Alex Vaslow
April 12, 2022

Predicting Future NBA Success from College Stats

Intro

NBA Teams spend a lot of money each year, scouting college players to determine who to draft. What if there was a way to predict NBA success based solely on publicly available data such as college statistics, height, weight, and age? NBA teams could save significant amounts of money on scouting and advanced testing for collegiate athletes. I am proposing a neural network that takes the aforementioned data to predict NBA success evaluated with a single measure.

Evaluating NBA Success

Ideally, we will have one metric to evaluate the career success of drafted players. Basketball Reference has a statistic called Win Shares, abbreviated as WS, which is an estimated number of wins contributed by the player. Since the main goal of NBA teams deciding which players to draft is drafting a player that will help them win, WS seems like a useful stat to grade draftees on. However, players drafted recently will have fewer years to rack up win-shares. To work around this problem, I will divide WS by the number of years since the draft occurred. I will cap the number of years at around 10-12, since few NBA careers are much longer than that (I could play around with this parameter if needed).

Input Vector Parameters

College Stats (use career stats):

• First Year • Last Year • Years Played • Games • Games Started • Minutes Played • Field Goals • Field Goal Percentage	• 2P • 2PA • 2P% • 3P • 3PA • 3P% • FT • PTS • Field Goal Attempts	• FTA • FT% • TRB • Assist • Steals • Blocks • PF • Strength of Schedule • Turnover

All of these stats are positive numbers, so we can scale them down to be between 0 and 1 by dividing by the NCAA record for each of these statistics.

Other Info (also available on Sports Reference)

- Position 0 for G, 0.5 for Forward, and 1.0 for Center or something like that to make position a numerical value
- Height (scaled by the max and min heights of NBA players ever)
- Weight (also scaled similar to height)

- Age when drafted (scaled by youngest and oldest players ever drafted)

Data Set

- Using data from 1990 onward, to account for the implementation of the NCAA 3 point line (so stats are consistent), we will have 30 drafts with about 60 players drafted each year. Out of these 60, at least ⅔ of these are usually college players (but this varies on the year). So, we will have 60 * 30 * ⅔ = **1,200 players to evaluate**. Even if we reserve 200 or so for validation and train on the other 1000 (or some similar ratio), we still have 1000 data points for about 30 input vector parameters.
- Something to consider, some NBA draftees never played a single game which could cause interesting things in the result. This could even be more disrupted by NBA draftees, like Len Bias, who died early in their career or even before they were slated to play a single NBA game. I will consider throwing out data of players who died before fulfilling their NBA potential (since it is impossible to predict freak events like that).
- I will use Python and Selenium to scrape Basketball Reference to get the raw data I need. I plan on saving it to a SQL database so I can easily keep track of it. Then, I will preprocess the data into values between 0 and 1 using python and save it in a format that is usable by Tensorflow. I will then use TensorFlow to fit the input vectors of playerdata to the output vector of WS/year.

How will we know if the project is successful?

Obviously, even the most advanced models or scouts make mistakes when evaluating college players. There is a human element that our network cannot account for. To determine if the network is worthwhile, we can check out the predicted WS/year for players in our validation set. Comparing players within the same draft class to each other, I will rank players by predicted WS/year. I will compare my ranking to the actual order of the NBA draft and evaluate both orders against the actual WS/year players accumulated, and see which one does better. Even if my network does slightly worse than the actual NBA draft order, I would consider the project successful considering that the network never saw any of the draftees play, had no advanced athletic metrics, and had no background knowledge of what makes someone successful in the NBA. If the network cannot predict novel players well, the extra scouting and metrics that NBA teams use are worthwhile.