Iris Borkovsky

Project 3 - McNulty

Robots vs. Humans

Competition Data Set

Objective:

I would like to use project McNulty as an opportunity not only to familiarize myself with classification algorithms, but to also make use of SQL and AWS. In order to ensure my use of the latter two, I did not shy away from larger datasets. I will be using a dataset from a past Kaggle competitions. The goal is to classify the users of a bidding website into real human customers and scripts ("bots") that are created to outbid people on items and therefore detract from the user experience.

Domain:

I am not familiar working with user data in a machine learning setting, so this will be an opportunity to grow my skill set.

Data:

The data consists of two large datasets. One contains information about all the users of the website while the other has information about separate bids that occured on the site in a given time period. The data, while vast, is cleaned and ready to be used.

What I like about this data set in particular is that I can think of a few features I'd like to create (for instance, the meantime between generated bids for a particular account), and the data will allow me to make them without encountering intermediate steps.

Known unknowns:

I do not currently know what percentage of site's users are real people vs. robots. This is something I would like to find out before proceeding with my investigation as it would aid me in choosing better suited classification methods and would allow for subsetting the data.

I am also not experienced with either SQL or EC2, so those will certainly require more attention from me in order for the project to move forward.

Shooting for the stars:

If everything works out smoothly, my "dream" achievement would be classifying the data into two distinct categories with a high degree of certainty. A more feasible outcome which I would be satisfied with is being able to detects a distinction in the data, indicative of the existence of two classes.