Around 50 thousand games are included in the training set for this competition. These games are not uniform in their rules, with different time settings, opponents, and even allowable words creating subtle differences in scoring capabilities; as well as differences in what a good player may be able to accomplish in these different game types.

With all these differences, it is important to provide a training set for the model that will not fixate and overfit on sparse combinations. Grouping several of these more sparse options into one will help reduce the overfitting issue there. However, one of the biggest issues with the dataset revolves around a handful of players and their gameplay frequency.

Games are either ranked or casual. Ranked games result in a player gaining or losing rank depending on the outcome of the game. Casual games do not result in a player changing rank due to the game’s outcome. When players register and sign up for the site, they are given a uniform 1500 rating. Should a player exclusively play casual games, their rating will stay at 1500, despite their ability perhaps being significantly different from that rating.

Much of the training data and test is tilted to several frequent players. The top 10 frequent rated players account for 20% of the training data. It is 36% in the test. One player, BB-8, had over 6000 games registered–all in casual (almost half of the training data’s casual games). Their rating was a flat 1500, despite most of our indications pointing to them being a stronger player. While more data is typically good for training, we fear the model would overfit on the attributes of this player, and would assign a rating near 1500 for anyone who plays similarly. That is to say, the grand truth isn’t very reflective of the truth!

Heavy players like this also exist on the test set. However, we believe that those in the test set have played some rated games in the past, and their rating is somewhat near their true strength.