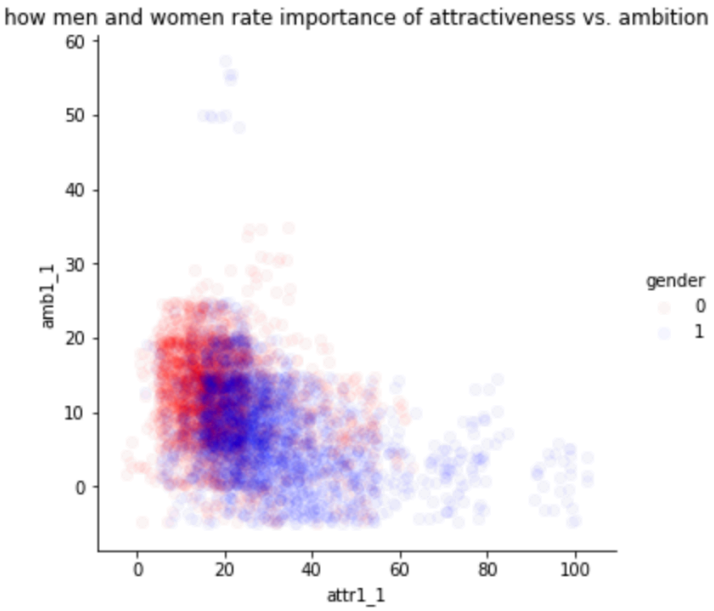
I'm using a Kaggle speed dating set to predict if a man and woman match or not. Another way to put it would be, do they decide to go out again or not? I've spent most of my time feature engineering so far. There was a lot to clean and new features to create in order to give some of the fields more salience. An initial pass at the fitting gave me the highest accuracy on random forest (accuracy = 0.827) and xgboost (accuracy = 0.847)
Some things I rated highest were if participants were from same zip code or state, and had similar career ambitions. I thought income and age would be also be important. I created some variables for how desirable someone is vs. how choosy they are. From the bar chart below, what order you were in for the speed dating event seems to reign supreme.
By the end of my cleaning, I still have ~500 features, many of them dummies.
My initial f1 score isn't too good but I'm going to try to use an fbeta score instead in case this might help me favor recall over precision? I'm thinking about false positives vs. false negatives in this scenario. The cost of a false positive isn't that bad, just a bad date, but a false negative may be a missed opportunity for a life partner.

Some initial graphs show a few interesting things, like women tend to rate ambition higher, while men rate attractiveness higher.



An initial bar chart shows relative feature importance.