Pip install pandas

Pip install sklearn

Choice

A\_follower\_count,

A\_following\_count,

A\_listed\_count,

A\_mentions\_received,

A\_retweets\_received,

A\_mentions\_sent,

A\_retweets\_sent,

A\_posts,

A\_network\_feature\_1,

A\_network\_feature\_2,

A\_network\_feature\_3,

B\_follower\_count,

B\_following\_count,

B\_listed\_count,

B\_mentions\_received,

B\_retweets\_received,

B\_mentions\_sent,

B\_retweets\_sent,

B\_posts,

B\_network\_feature\_1,

B\_network\_feature\_2,

B\_network\_feature\_3

1 means select A

0 means select B

A solution used average of RF and GBM using original features.

Also created a 1AB for each 0BA. Plus, I assumed that if A>B>C than A>C

Does doubling the training data help? And then the testing data?

Someone normalized all features by median and std then scaled to 0-1

Someone derived a feature that was the subtraction of following\_count from follower\_count because they saw a few where both follower & following count were high and thought it would be more useful to do this??? Also they left out retweets sent as not useful.

From Kaggle on Test data using their submission thing. These are not percentages but areas under the curve or something. IDK check out the evaluation part

Baseline Solution using SciKit = 0.85586

Follower Count Benchmark = 0.75175

Random Benchmark = 0.50438