Winning Fantasy Sports with Apache Spark (Appendix)

By Jordan Volz

Statistical Overview

Determining the value of a player in fantasy sports is not a trivial task. Our ultimate goal should be to produce a ranked list of players in terms of their value, but being able to assign a numerical value in order to compare players directly would also be useful for many use cases. Such a list would enable managers to select players for their team with confidence. Most fantasy leagues will provide player rankings preseason and in-season, but it's not always transparent how those rankings are calculated and for which types of league they apply to. Even fewer actually provide the numerical value that is driving that ranking or explain the process involved in calculating it. In this space we'll attempt to make some sense of what goes on behind the scenes and formulate a reasonable method for determining player value.

There are many variants between draft styles (e.g. traditional or auction/salary), league types (e.g. rotisserie, head to head, points), and which statistics are scored (dependent on which sport is played). Additionally, leagues may implement keeper or dynasty rules that allow teams to carry over a portion of their team to the next year, which adds another layer of complexity. For our purposes, we'll assume a fairly standard league that employs traditional drafting and is a head to head league (H2H). The objective of the draft is to assemble the team with the largest expected value; auction drafts complicate matters in that decisions early in the draft (such as paying too much for good players) can negatively affect a manager later in the draft (such as when mid-tier players can't be afforded and the team has to settle for too many low end players) -- in this case being economical is much more important. Traditional drafting avoids this complexity by allowing you to select the best available player at every juncture of the draft.

Regardless of what happens during the draft, we'll focus primarily on assessing player value during the actual season. Each type of league will also influence how value is determined. In head to head leagues, the objective is to win more categories than an opponent each week. Those categories are then translated to "wins" and "losses" and accumulate over a season. Rotisserie leagues are similar, but there are no weekly matchups -- whoever is leading in the most categories at the end of the season is simply declared the victor. In both cases, there is an argument that having a well-rounded team that is competitive in all or most categories gives a manager the greatest chance of claiming victory. In points leagues, statistics are converted to points in some manner, and those points are either accumulated over the course of a season (like rotisserie) or weekly vs an opponent (like head to head). In this format, how points are assigned to statistics completely determines which players are valuable in the league.

For our analysis, we will concern ourselves with the following 9 basketball statistics: Field Goal Percentage (FG%), Free Throw Percentage (FT%), Three Pointers Made (3P), Total Rebounds (TRB), Assists (AST), Steals (STL), Blocks (BLK), Turnovers (TOV), and Points (PTS). These are often referred to as the "standard-nine" in fantasy basketball circuits; another popular set of stats consists of the same set minus turnovers. Other collections of statistics can be analyzed in a similar manner as we show below, simply by substituting for the desired statistics.

Z-scores (aka standard score) are a popular statistic in sports that gauge how close a value is to the mean, in terms of standard deviations. One of the main reason they are used in fantasy sports is that they're able to take all statistics and convert them to numbers which you can them begin to compare across statistics. For example, knowing that player A averages 10 rebounds a game and player B averages 10 assists a game is not very useful in and of itself, but if you know that player's A z-score for rebounds is 1.5 and player B's z-score for assists is 3, then we know that player B is better at assisting than player A is at rebounding, compared to the rest of the league. Be careful, however; raw scores are not necessarily distributed normally nor similarly, and z-scores are a linear transformation; hence, the relative positions of the values are not altered under z-scores nor is the shape of the distribution changed.

For each player in every year, we will calculate a z-score for each statistic.
Z-scores is calculated as follows:

For every element i in the set {3P, TRB, AST, STL, BLK, TOV, PTS} and each element j in the set of all players for a given year, we calculate

$$statZ_{(i,j)} = \frac{(stat_{(i,j)} - \mu_i)}{\sigma_i}$$

where μ_i is the mean and σ_i is the standard deviation of $stat_i$ in a given year. For example, if $\mu = 4.5$ and $\sigma = 1.3$ for TRB in a given year where John Doe averaged 7.1 TRB a game, we would calculate:

$$statZ_{(TRB,John\ Doe)} = \frac{(stat_{(TRB,\ John\ Doe)} - \mu)}{\sigma} = \frac{(7.1 - 4.5)}{1.3} = \frac{2.6}{1.3} = 2$$

Here we find the z-score for John Doe is 2 for total rebounds, meaning his average is 2 standard deviations away from the league average.

The one caveat we have in this calculation is that z-scores for FG% and FT% don't take into account the volume of shots that are taken. For example, a player who makes 60% of his field goals but only shoots 1 a game is likely less useful than a player who makes 55% of his field goals but shoots 20 times a game (taking as a given that the league average for FG% is much less than 55%). Different methods have been proposed in order to take shot volume into consideration for this calculation (see: field goal impact and fgscore, for example). Historically, I calculated a weighted FG% (FGW) by taking the z-score for FG%, multiplying it by FGA divided by the league average for FGA, and then calculating the z-score of *that* result. In other words, I tend to do the following:

$$FGW_j = \left(\frac{(stat_{(FG\%,j)} - \mu_{FG\%})}{\sigma_{FG\%}} * \frac{(stat_{(FGA,j)})}{\mu_{FGA}}\right)$$
$$FTW_j = \left(\frac{(stat_{(FT\%,j)} - \mu_{FT\%})}{\sigma_{FT\%}} * \frac{(stat_{(FTA,j)})}{\mu_{FTA}}\right)$$

$$statZ_{(FG\%,j)} = \frac{(FGW_j - \mu_{FGW_j})}{\sigma_{FGW_j}}$$
$$statZ_{(FG\%,j)} = \frac{(FTW_j - \mu_{FTW_j})}{\sigma_{FTW_j}}$$

where FGA and FTA are field goal attempts and free throw attempts, respectively. You may notice that my formulation, field goal impact (FGI), and fgscore (FGS) are all very similar -- in fact, the following is true:

$$FGW_j = \text{FGS}/\sigma_{FT\%} = \text{FGI}/(\mu_{FGA} * \sigma_{FT\%})$$

What is nice here is that means and standard deviations are invariant under scalars, so all three of these metrics end up giving you the same z-score. For computational simplicity, we'll adopt field goal impact (and free throw impact).

$$statZ_{(FG\%,j)} = (stat_{(FG\%,j)} - \mu_{FG\%}) * stat_{(FGA,j)}$$
$$statZ_{(FT\%,j)} = (stat_{(FT\%,j)} - \mu_{FT\%}) * stat_{(FTA,j)}$$

Once we have all the z-scores in a given year for each statistical category, we can sum them together to get an aggregate score for each player j:

$$totZ_j = \Sigma_i \, statZ_{(i,j)}$$

This gives us a single number which we can use to compare players across years. Note that since the z-scores are calculated within a given year, what we are measuring is a player's performance relative to other players playing at the same time. I.e. we don't need to account for changes in the style or tempo of a game to account for why the raw numbers are different across years, we are merely interested in how a player performed relative to others in the same situation. We also have a number that tells us how much the player deviates from mean across all 9 categories. For example, if player A has a value of +0.5 in nine categories, he ends up with a totZ of +4.5. There are many ways to get the same value, however. Player B could have a value of +12.5 in one category and -1 in eight categories to end up at the same score.

This last observation is troubling in some fantasy formats. If a player is an extreme outlier in a few categories and doesn't contribute or hurts in other categories, then his value may actually be reduced quite a bit. Head to head formats, for instance, don't care what the margin of victory is, so accumulating record-setting stats in one category at the expense of several others is not generally a good strategy (although this can work out quite nicely in points leagues). Because of this, it may be helpful to consider another statistics, a normalized z-score. We calculate this as follows:

$$statN_{(i,j)} = \frac{statZ_{(i,j)}}{abs(max(statZ_{(i,j)}), min(statZ_{(i,j)}))}$$

That is, we take each z-score and divide it by the largest possible value to yield a result within the range [-1,1]. Similarly, we can calculate:

$$totN_j = \Sigma_i \, statN_{(i,j)}$$

totN will always be a value within [-9,9]. Going back to our example, Player A is likely to be capped at a statN value of +1 for the category in which he was previously getting +12.5, meanwhile the other 8 categories will get mapped to a value between -1 and 0, so it's reasonable to assume that his totN ends up somewhere around 0, or maybe below. Player B, on the other hand, definitely ends up with a positive score, one that's likely higher than player A.

Note that what we lose with statN is the concept of the z-score itself. Values are now some kind of fractional representation of standard deviations away from the mean, and this fractional representation is different for each stat. However, since we merely performed a linear transformation, we haven't changed the order or the shape of the distribution, merely ensured that the magnitude is the same across all statistics.

Understanding your league and how to interpret the various statistics available to evaluate players can be important in determine player value and optimizing the strength of your team.