I used test.csv only to compute a mean opponent rating for each player for the current month. I realized this was very naive only when I saw Tim sending 0.246 score (which I believe is unlikely using only the mean opponent rating) I realized I should have tried reconstructing more information out of the test.csv, there was only couple of days left, not enough to do this. Without further use of test.csv I believe I was reasonably close to the predictive power what could be expected out of an $10k competition. I think improving further would have required larger prize for motivation.

My model is an ensemble method that includes many known rating systems, elo, glicko, chessmetrics, etc. (I ended up modifying them significantly to optimize for the competition data) and predictors such as different "ratings" for each player pair and variants of these predictors with limited window (i.e. they reset or remember only few months back) also some experimental predictors that try to compute prediction from: recent opponents, mean opponent rating for the month (test.csv), the absolute history win/draw/loss, time since last game, delta from min/max historical elo, etc. various combination of these features. Because of the competition design (noise and shuffling of games) it was very difficult to make more complex predictors (than known rating systems) work. My model was very much worsened by the noise as I require precise order of games that actually happened to compute a new prediction. I was expecting a leaderboard score of 0.243 but got 0.247 instead.

My model wasn't a rating system because a rating is a weak predictor. To my knowledge Shirov never beat Kasparov, even though their ratings were close enough, hence it will not work to try predicting a score for a Shirov vs Kasparov and Shirov vs Kramnik games from the same rating.