OUTLINE

* Introduction
  1. Experience with machine learning before starting/Expectations
  2. Game plan
     1. Data Processing
     2. Model Selection
* Data Processing
  1. Attempted to fit model on raw data
  2. Apparent from results/forums data processing was required
  3. Imputation played a big role, perhaps more research needs to be done on an optimal way to deal with NA’s
  4. Our (and every other groups) CV errors seem to underestimate what we get as a test error – due to over processing/using test data? Just overfitting models? Important to determine the cause so we can trust that optimizing cv error optimizes test error.
* Model selection
  1. Thanh knows better which models should fit well
  2. Tanner Carbonati’s kernel ensembled xgboost, ridge, lasso, and elastic net which seemed redundant. Ensemble of xgboost and PLS worked better (why?), using pls instead moved us up about 5% in leaderboard.
* Future
  1. Optimal imputation techniques, fixing cv error underestimation
  2. Finding better features
  3. Better use of ensembling (I just averaged two predicted y’s)
  4. …