Sejal Dua
Megan Monroe
COMP152: Sports Analytics
9 March 2021

DRAFT DAY

"A day where lives are changed, fates are decided, dynasties are born, and the clock is always talking. Of course, I'm talking about... Draft Day."
––Chris Berman (ESPN sportscaster)

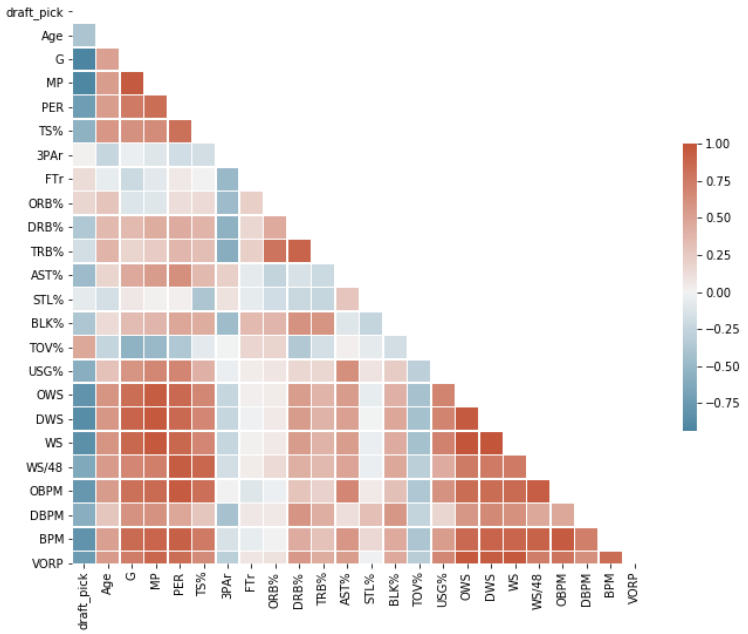
This lab primarily served as an exploration of how historical data can be used to generate insights for future scenarios in the field of sports analytics. The NBA is a prime example of how the relationship between draft pick and player performance is generally non-linear, featuring a steep dropoff after the first ten or so picks are selected, and then a gradual decline thereafter. The objective for this lab will be to attempt to model this complex relationship in a way that can be insightful when comparing relative draft pick value in the context of trade proposals. It is worth mentioning that through the implementation process, I came to discover just how noisy and complex this problem is and why context (e.g. positions needed, salary cap, available / undrafted players, scouting reports, game film, etc.) is of utmost importance when it comes to management-related decisions such as NBA draft negotiations. Ultimately, I would like to use the outputs of a simple linear regression model to showcase why the draft is a beast that must not be over-simplified. That said, this model can be a starting point for more complex, context-aware endeavors in the future.

Data Cleaning and Preprocessing

To begin the project, the player databases and draft databases were read into pandas dataframes, and then the two tables were merged on the foreign key which links players and draft selections: playerID (e.g. abdulka01 for Kareem Abdul-Jabbar). Following this operation, the merged data source was converted into a dictionary in order to speed up the data extraction process and eliminate unnecessary iteration. Grouping the dataframe by playerID created a dictionary with the helpful property of constant time key-indexing access to each player's per-season stats, which were automatically aggregated into lists. In representing a player's contributed value over the course of their NBA career, a design choice was made to average their per-season statistics. However, depending on what a given team needs to get out of a draft, there are arguments to be made for weighting these statistics in such a way that certain periods within their career matter more than others. Finally, a loop was implemented to iterate through each draft pick since 1949 and get the average performance-related statistic for the value that the player contributes on the court. This data was crunched down one step further by binning each player based on his draft pick number and averaging statistics like Age, VORP, BPM, WS, etc. to get a rough snapshot of how draft pick number relates to player value. Using this consolidated data, it was time to start selecting features!

Selecting Strongly Correlated Features

The key aspect of this project was to figure out, among the available per-season statistics of each NBA player from 1980-2019, which variables illustrate the relationship between draft pick and performance value. To do so, a player's draft pick number was used as the target or response variable, and the independent variables or features were selected using Pearsons' correlation among all independent variables. In the lower triangular correlation matrix below, the first column illustrates the correlation between each candidate feature variable in the target variable of draft_pick

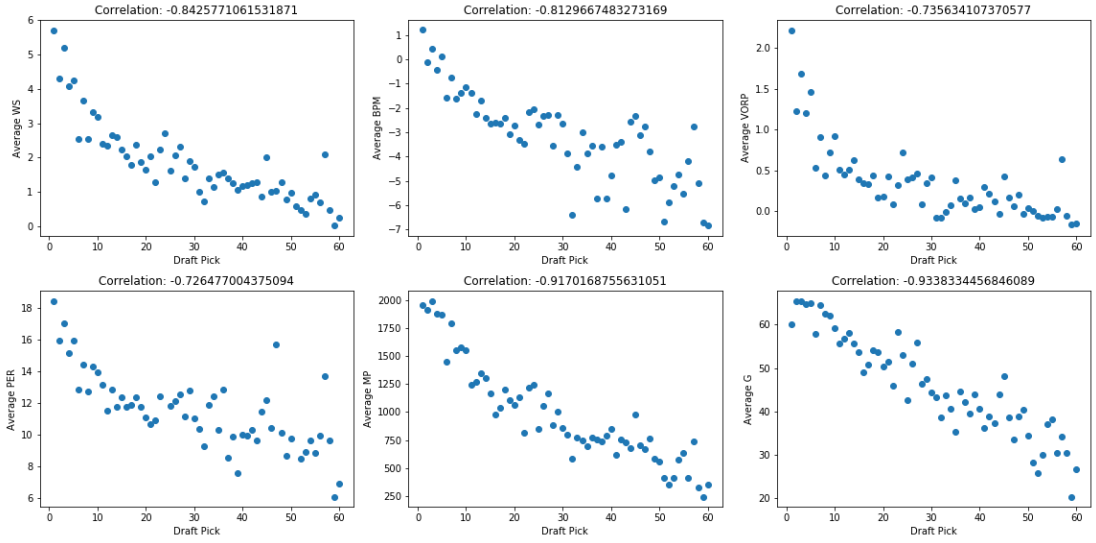


Since we expect players who were drafted early on in the draft to contribute more value than players who were not drafted as high, we want to identify the most negative correlation scores, which are depicted by dark blue squares. Some strong candidates for features include G, MP, PER, OWS, DWS, WS, OBPM, BPM, and VORP.

Linear Regression Assumptions

In picking the most negatively correlated features, we must remember a few key assumptions about linear regressions. Multicollinearity is an issue that occurs when independent variables are too highly correlated with each other. As we can imagine, OWS, DWS, WS, and WS/48 all likely exhibit multicollinearity, since they are computed using the same equation with slight variations. This can be formally tested using a metric called the variance inflation factor (VIF), but for a naive approach, selecting one that logically makes sense and has a strong negative correlation will suffice. We will go with win shares (WS) here because it is a summation of offensive win shares (OWS) and defensive win shares (DWS). Though DWS may have a more negative correlation with draft pick number *on paper*, it seems very wrong to let this noisy data tell us to use DWS over WS, which is logically the more holistic metric. Sports analytics is better suited for problems towards which human judgment can be applied.

Another assumption of linear regression is homoscedasticity. That is, the residuals must all be equal across the regression line; the scatter plot should not depict a funnel shape. When homoscedasticity is present, a non-linear correction can fix the problem. In the context of this project, DWS and WS/48 were both good candidates for feature variables, but WS/48 exhibited heteroscedasticity. A correction could have been applied here, but since WS/48 = 48 * WS / MP, where MP = minutes played, this particular issue could be resolved by doing away with WS/48, and using both WS and MP instead. The heteroscedasticity of average BPM scores and average PER scores was corrected for with a log transformation.



The 6 most negatively correlated features, after applying simple logic to avoid multicollinearity among the variables, were WS, BPM, VORP, PER, MP, and G. For each subset of players drafted at each pick number from 1 through 60, average feature scores were computed in order to yield the scatter plots below. Correlation scores are included as the title for each relationship between x and y. From these graphs, we can expect that win shares, minutes played, and games played are among the most important factors which help us predict the draft pick number of a player given their career-long performance.

Strategy for Estimating Relative Value of Draft Picks

The approach is as follows:

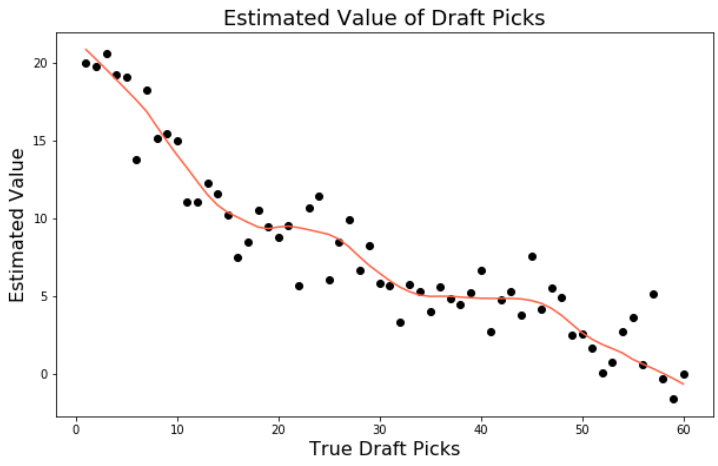
1. Implement a linear regression which will take N observations (unique NBA players) and fit these players' 6 career-long performance variables (see X below) to their true draft pick number, y.
   `X = [x1: WS, x2: BPM, x3: VORP, x4: PER, x5: MP, x6: G]`
   Note: there will also be a constant in the implementation of the linear regression.
2. Feed the same N observations into our fitted model to obtain an estimated draft pick ŷ for each player. ŷ is just a linear combination with weights and features like so:
   `ŷ = w1•x1 + w2•x2 + w3•x3 + w4•x4 + w5•x5 + w6•x6 + const`
3. Create bins from 1 through 60 so that we can implement a mapping from discrete draft pick numbers to estimated value.
   `mappings = {1: [], 2: [], 3: [], ... 58: [], 59: [], 60: []}`
4. Loop through each player, place their ŷ draft pick evaluation into the bin corresponding to their true draft pick number.

5. For each input draft pick number, average the estimated ŷ values to finalize our mapping.
   `mappings = {1: #, 2: #, 3: #, ... 58: #, 59: #, 60: #}`
6. Given user input (e.g. 1, 3, 25), sum or average (depends on if there are top picks being considered or not) the estimated values corresponding to each pick number.
7. If the estimated value of the picks to receive outweighs the estimated value of the picks to give away, designate the trade proposal as advantageous / successful; else designate it as a mistake.

The weights from the linear regression are listed under the `coef` column in the figure below. Also listed are standard error, t statistic, p values, and confidence intervals. It would be wise to do some feature engineering and standardization of the variables here, but I would rather spend time delving into some tricks that I used to make my program more robust and an analytical discussion of the intricacies in the data that make this an incredibly complex problem.

	coef	std err	t	P>|t|	[0.025	0.975]
WS	0.4738	0.694	0.682	0.495	-0.888	1.836
BPM	0.3490	0.219	1.596	0.111	-0.080	0.778
VORP	0.4161	1.228	0.339	0.735	-1.994	2.826
PER	-0.3453	0.140	-2.466	0.014	-0.620	-0.071
MP	-0.0149	0.002	-8.731	0.000	-0.018	-0.012
G	0.0436	0.054	0.806	0.420	-0.063	0.150
const	43.7241	3.051	14.330	0.000	37.738	49.711

After fitting the linear regression model using the statsmodels Python package, the ŷ values were obtained by calling the predict() method, which essentially just computes a linear combination for the input variables given the model's trained weights. The issue is that ŷ is actually an estimate of what a player's draft pick number *should* be given their career-long performance in the NBA. Therefore, a lower ŷ ("estimated value") is actually better. In order to flip it so that a lower draft pick number maps to higher estimated value, we need to set a baseline. The value of each pick should be relative to the worst possible value that is available in the draft: pick number 60. Thus, each mapping was subtracted by the average ŷ value for 60th picks. The result is as follows:



Locally Weighted Scatterplot Smoothing (LOWESS)

Focusing on just the black points on the graph, we can notice that the historical data suggests that the 3rd pick is of higher value than the 2nd pick. It would be absurd to suggest that trading away the 2nd pick for a 3rd pick is a desirable trade. Thus, some form of a correction is necessary. I decided to use a technique called Locally Weighted Scatterplot Smoothing in order to both recognize the non-linear nature of our model's predictions and also to correct for noise caused by standout players who ended up being very valuable despite being snagged later in the draft. Given each draft pick number x_i, LOWESS works by taking the span (frac*N) closest points to (x_i, y_i) based on their x values and estimating y_i using a weighted linear regression. The frac variable as part of the span of closest data points to use for this smoothing estimation was set to 0.2, meaning approximately 12 nearby estimates were used to formulate each part of the red line in the graph above.

While this model is, by no means, perfect, what is nice about it is that it reasonably demonstrates the phenomenon that the first 14 picks in the NBA draft are worth far more than later picks. We can observe that the picks toward the end of the first round (15th-30th) do depreciate in value slightly, but not nearly as much as the top picks. The second round of the draft is pretty hit or miss when it comes to selecting players who will add value to a given organization. The 30th to 45th picks are all roughly interchangeable in terms of trading value, and then value trails off toward the end of the draft, though there have certainly been some players who earned their keep despite being brushed off in the draft.

Discussion of Limitations and Ideas for Optimization

NORMALIZATION → One huge limitation of my draft curve is that it is not normalized. After stumbling upon Kevin Pelton's 2017 draft curve on nbasense.com, I realized that I should have normalized / rescaled my y axis so that the range of values for each draft pick is from 1 to 100. Without doing that, my comparisons between summed give value and summed receive value may be off-base due to the scale being off. Furthermore, lots of other feature engineering could have been applied to this problem. For instance, expected value differentials for both VORP and WS have been shown to be quite indicative of draft value proposition, as well as many variations of box plus-minus (BPM).

COLLEGES → It occurred to me that lots of athletes who are chalked up to be superstars in the NBA end up flopping because they just have a hard time adjusting to the "big leagues," where the 3 point line is further back and the competition is much more elite. That said, athletes from the Blue Bloods tend to acclimate pretty well to the NBA. If there was some way that the textual College / University field could be turned into a qualitative feature, that feature might be valuable in this context. After doing a bit of digging, turns out there is a Talent Score metric that has been ascribed to a few well-known schools. This may be a good starting point. *Reference*: Ranking the Top 15 College Basketball Programs by Their Current NBA Talent

COST → The NBA Rookie Scale is also something to consider in the context of a trade proposal. First round picks can sign for as much as 120% and as little as 80% of the rookie scale. Using this scale as a feature in the linear regression would likely give weight to the fact that mid-round picks are better bargains (i.e. they are cost-effective relative to the other picks). Moreover, smaller franchises like the Portland Trailblazers are perpetually in a state of draft *purgatory*-- not good enough to win a championship, not bad enough to select the cream of the crop, and not big enough to afford multiple elite picks.

Young & International Talent (*Reference*: On the NBA's G League Ignite, Jalen Green gets paid to prep for the draft)

Ripeness upon entering the draft is a huge topic of discussion right now as the NBA is currently undergoing an experiment to enrich talent development. The NBA G League Ignite program is an "ambitious and unprecedented project in which teenage prospects are getting groomed for the NBA by an unlikely source: the NBA." The G League is NBA' long underutilized developmental league which is now being revamped to give young hopefuls a head start and an opportunity to get paid for their abilities in the 1+ year between high school and joining the league. In doing so, the project helps the NBA fend off global competitors hoping to poach talent. The challenge with young high school prospects is that you could strike gold with a Kevin Garnett, Kobe, or LeBron, or you could get an underdeveloped high school star with big dreams and insane handles, but no minutes on the court. It is a huge risk that is very difficult to make. On the contrary, international talent is also quite difficult to translate to NBA value. The famous story of Jerry Krause recruiting Toni Kukoc while Michael Jordan was putting the team on his back in Chicago highlights this phenomenon quite well. Most notably, though, Manu Ginobili being drafted out of Argentina at 57th pick overall speaks to the fact that there are many outliers in this data. Who would have seen that coming? I don't think anyone would call him underrated at the time. While we are on the topic of the Spurs, Tony Parker went 28th in the draft whereas Tim Duncan went 1st, but I think it is fair to say that both were instrumental in leading the Spurs to a championship, so it is really tough to use a model like this for anything more than a sanity check. What is, perhaps, more worthwhile is building a model to predict who is likely to be a standout player relative to their projected draft pick number, and who is likely to be a bust. If it is proposed to trade a number 10 pick away in exchange for a number 11 pick, using this model alone is pretty meaningless, but making sure that the GM makes the most optimal pick of the options available is what is key. A bust like Greg Oden can have long-term ramifications for the morale of sports fans for years to come.

STEALS (*Reference*: Manu Ginobili and the 10 Biggest Round 2 Steals in NBA Draft History)

- Rashard Lewis (1998) - No. 32
- Carlos Boozer (2002) - No. 34
- Maurice Cheeks (1978) - No. 36
- Dennis Rodman (1986) - No. 27
- Manu Ginobili (1999) - No. 57

INJURY (*Reference*: Eight NBA stars who were robbed of greatness by big injuries)

- Anfernee Hardaway (1993) - No. 3 - legs deteriorated
- Grant Hill (1994) - No. 3 - started off with 6 All-Star visits, broke ankle in 7th season, still Hall of Fame
- Jay Williams (2002) - No. 2 - severe leg injuries in motorcycle accident
- Chris Bosh (2003) - No. 4 - third member of the Big Three - blood-clotting condition
- Shaun Livingston (2004) - No. 4 - 18 years old - knee injury - was able to recover & reinvent
- Brandon Roy (2006) - No. 6 - knee injury, retired after 5 seasons, then played 5 games with T-Wolves
- Greg Oden (2007) - No. 1 - knee issues - unsuccessful comeback with Heat
- Derrick Rose (2008) - No. 1 - Rookie of the Year, MVP, 3 x All Star - torn ACL - revitalized career

BUSTS (according to my brother and I because these players went in the top three, yet nobody talks about them… literally ever):

- Anthony Bennett (2015) - No. 1
- Derrick Williams (2011) - No. 2

Darko Milicic (2003) - No. 2
Hasheem Thabeet (2009) - No. 2
Adam Morrison (2006) - No. 3
Kwame Brown (2001) - No. 1