Valley Boyz Hackers

Karthik Rajkumar, Shardul Kothapalli, Sachin Konan

August 4, 2019

Abstract

We propose completely eliminating the lottery system in favor of a machine learning-based ranking system which is able to account for deliberate losing (tanking) throughout the season.

Introduction: The NBA Draft serves as a lifeline for bottoming teams, replete with fresh young talent that they hope can bring them back to relevance. Unfortunately, while attempting to ameliorate the prevalence of tanking, the lottery system as well as its recent revision has somewhat contributed to perpetual tanking: the 2019 lottery is a prime example. Developing, young, but poorly constructed teams such as the Atlanta Hawks, Chicago Bulls, and the Phoenix Suns, have looked to the draft to build their teams, but were only able to land the 8th, 7th, and 6th picks, respectively. Meanwhile, teams like the Memphis Grizzlies and the New Orleans Pelicans landed the 2nd and 1st overall picks in the draft, which is a testament to the infallibility of the system for the simple fact that these two teams were the ones tanking: the Pelicans rested their star player for 4th quarters for over half the season, and as soon as the Grizzlies entered a mid-season slump, sent away their franchise cornerstone, Marc Gasol, and were simultaneously actively shopping Mike Conley Jr.

We seek to determine an effective win-loss record that accounts for tanking throughout the season. This adjusted win-loss ratio will be utilized to determine drafting order. Hypothesis: We posit that by determining a quantitative measure of tanking and non-tanking games the NBA can better calculate the true win-loss ratios for teams near the end of a season. Furthermore, we predict that games where a given team has lower average lineup age, minutes played of the best lineup, offensive/defensive rating, and Pythagorean win shares, can generally be classified as a tanking game.

Potential applications: This project can provide key insights into how to reform the nba draft process. The NBAs current method to curb tanking with weighted odds is rather rudimentary considering we have modern analytics to accurately assess the quality of teams. However, even if the NBA is unwilling to let go of the lottery completely, our metrics can still help create better lottery odds that is more fair for teams. We envision a draft system where the distribution of high quality draft picks is determined by a win-loss record independent of any tanking a team might enact.

Pre-Processing Data/Types of Data Used: A crucial step to classifying whether a game can be classified as tanking or non-tanking is analyzing the games statistics in relation to cumulative/running performance. This is why several of the variables we have chosen are differential to a sliding window or cumulative total. The amount of variables explained below is open to an increase during actual development, depending on how they would bolster the model.

1. *Age Differential:* The difference between the average weighted age of a team across a season and for a particular game. The weightage originates from the minutes played by each player on the roster, so if a 40 year old player plays zero mins, then his age isnt accounted for.

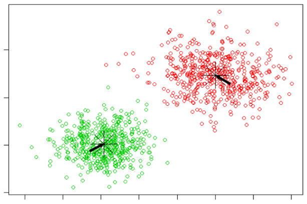
2. *Golden Lineup Play %:* a) For each game, create a list of active players. b) Filter NBA data for golden lineup that is a subset of active players in a game. c) Calculate total minutes of that lineup/ (48 minutes * 5 players). We can figure out when teams are resting their best possible lineup or conversely when they are playing them.

3/4. *Team offensive/defensive rating differential:* (difference between rolling average offensive rating/defensive rating and game). We can figure out when overall team OR/DR is down.

5. *Pythagorean win differential:* a) look at win loss within current sliding window. b) look at the Pythagorean win loss during current sliding window. c) apply that to every single game (stride length variable). We can figure out if a team is underperforming, performing well, or at average.

Statistics 3-5 are differentials that are compared to rolling average. This rolling average will be calculated every 10 games at a stride of 1, to include the inherent ups-and-downs of a team through a season

Proposed methodology: Our training set will consist of several years of game/team data. We choose to only analyze losses because even our model classifies a win as a tanking game, we do not want to double count it when we conduct the W-L record adjustment. We assume an input tensor size of (# of lost games * 30 teams) x 5 (# of statistics stated above). This prompt essentially boils down to an unsupervised learning problem, because we are trying to classify whether the result of each game can be attributed to tanking or true failure to perform, without any knowledge of the game being a tanked game.



To determine whether a game is tanked or not, we will investigate all the effectiveness of various clustering algorithms such as k-means clustering, self-organizing maps, neural autoencoders, etc on all games that resulted in a loss. Once two discrete clusters can be generated, we will remove the lost games attributed to tanking so theoretically we can calculate more accurate win-loss records.

The idea that teams will begin optimizing their losses in order to end up in the tanking cluster is a legitimate concern; however, since machine learning algorithms are black boxes, teams should not be able to optimize according to a formula, as these models consist of complex convolutions and optimizations that make it nearly impossible to track outputs back to inputs.