Hi there ,

My name is marwan takrouri , I am a civil engineering degree holder and Data Analyst , graduated recently from university of Toronto , versatile data analytics and valuations boot camp ,

I am writing a small script to introduce you to our final project , Group I ,

The project was : Human or Bot .

My Role in the Project was :

* Initiate, control , monitor and master the generation of our GITHUB repo , and helping my peers in the team with generation of the branches .
* Selection of the project title itself from three options were short listed by the team members .
* Combined the data in a google slide for a rough sketch of our data in the beginning
* Working out our project tableau inter active dashboard.
* The Main presenter for the project in the final presentation day.

Project link :

<https://github.com/Marwan-Takrouri/Human-or-Robot-Facebook-Kaggle>

Purpose of the Project :

A in depth data analysis using machine learning, and all the take outs of our data analytics boot camp practices from SQL to python to Train, Test and Splitting data combined to make out a brief state of the Art , Zero Errors analysis , with a collaboration of a very talented and professional team members of five including myself .

We wanted analysis and find how Many online bids are placed by robotic/automatic bidders

Using ML to flag these robotic users and also :

* Can we identify bot behavior using ML?
* In order to improve fairness of bidding competition, can we eliminate (if not greatly reduce) bot generated bids in auctions?

Data collection:

* Our data is from Kaggle Competition posted from 7 years ago.
* Bidder datasets
  + Training dataset
  + Test dataset
* Bid dataset
  + Bid dataset
* Data given: bidder id, payment account, address, auction, merchandise, device, time, country, IP, URL and outcome.

Out come and Results :

* Mean of time difference between bids by each bidder (robots have a lower mean time)
* Dataset consisted mostly of categorical variables, making the data cleaning harder than initially expected
* We used the model for imbalanced data, but we can also try undersampling or oversampling before using the model itself
* The model can be further improved, should be able to get more than 90% balanced accuracy score by making some changes with preprocessing
* Using MaxMind GeoIP2 Python API to fill in empty countries