**Trainer notebook:**

The trainer note book is the wrangling of a data taken from botometer site to train our Machine learning model so

What is wrangling?

First why to wrangle the data ?

We wrangled it so it could fit the feature of ML model

then

Wrangling is data processing and consists of 3 parts

1. Data gathering (the process of gathering data either provided by the client or imported from the internet) the botometer dataset
2. 2-Data assessing (the process of assessing quality and tidiness of data )
3. 3-Data cleansing (the process of cleaning data either due to quality or tidiness problems to make the data more reliable)

So what we did is that we gathered the data which was in JSON file and in text file

The JSON file was corrupted in spacing so pandas liberary couldn’t make a good data frame what happened actually is that pandas read file as a 2 dictionaries (2 columns user and created\_at)

I saved the file in df\_corrupted dataframe then divided the data frame into 2 dataframe each consists of one column df\_1\_s = created\_at and df\_2\_s = user

Then df\_1\_s which was a series was changed into dataframe and saved it to df\_1 then changed the name of columns from created\_at to created\_at\_1

Df\_s\_1 was a dataframe so in order to makes every feature in a column first I converted it into list df\_2\_l then converted list again into dataframe df\_2 finally I add the 2 dataframe (with concat) df\_1 and df\_2 to df\_full

Now we have a full data frame with each feature in a column we proceed to the txt file   
the txt file just contain IDs and if it’s a bot or not so I read it into dataframe df\_bot with columns’ names ID and bot then I merged (not concatenate) with df\_full (how = left “df\_full” as id is common in both so we got dataframe according to ids )

Now we have a full data and with also bot column   
we proceed to the final modification

First we change type of dates (created at and created at 1 )to datetime

From that we took the difference from created\_at “account creation date” and created\_at\_1 “tweet creation date” to get the age of account and to get it into days (for example if 1 year =365 days) we add .dt.days

Why in days so we can get average tweets per day “a feature in ML model” by dividing statuses/(age in days)

Lastly we saved the dataframe into new CSV file “datafull.csv” to run it in the model

**The original data (tweets\_2020.csv)**

The dataset tweets\_2020 was loaded in tweets\_2020\_wrangling notebook

What I’ve done in data was similar to what I have done in the training data and well be marked down in note book but as a summary:

1. I have removed duplicates according to creation date (i.e. from most recent) so I can got the age of each users
2. I have created also hour based on the recent tweet
3. I have created the other feature like tweets per day and so on
4. The data then saved to csv file called ‘tweets2020\_ML.csv’

*Please note that the data lacked the favourites count geo enabled default profile and default profile image I gathered these data using tweepy but tweepy got most recent data (of these feature) i.e. the data of June 2021*

**Running the model**

The model\_loaded notebook for is saved in folder ML on users and that folder contains the dataset of tweepy ”users\_2020.csv” (the 4 missed features the have been gathered using tweepy) and the wrangled data of tweets\_2020.csv file (tweets2020\_ML.csv)

How models work?  
after training model it’s saved as \_\_\_.pickle file

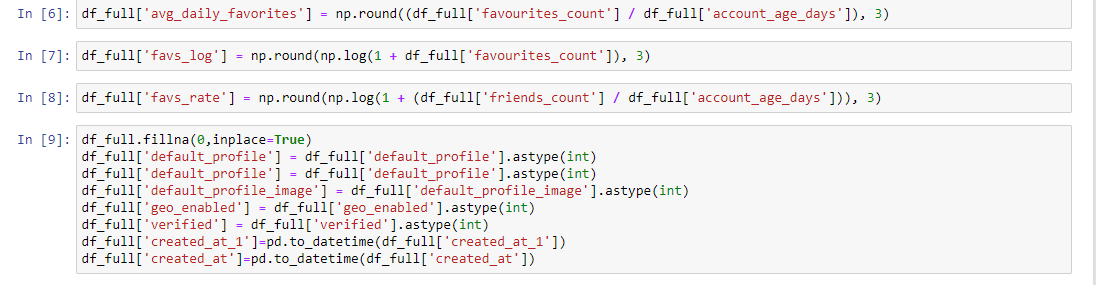
Then you recall that .pickle file to run it

Before running the .picke file we will first call the wrangled dataset as df\_users\_1 and tweepy dataset as df\_users\_2

After that we will merge them together how=’left’ and on=’screen\_name’

*Note the left dataFrame must be the tweepy dataset (df\_users\_2) and save the to df\_full*

After that I copied the calculation that’s present in Machine learning model itself

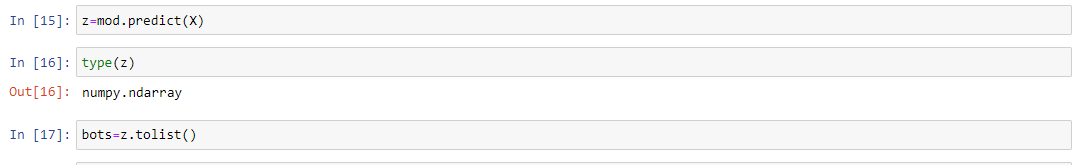


I filled null values to 0 then continued

Lastly I’ve selected the features then imported pickle library to call the model.pickle file with open functionI



I saved model to mod then run it with mod.predict(X) and saved the result into z variable

The results were in np.array so I’ve got the list of results by using.to\_list() method and saved it to bots variable

Lastly I have made new DataFrame called bots\_2020 with just to columns one is the screen name of df\_full and the other the 0/1 number of bot prediction (bots list)



*Please don’t change or set the list bots because the value of ML prediction are saved in the same order of screen\_name otherwise you’ll put 1 (bot) for non bot users and 0 (human) for bot users*

*The rest were just visualization and lastly saving result in csv file*