Milestone 1

NETIDs: el3418 ac8480

Motivation

We are interested in researching basketball prospects. Specifically, we want to examine the value of draft picks in the NBA. In the NBA, teams take turns picking players from the incoming draft class. Pick order is determined by a lottery - generally, pick order favors the teams with the worst records from the previous season. A wrinkle in this scheme is that teams can elect to trade their draft pick to other teams, and the pick order is determined by the record of the team that originally owned the pick. This incentives teams to go to two extremes - either go all in on a championship by trading away future draft capital for established veteran players in the current market, or commit to a losing record by trading away potentially valuable players to gain future picks from other teams. The questions to be examined are two-fold: how well do a player's abilities transfer from college to the NBA, and, expanding on this question, can we predict the future value a player will bring to their team?

From a data perspective, this question is interesting because of the challenge of quantifying sports performance. There are many ways for players to bring 'value' to a team. Some ways, such as scoring, rebounding, or passing, are more easily quantifiable. Others, such as defensive communication or leadership, are not as quantifiable. In addition, a player's contributions over the course of his career do not always remain constant or grow linearly. Some players' performances fluctuate rapidly depending on their roster or management situation. Yet, others may be late bloomers that take some time to develop until they are ready to contribute meaningfully.

There are multiple ways to approach this problem, but this project will mainly focus on the more statistically quantifiable aspects of basketball, while taking a holistic approach to player performance in order to at least somewhat capture non-obvious aspects of game impact.

Dataset

The data will mostly be scraped from basketball-reference, a popular site for both basic and advanced basketball statistics, as well as its college basketball equivalent, sports-reference. For this project, we will mainly be using the basic counting stats from both a player's college and NBA career (points, assists, rebounds, etc.) Some 'advanced' statistics, such as usage rate, will also be used. However, many of these advanced statistics will be omitted. This is intentional, as many of these advanced statistics are formulated using basic statistics, and would thus bring up issues of multicollinearity. This is because their college stats form a basis for how they are projected to perform in the NBA, thus contributing to their draft pick value. Something that can be taken into consideration is a single year perspective or a sliding moving average for these player statistics.

In 1979, the 3 point line was introduced, which is where we will begin our comparative exploration of how college players are likely to perform in the NBA by starting with the 1979-80 season. To narrow down the scope of the data and make this problem more tractable, certain restrictions will be placed on the dataset. College statistics will be the entirety of a player's college career, as well as all of their NBA career. Players that did not play college basketball will be omitted. Second round draft picks will be omitted from the dataset as well due to most of these players being sent to the G-league for further development. This will cut down on a huge part of the variability that would hinder model performance. The trade-off of these restrictions is that a model trained on this dataset would not be able to predict well on players projected to be drafted in the second round, and would not be able to predict at all on international players that did not play college basketball. However, the majority of NBA players were highly rated college players, and that should still continue to be the case in the future, so any models produced would not suffer massively from these exclusions. A final restriction is that the dataset will only include player's that have played at least 5 seasons in the NBA (approximately the average length of an NBA player's career). This will allow us to create a model that rewards a player for a lengthy playing career (and thus more value added), while not punishing players such as rookies that have not played a large number of seasons yet.

In order to measure player value, we will be using a measure known as VORP (Value Over Replacement Player). This is an estimate of the points a player contributes above a replacement level player. By using this metric as the ground truth for player value, we can see how much more value a player brings compared to an average NBA player. The VORP should be adjusted by weighting based on how many consecutive seasons they played. Win-share calculations or RAPTOR can also be utilized for this purpose for college players to see how much they contributed to winning, both being measures that seek to encapsulate player value. There can also be a way to combine these metrics. We can then test our results using a holdout set of college players with their known ground-truth VORP.

Methodology - What methods do you plan to apply? Why are they appropriate?

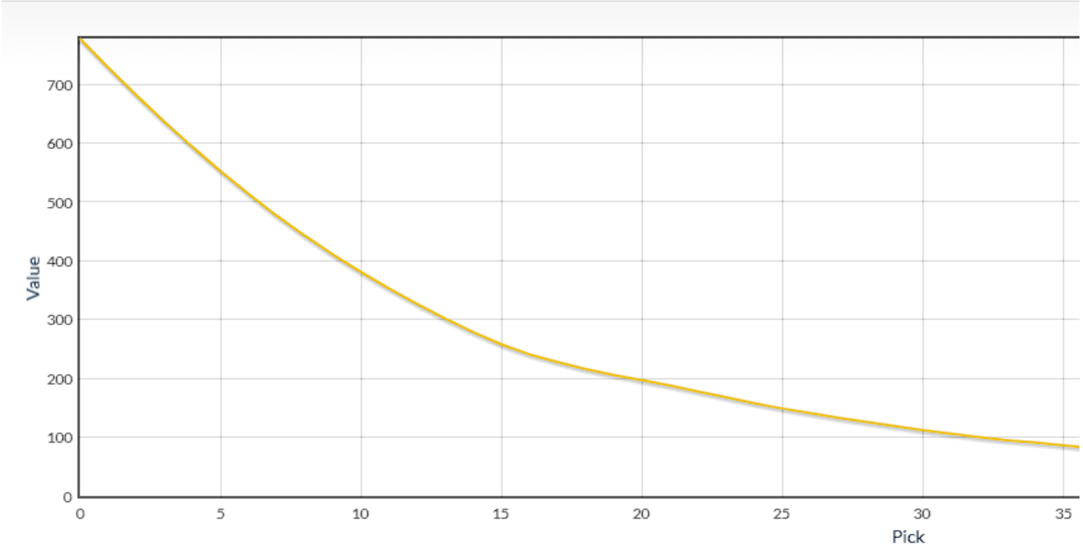
We will check for multicollinearity because most metrics are attempting to accomplish a similar goal in quantifying player performance. Because of this, correlation between certain features will be incredibly large. Correlation heatmaps can be used to view high correlated variables, drop certain variables, or combine them with an aggregate method. Due to the highly correlated nature of the features, it would also be important to examine causal effects and potential outcomes, as well as consider the unobserved counterfactuals.

In addition to the basic stats collected throughout their career, we can also incorporate alternative features such as the altitude of playing locations, game conditions (time remaining, score differential), shot controls (shot type, shot distance), and defense (distance between the shooter and the closest defender, percentage of time a player is the closest defender). These can act as confounders on a player's real potential. We also have a separate dataset that includes player salaries - another feature that may be worth exploring.

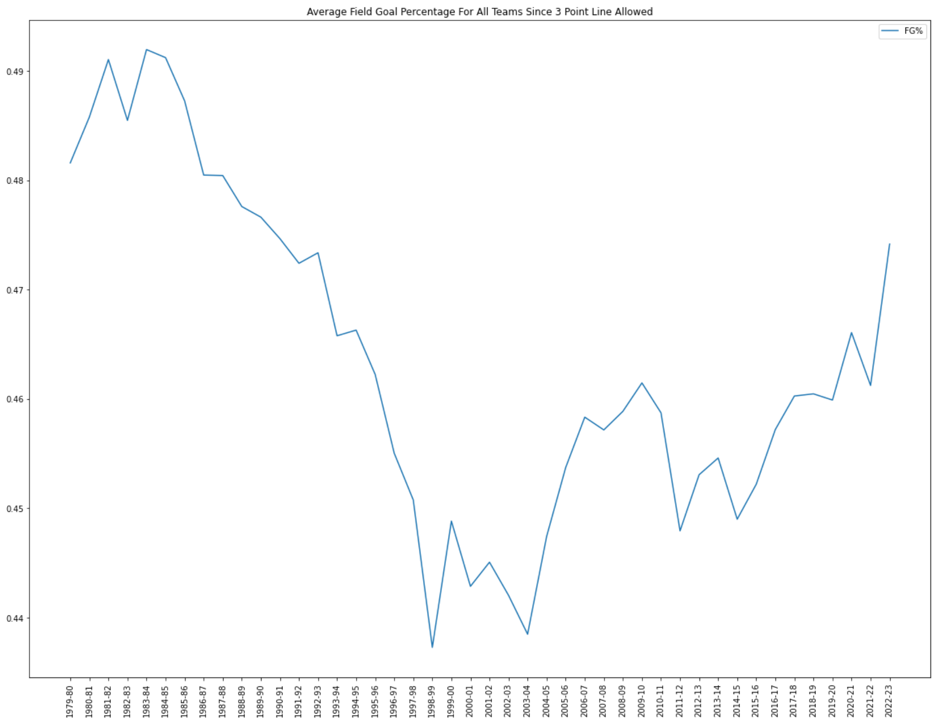
Appendix

Hot hand fallacy:
https://www.sloansportsconference.com/research-papers/the-hot-hand-a-new-approach-to-an-old-fallacy

Average draft value calculated using a 2-year average VORP, sorted by pick order



Average FG% by year since the 3 point line was created


Average Field Goal Percentage For All Teams Since 3 Point Line Allowed
FG%