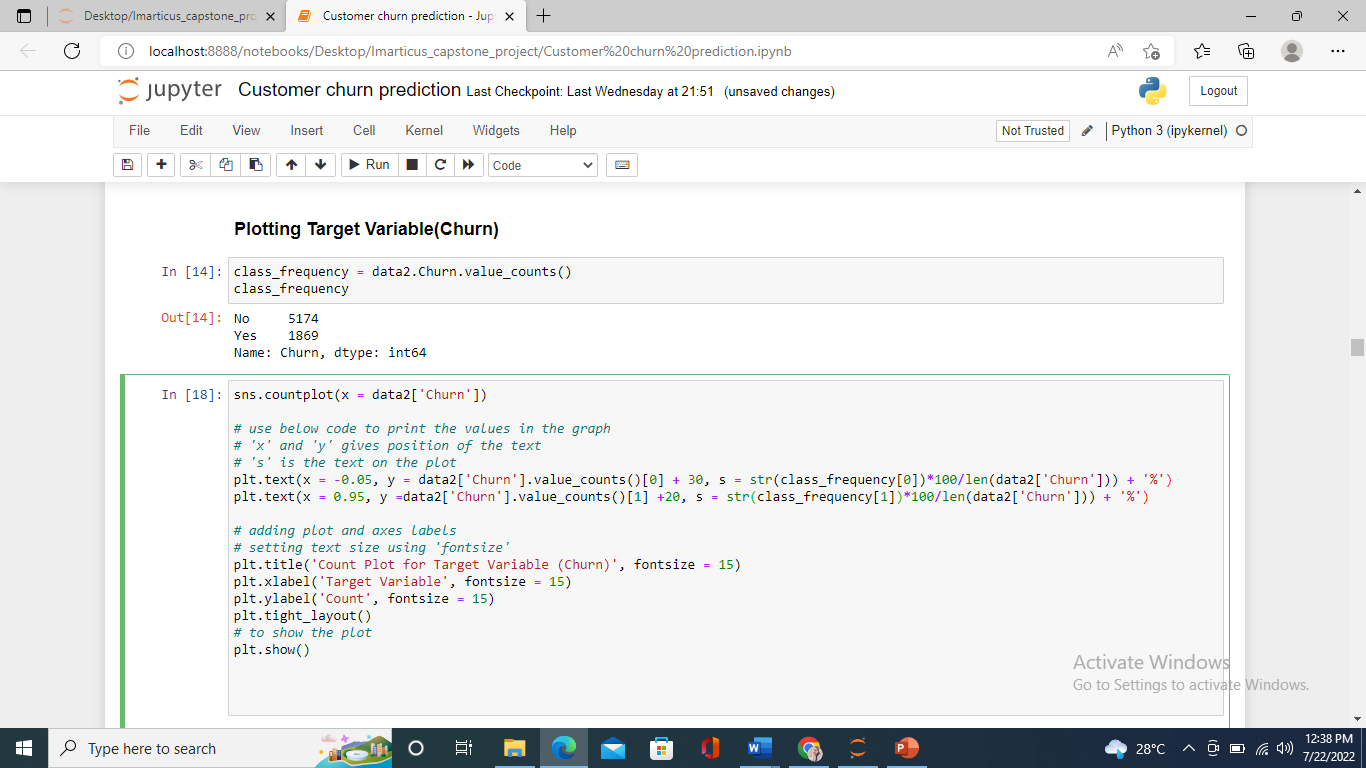
Now the data looks clean without any missing values. Further we had dropped a

‘customerID’ column which is not necessary for further analysis.

**EXPLORATORY DATA ANALYSIS:**

It is an approach to analyse the datasets to summarize the data, discover patterns, etc. often using data visualisation techniques.

PLOTTING TARGET VARIABLE:

Chart

Description automatically generated

From the above pic, we can observe that the data is imbalanced.

**DISTRIBUTION OF INDEPENDENT VARIABLES w.r.t CHURN:**

**Graphical user interface, application

Description automatically generated**

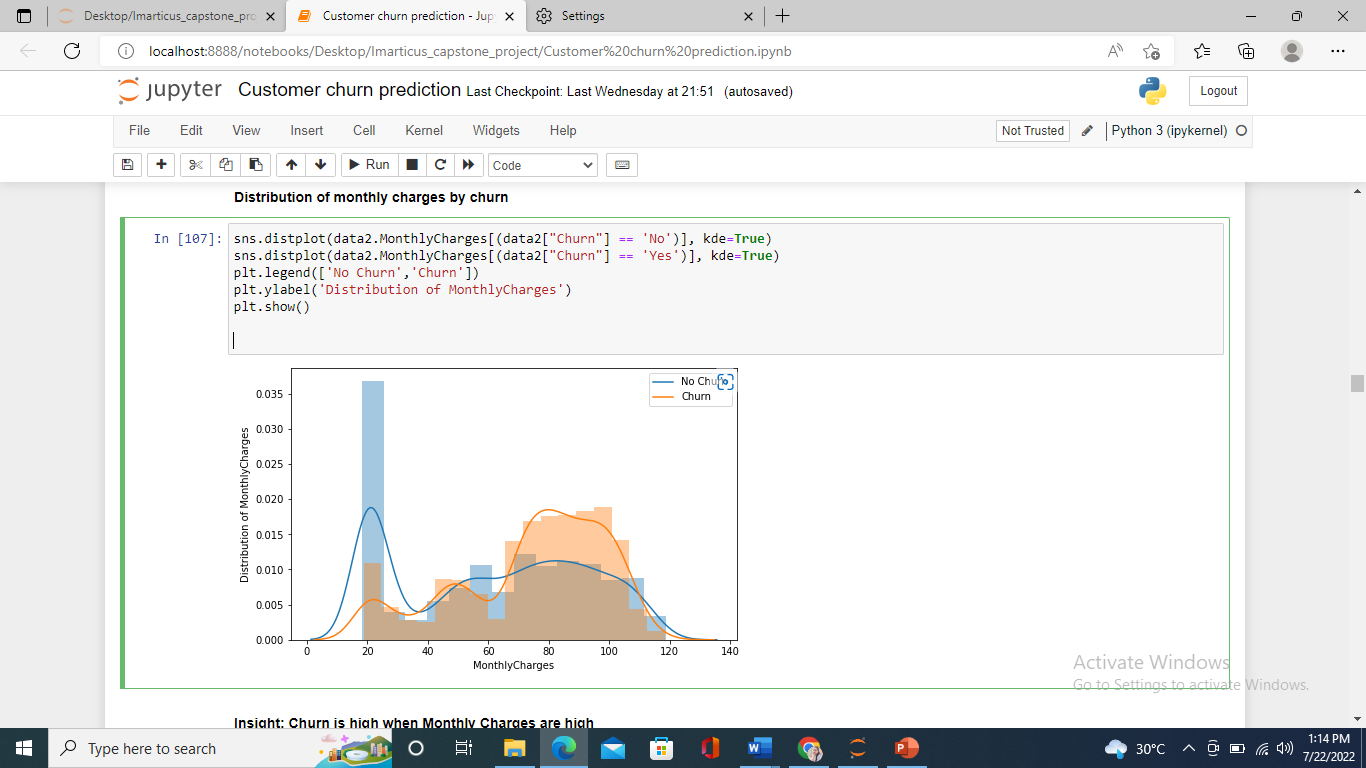
1.Electronic check medium are the highest churners.

2.Contract Type - Monthly customers are more likely to churn because of no contract terms, as they are free to go customers.

3.No Online security, No Tech Support category are high churners.

4.Non senior Citizens are high churners.

**DISTRIBUTION OF INDEPENDENT NUMERICAL VARIABLES**:

 Graphical user interface, application

Description automatically generated

Graphical user interface, application

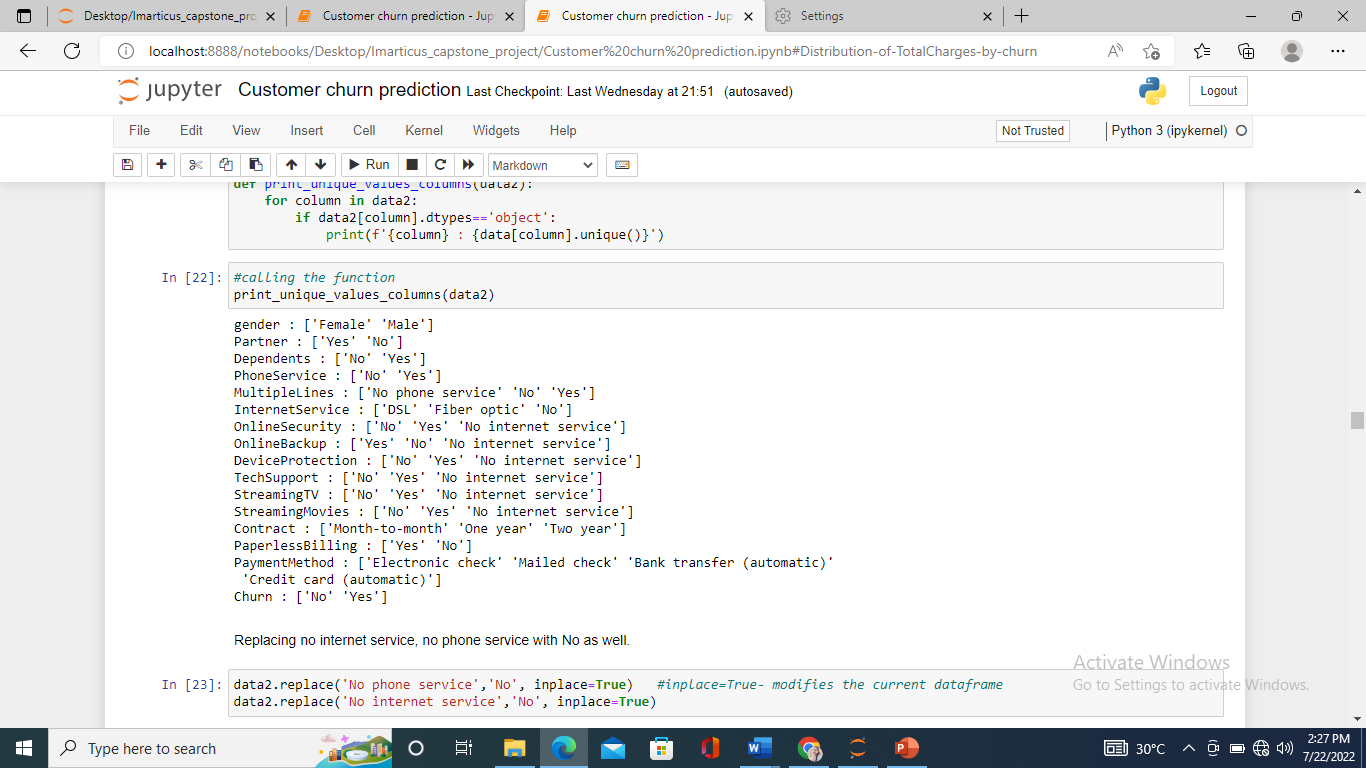
Description automatically generated 

The tenure histogram is rightly skewed and shows that most customers have been with the telecom company for just the first few months (0-9 months). The highest rate of churn is also in the first few months (0-9months). 75% of customers who end up leaving the Telco company do so within their first 30 months. The monthly charge histogram shows that clients with higher monthly charges have a higher churn rate. This suggests that discounts and promotions can be an enticing reason for customers to stay. Surprising insight is that higher Churn at lower Total Charges.

Hence, all these 3 factors viz Higher Monthly Charge, Lower tenure and Lower Total Charge are linked to High Churn.

As expected, MonthlyCharges and TotalCharges are highly correlated.

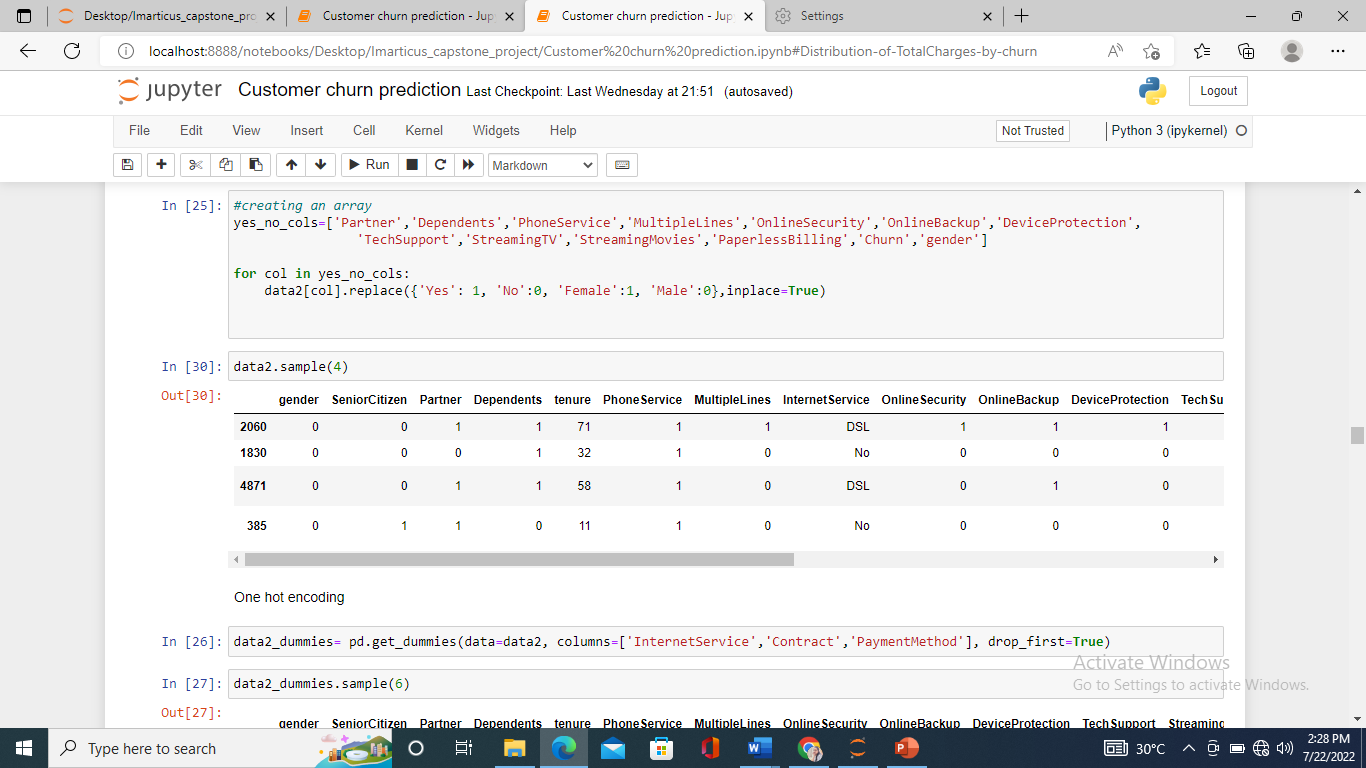
**DATA CLEANING:**

****

Created a user-defined function to print all the columns with their unique values. In some of the features we observe there is a unique value called ‘No internet service’, ‘No phone service’ which means the service is not available. So, replacing it with ‘No’ as well.

Graphical user interface, text, application

Description automatically generated



Replacing ‘yes’ with ‘1’, ‘No’ with ‘0’, and ‘Female’ with ‘1’, ‘Male’ with ‘No’ respectively using for loop function. Iterates through all the columns in the yes\_no\_columns list amd then replaces the values as we provided with replace().

**ONE-HOT ENCODING:**

One-hot encoding is used to convert categorical variables into a format that can be readily used by machine learning algorithms. The basic idea of one-hot encoding is to create new variables that take on values 0 and 1 to represent the original categorical values.

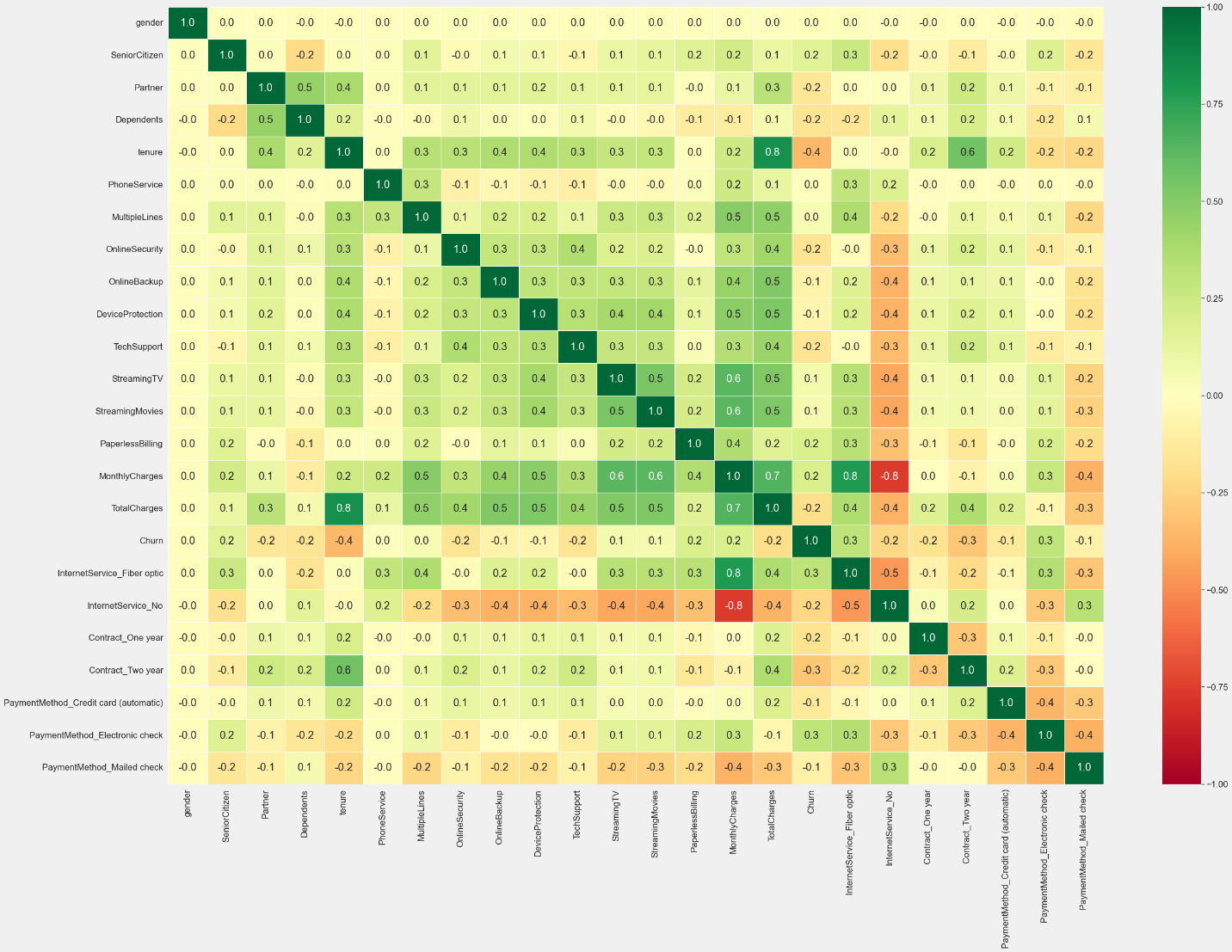
Graphical user interface, application

Description automatically generated

Now the data looks clean with all the observations in integer type. Let’s find correlation for all the features.

**CORRELATION MATRIX:**

Correlation is the measure of how the changes in one variable affects another. A correlation matrix is simply a table which displays the correlation coefficients for different variables.



The plot shows high correlations between Total charges & tenure and between Total Charges & Monthly Charges.

**DATA PREPARATION:**

Prepare the data by separating dependent and independent features by storing it in X and y variables. Further, we can split the data into training and testing sets to build a model using various algorithms. The training and testing data in the ratio of 80:20 or 70:30 0r 75:25 according to the knowledge of the domain.

We have used 30% for testing and 70% for training. We can build a model using training data and validate it in testing set.

Graphical user interface, text, application, email

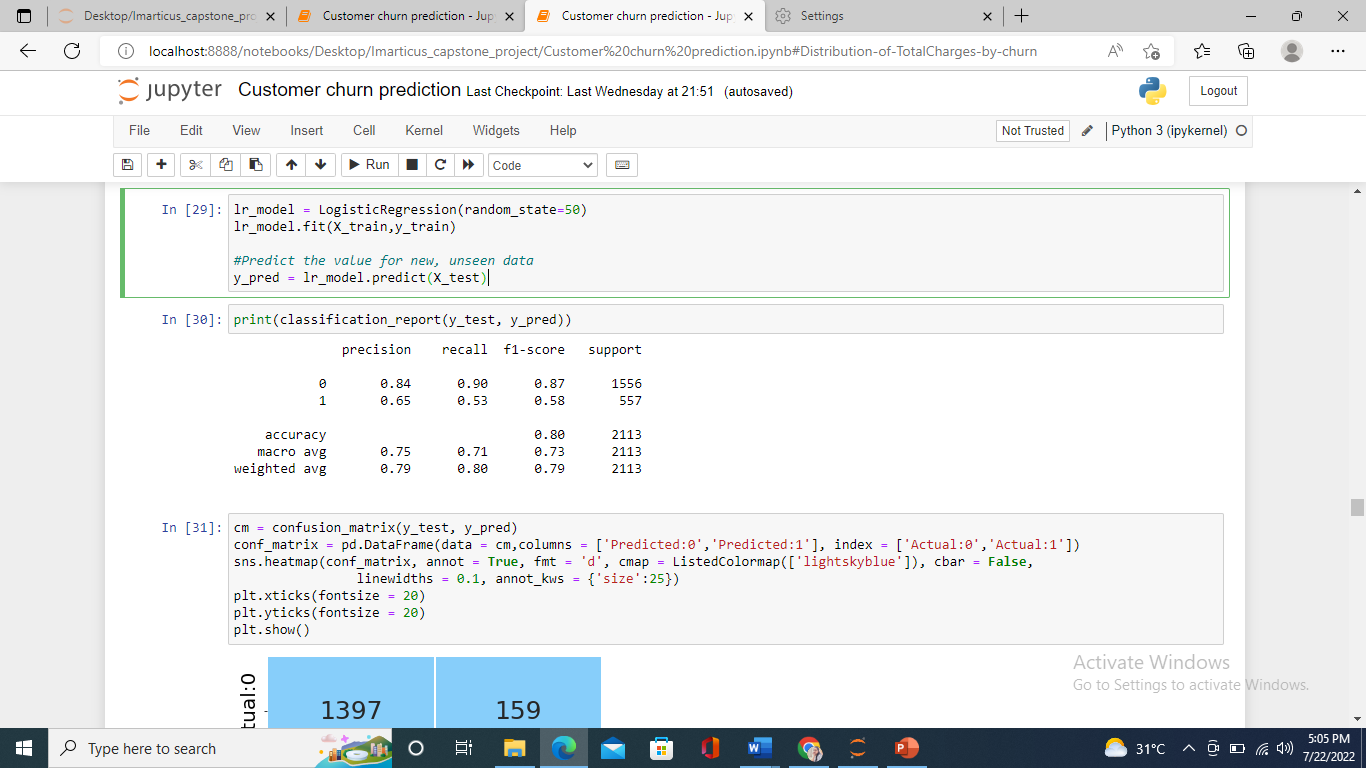
Description automatically generated

**MODEL BUILDING:**

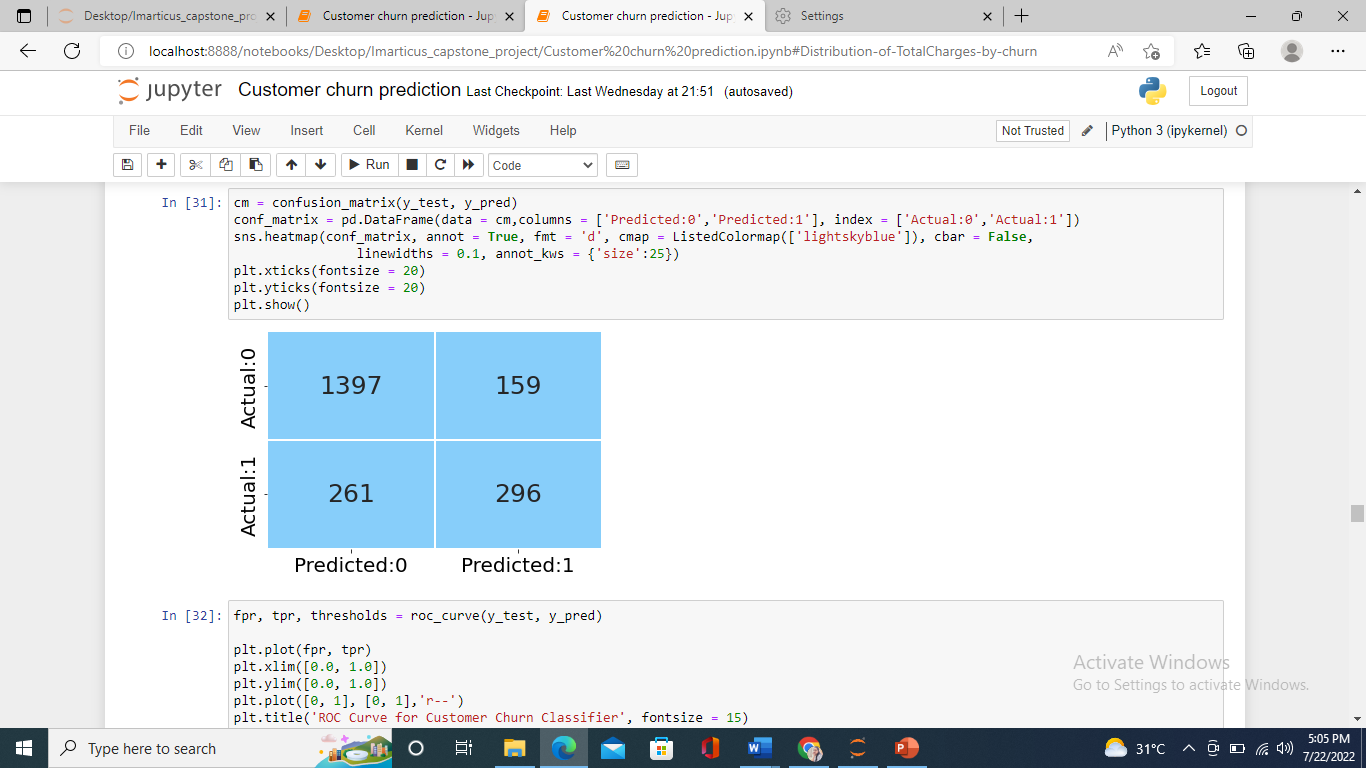
1. **Logistic Regression:**

It’s a classification algorithm, that is used where the response variable is categorical. The idea of Logistic Regression is to find a **relationship between features and probability of particular outcome.** According to our data, we need to predict whether the customer will churn or not when the new dataset is provided to the machine. Here, the target variable has 2 values. This kind of problem is Binomial Logistic Regression.

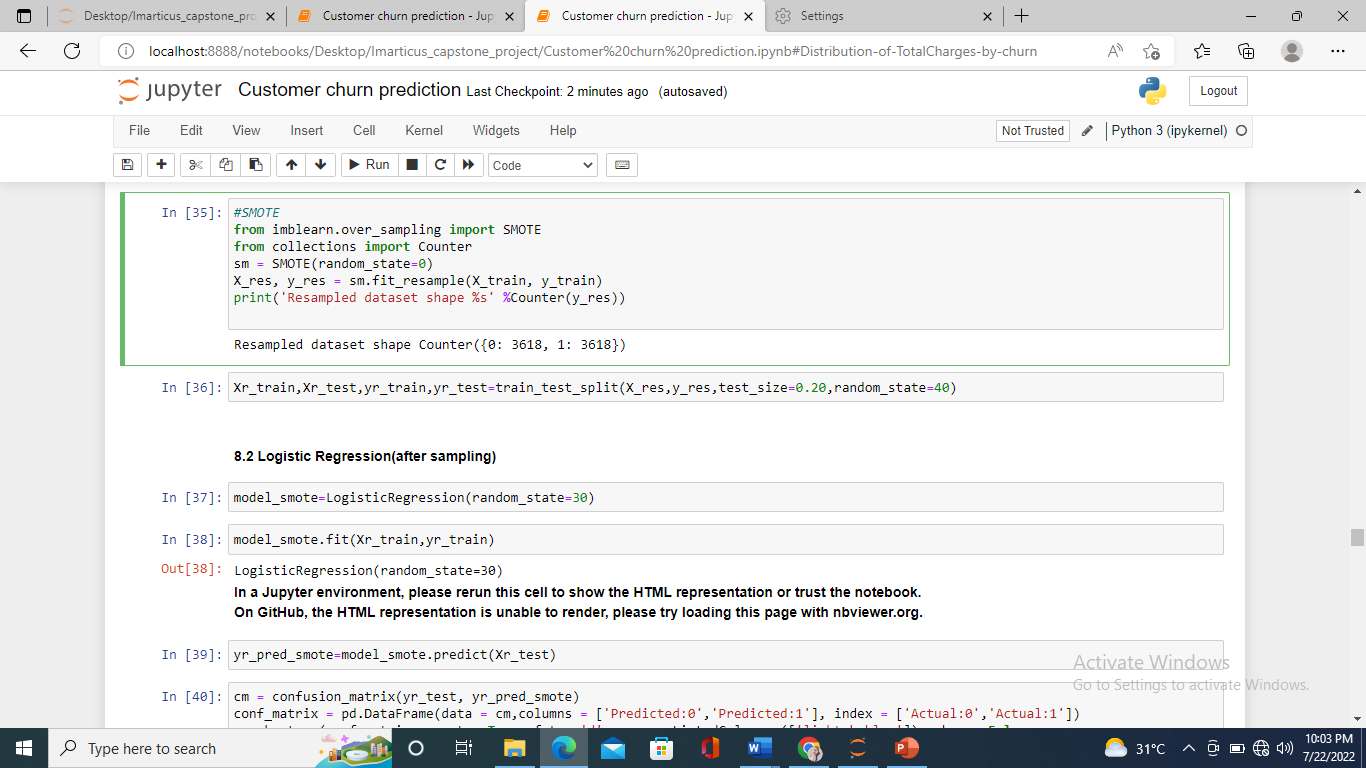
Performance of Logistic Regression:

 A picture containing chart

Description automatically generated

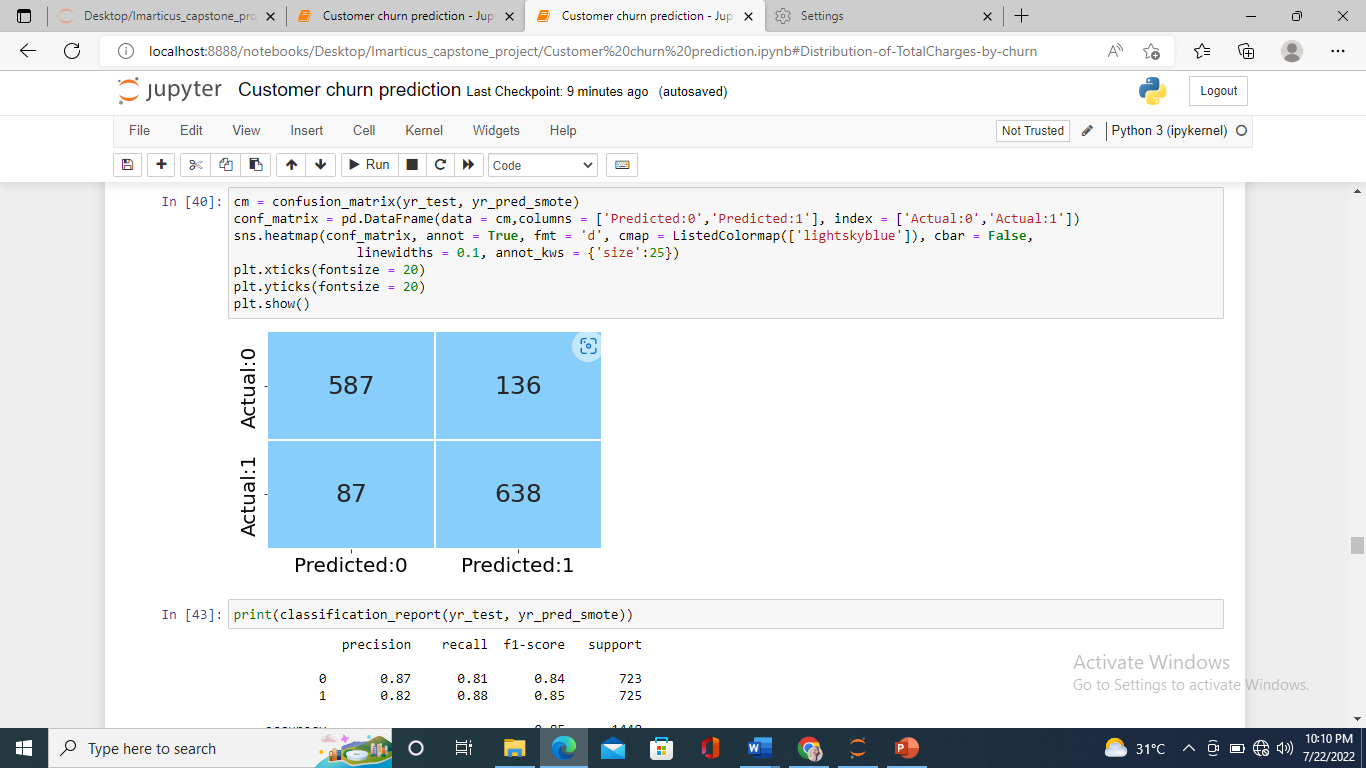


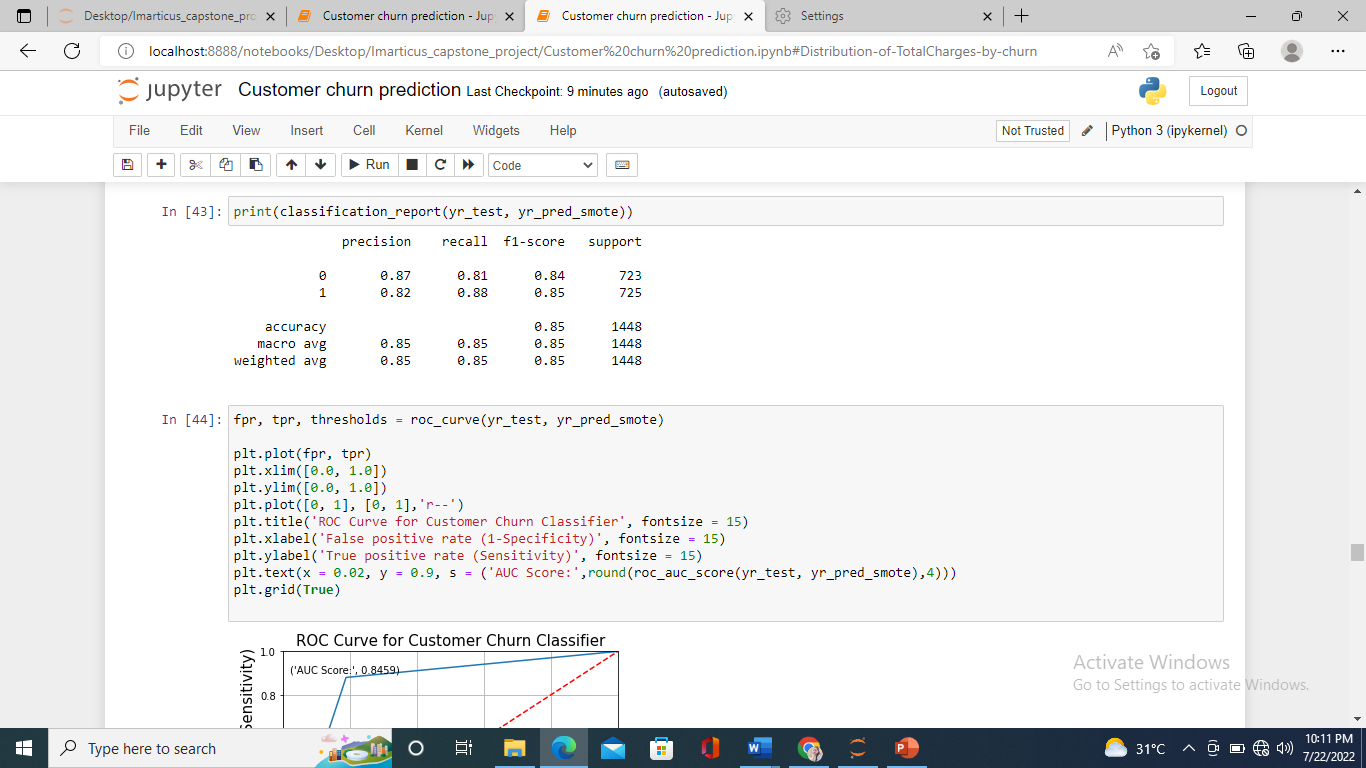
Even though the accuracy we achieved by logistic regression is 80% , the precision,f1,recall is extremely low in minority class. This is because our data is imbalanced. I had used SMOTE technique to balance it.

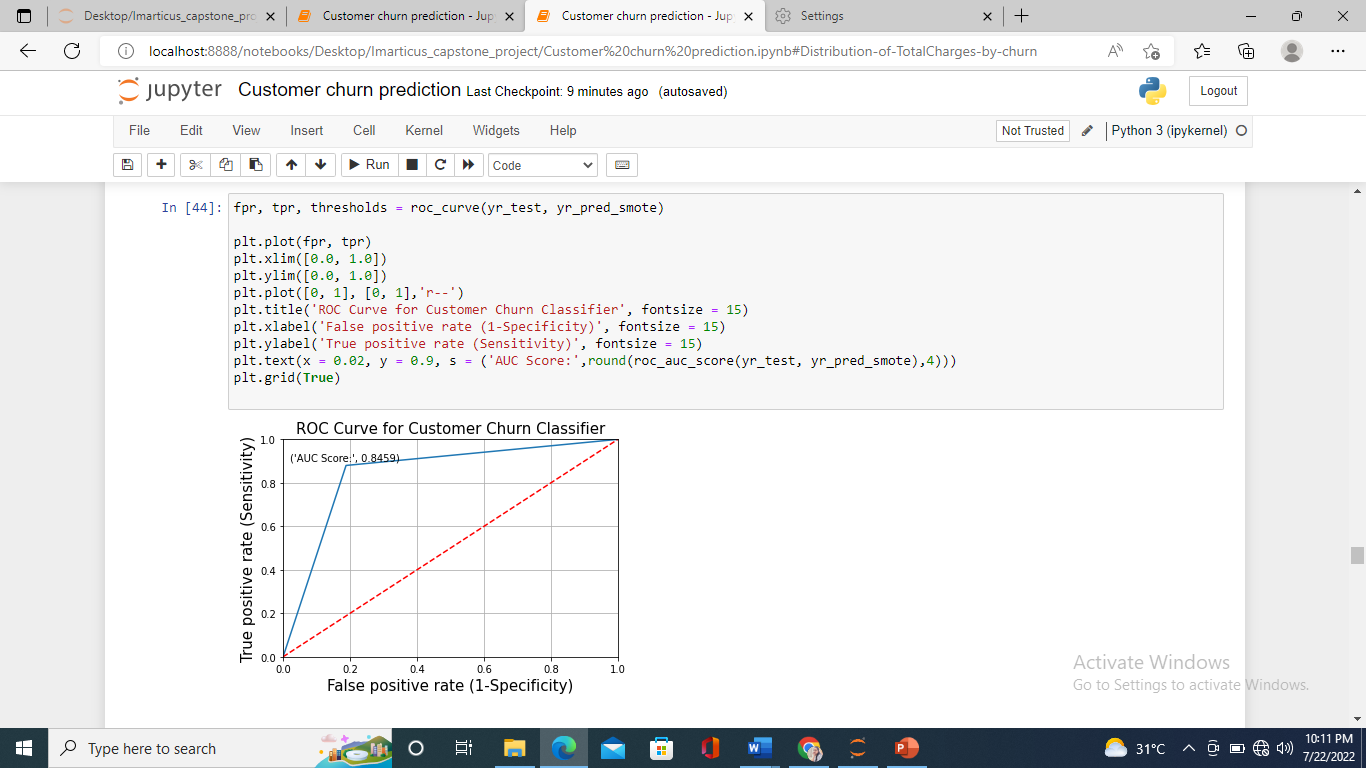


**LOGISTIC REGRESSION (AFTER SAMPLING):**

Fitting model with the balanced data.







From the above fig, we can clearly see the AUC score getting higher with 0.8459 which is a good value.

1. Random Forest Classifier:

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

1. Grid Search CV:

Grid-search is used to find the optimal hyperparameters of a model which results in the most ‘accurate’ predictions.

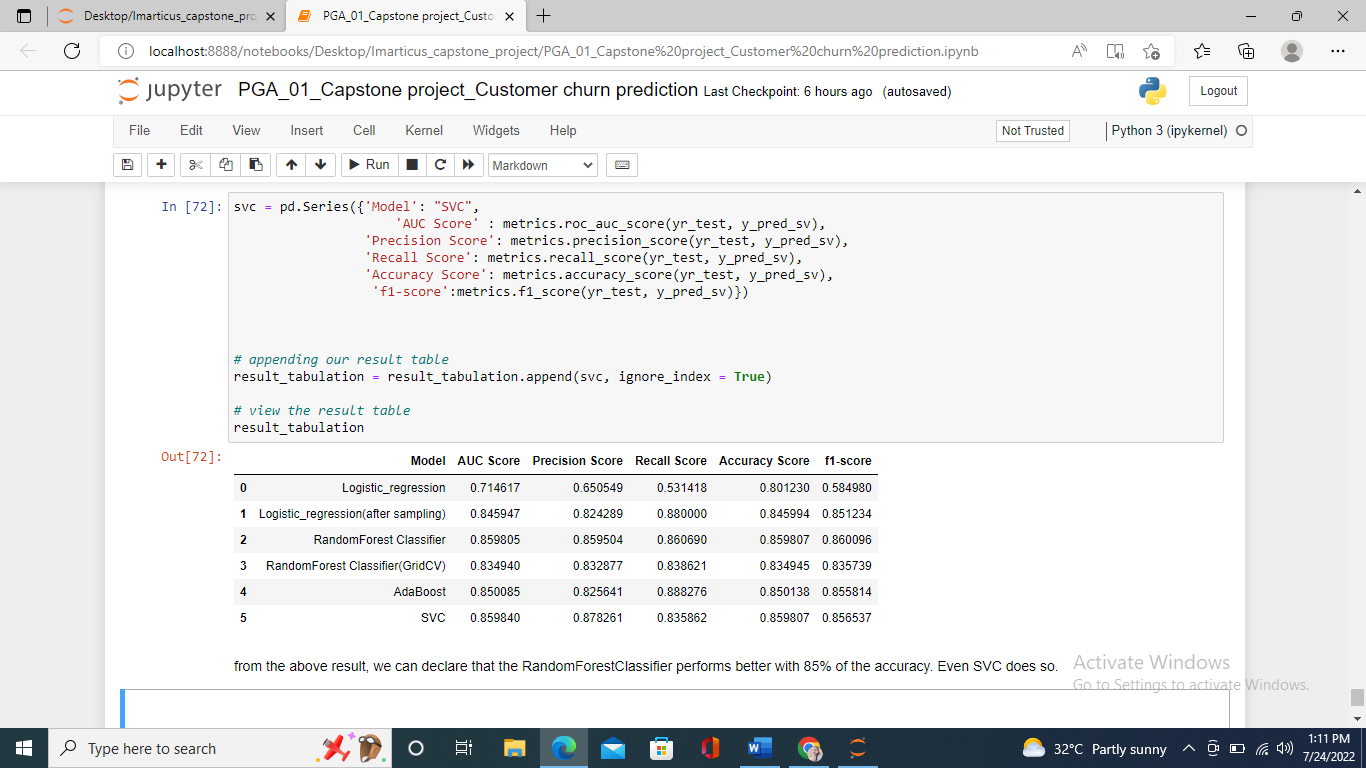
1. Ada-Boost:

AdaBoost algorithm, short for Adaptive Boosting, is a boosting technique used as an Ensemble Method in machine learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances. Boosting is used to reduce bias as well as variance for supervised learning.

1. SVC:

The objective of a Linear SVC (Support Vector Classifier) is to fit to the data, returning a ‘best fit’ hyperplane that divides, or categorizes the data. After getting the hyperplane, one can feed features to the classifier to what the ‘predicted’ class is.

**RESULT TABULATION:**

****

Looking at the table of metrics from all the five models, we can figure out that Random Forest Classifier has higher accuracy score with 85% and highest Area under ROC (Receiver optimum characteristics curve). Even SVC performed better.

This would be the best fit model with better performances and gives higher accuracy when we give new data to the model for predictions.

**CONCLUSION:**

Churn rate is an important indicator for subscription-based companies. Identifying customers who aren’t happy can help managers identify product or pricing plan weak points, operation issues, as well as customer preferences and expectations. As a result, by addressing churn, these businesses may not only preserve their market position, but also grow and thrive. More customers they have in their network, the lower the cost of initiation and the larger the profit. As a result, the company's key focus for success is reducing client attrition and implementing effective retention strategy.