Algorithmic NBA Player Acquisition

Ryan S. Brill*, Justin Hughes†, and Nathan Waldbaum†

November 21, 2023

Abstract

Player acquisition is one of the fundamental problems of basektball analytics. An analyst may be tempted to recommend simply acquiring the best available player, where best is defined by an all-encompassing skill metric. How a player fits with his teammates, however, is also important in determining the effectiveness of a lineup. Thus in this paper we model the effectiveness of a lineup as an interacting function of the offensive skill, defensive skill, and player archetype of all ten players on the court. We find that fit is indeed just as essential as skill in crafting a lineup.

1 Introduction

Roster construction is one of the fundamental problems that an NBA front office faces. NBA general managers want to sign free agents and trade for players who improve the team's ability to win games. Also, given the roster NBA coaches want to start the best available five-man combination of players. Mathematically, each of these problems (free agency, trading, and setting a lineup) rests on being able to estimate the effectiveness of a (potentially unseen) five-man lineup. For instance, given a solidifed set of four players in the starting lineup, a general manager may want to add a free agent who maximizes the effectiveness of the lineup.

An analyst may be tempted to recommend simply adding the best available player, where best is defined by an omnipotent skill metric. All-in-one metrics like RAPM, RPM, PIPM (RIP), LEBRON, BPM, and DARKO attempt to distill player skill into just one number. Given a team's current players, the most skilled available player is not necessarily the best acquisition. Beyond skill, a player's fit, or the interaction of his role with the roles of his teammates, contributes to

*Graduate Group in Applied Mathematics and Computational Science, University of Pennsylvania. Correspondence to: ryguy123@sas.upenn.edu

†Wharton School, University of Pennsylvania.

the effectiveness of a lineup. Hence in this work, we estimate the effectiveness of a lineup as a function of the offensive skill, defensive skill, and player archetype of all ten players on the court. From this model we create a player acquisiton algorithm. We find that fit is just as essential as skill in crafting a lineup.

2 Data and model specification

We begin with a brief overview of our dataset of NBA possessions and identify several variables that may be predictive of the outcome of a possession. We then create player archetypes and introduce our Bayesian regression model.

2.1 Data

First we obtained data that we use to cluster players into *player archetypes*. A player archetype should capture a player's role in a lineup, or *how* he plays, not *how well* he plays. Hence we obtain data for each player-season that quantify how he plays, including totals, per 100 possession stats, shooting distances, time with ball, rated-based location data, advanced passing, driving/catch and shoot/pull-up rates, and post-up and paint frequencies. More specifically, using player-season data from NBA.com, BasketballReference, and CraftedNBA, we compiled 48 metrics for each player-season that capture a player's style: FTPCT, TSPCT, THPAr, FTr, TRBPCT, ASTPCT, AVGDIST, Zto3r, THto10r, TENto16r, SIXTto3PTr, HEIGHT, WINGSPAN, FRNTCTTCH, TOP, AVGSECPERTCH, AVGDRIBPERTCH, ELBWTCH, POSTUPS, PNTTOUCH, DRIVES, DRFGA, DRPTSPCT, DRPASSPCT, DRASTPCT, DRTOVPCT, DRPFPCT, DRIMFGPCT, CSFGA, CS3PA, PASSESMADE, SECAST, POTAST, PUFGA, PU3PA, PSTUPFGA, PSTUPPTSPCT, PSTUPPASSPCT, PSTUPASTPCT, PSTUPTOVPCT, PNTTCHS, PNTFGA, PNTPTSPCT, PNTPASSPCT, PNTASTPCT, PNTTVPCT, and AVGFGATTEMPTEDAGAINSTPERGAME. We exclude players who played fewer than 1000 minutes in a season to make sure each player has a representative sample. From this data we fit K player archetypes, described in Section 2.4.

Next we obtained data that we use to cluster five-man lineups into *lineup superclusters*. A lineup supercluster should capture *how* a five-man lineup plays together as a unit, not *how well* it plays together. Hence we obtain data for each lineup-season that quantify how it plays, including traditional, advanced, miscellaneous, four factor, and scoring data per possession. More specifically, using lineup-season data from the NBA.com lineup tool, we compiled the following weighted metrics that capture a lineup's style: WFGMPercUAST, WFGMPercAST, WThreeFGMPercUAST, WThreeFGMPercAST, WTwoFGMPercUAST, WTwoFGMPercAST, WPercPTSPITP, WPercPTSOffTO, WPercPTSFT, WPercPTSFBPS, WPercPTS3PT, WPercPTSMR, WPercPTS2PT, WPer-

cFGA3PT, WPercFGA2PT, WOppTORATIOpercent, WOppFTARATE, WPACE. From this data we fit lineup superclusters, described in Section 2.5.

We use DARKO to measure player offensive and defensive skill. DARKO (Daily Adjusted and Regressed Kalman Optimized projections) combines bayesian inference with machine learning to develop game-by-game estimates of how well a player is likely to perform in the future (Medvedovsky and Patton, 2022). We use DARKO because, to our knowledge, it is the most predictive all-in-one skill metric for NBA players (see Figure 1, a Tweet from DARKO creator Kostya Medvedovsky). We obtained the pre-season DARKO for each player-season in our possession-level dataset (described below). We also obtained player salary data from HoopsHype.

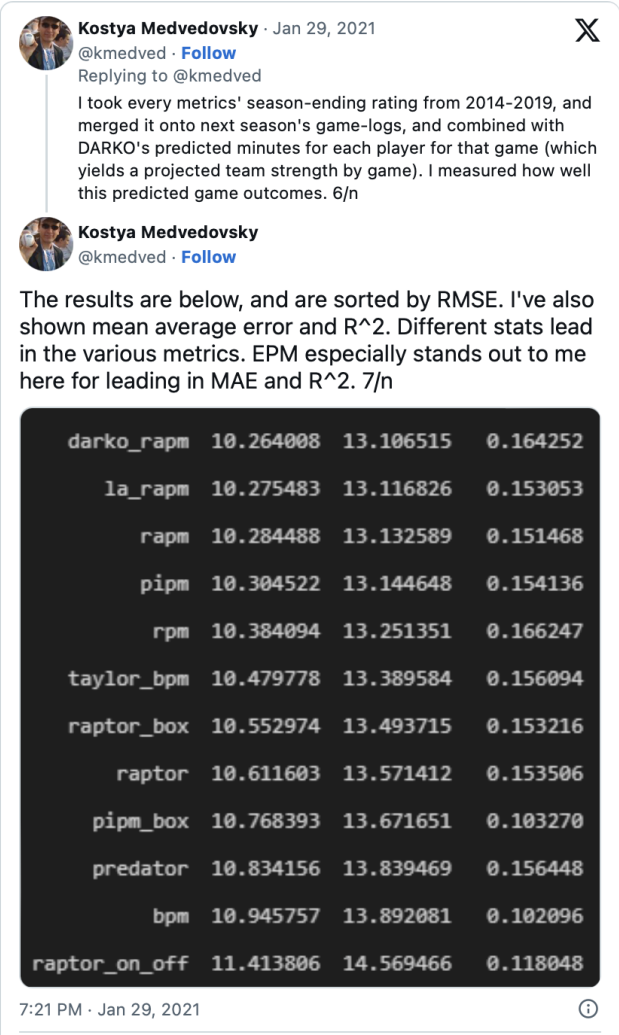


Figure 1: Evidence that DARKO is the most predictive all-in-one NBA player skill metric.

Finally, we obtained possession-by-possession data, which we use to fit our model that estimates the outcome of a possession. From Ramiro Bentes' GitHub page we got a PBP dataset that includes

lineups, teams, scores, time, and playtype data for every event from the 2022-23 season. For simplicity, in this paper we use data from just the 2022-2023 NBA season.

To measure the outcome of the i^{th} possession y_i we use expected net points, or the points scored by the offensive team minus points given up in transition. We subtract points given up in transition because transition points are generally regarded to be the result of bad play by the offense. We define transition play as within 7 seconds of a turnover.[1]

2.2 Modeling the outcome of a possession

Our goal is to model the outcome of a possession as a function of the skills and archetypes of all ten players on the court. If the outcome of the i^{th} possession y_i were purely a function of the offensive skills (DARKO) X_{ij}^{off} and defensive skills (DARKO) X_{ij}^{def} of each player j on the court, we could model

$$\mathbb{E}y_i = \beta_0 + \beta^{off} \cdot \left(\sum_{\substack{\text{off. player} \\ j=1,\ldots,5}} X_{ij}^{off} \right) - \beta^{def} \cdot \left(\sum_{\substack{\text{def. player} \\ j=1,\ldots,5}} X_{ij}^{def} \right). \tag{2.1}$$

This is an additive model that says the expected outcome of a possession is a weighted difference between the combined offensive skills of the offensive lineup and the combined defensive skills of the defensive lineup. Having differing coefficients β^{off} and β^{def} for the offensive and defensive lineups, respectively, allows the relative impact of offensive and defensive skills to differ. In the NBA, we expect offensive skill to have a larger impact than defensive skill.

The primary weakness with this simple model is the multiplicative impact of an offensive player's offensive skill β^{off} is constrained to be the same for all five players on the court (similarly for the defense). This is not true in the NBA: the marginal impact of a player's skill on the outcome of a possession depends on his role or position in the context of his five-man lineup and the opposing five-man lineup. For instance, consider a lineup with "offensive juggernaut" Lebron James. James usually has the ball in his hand and tends to draw extra defenders towards him, which increases the impact of "3&D" players who can catch-and-shoot and decreases the impact of "playmaking initiating guards" whose skillset is redundant to James'. Now consider a lineup with Giannis Antetokounmpo, the best "non-shooting, defensive minded big" in the NBA. The impact of Antetokounmpo's offensive skills, dominating the paint area, is muted against a Miami Heat defense that crowds the paint area. Finally, consider another "non-shooting, defensive minded big" Rudy Gobert. The impact of Gobert's defensive skills, rim protecting, is muted against a great shooting team like the Warriors who can spread the floor and shoot from distance. Hence the impacts β^{off}

[1] https://halfcourthoops.substack.com/p/nba-defense-transition

and β^{def} of a player's skill should depend on his archetype a in the context of all ten players of the court, or the matchup.

A matchup $M = (L^{off}, L^{def})$ is a combination of an offensive archetype-lineup $L^{off} = \{a_1^{off}, ..., a_5^{off}\}$ and a defensive archetype-lineup $L^{def} = \{a_1^{def}, ..., a_5^{def}\}$, where a_j^{off} is the archetype of offensive player j (similarly for defensive player j). Given enough data for each matchup, we would fit separate coefficients $\beta_{a,m}^{off}$ and $\beta_{a,m}^{def}$ for each archetype a in matchup M. In that case, we would model

$$\mathbb{E}y_i = \beta_{0,M_i} + \sum_{\substack{\text{off. player} \\ j=1,...,5}} \beta_{a,M_i}^{off} \cdot X_{ij}^{off} \cdot \mathbb{I}\left(\begin{smallmatrix}\text{off. player } j \\ \text{has archetype } a\end{smallmatrix}\right) - \sum_{\substack{\text{def. player} \\ j=1,...,5}} \beta_{a,M_i}^{def} \cdot X_{ij}^{def} \cdot \mathbb{I}\left(\begin{smallmatrix}\text{def. player } j \\ \text{has archetype } a\end{smallmatrix}\right), \quad (2.2)$$

where M_i is the matchup in possession i.

In practice, given the amount of data we have, there are far too many unique matchups to fit separate coefficients for each matchup. Specifically, we fit $K = 8$ player archetypes in Section 2.4, so there are $K^5 = 32,768$ unique archetype-lineups. In practice the vast majority of archetype-lineups feature at most two players of the same archetype, so there are effectively

$$\binom{K}{5} + \binom{K}{4} \cdot \binom{4}{1} + \binom{K}{3} \cdot \binom{3}{2} = 504 \quad (2.3)$$

unique archetype-lineups. In the the 2022-23 season there were 182 observed unique archetype-lineups and 7,203 unique matchups.

To make fitting our model tractable we reduce the dimensionality of the set of matchups. The idea is to cluster archetype-lineups into *lineup superclusters* who play similar styles of basketball. In Section 2.5 we create $K' = 6$ lineup superclusters, capturing how a five-man archetype-lineup plays together as a unit (not how well it plays together). Then, a matchup $m = (l^{off}, l^{def})$ is one of 36 combinations of an offensive supercluster l^{off} and a defensive supercluster l^{def}. We model

$$\mathbb{E}y_i = \beta_{0,m_i} + \sum_{\substack{\text{off. player} \\ j=1,...,5}} \beta_{a,m_i}^{off} \cdot X_{ij}^{off} \cdot \mathbb{I}\left(\begin{smallmatrix}\text{off. player } j \\ \text{has archetype } a\end{smallmatrix}\right) - \sum_{\substack{\text{def. player} \\ j=1,...,5}} \beta_{a,m_i}^{def} \cdot X_{ij}^{def} \cdot \mathbb{I}\left(\begin{smallmatrix}\text{def. player } j \\ \text{has archetype } a\end{smallmatrix}\right), \quad (2.4)$$

where m_i is the matchup in possession i. Combining the offensive skills X_{ij}^{off} of all offensive players j of archetype a on the court during possession i into Z_{ia}^{off} (similarly for the defense), our model is equivalently expressed by

$$\mathbb{E}y_i = \beta_{0,m_i} + \sum_a \beta_{a,m_i}^{off} \cdot Z_{ia}^{off} - \sum_a \beta_{a,m_i}^{def} \cdot Z_{ia}^{def}. \quad (2.5)$$

We initially fit this regression model using ordinary least squares but found that some of the signs of the β parameters were wrong. The parameters β^{off} should be positive because an increase in offensive skill should be lead to an increase in the predicted net points of a possession. Similarly, the parameters β^{def} should be positive because an increase in defensive skill should be lead to a decrease in the predicted net points of a possession. To constrain the signs of these parameters, we use weakly-informative diffuse priors

$$\beta^{off}_{a,m_i} \sim \mathcal{N}_+(0,5^2) \quad \text{and} \quad \beta^{def}_{a,m_i} \sim \mathcal{N}_+(0,5^2). \tag{2.6}$$

We use standard normal priors for all other parameters.

2.3 Modeling the effectiveness of a lineup

The expected outcome of a possession $\mathbb{E}y$, which we model in Equations (2.5) and (2.6), is implicitly a function $\mathbb{E}[y(\ell_1,\ell_2)]$ of the offensive lineup ℓ_1 and the defensive lineup ℓ_2. A lineup ℓ is a set of five players' offensive skills, defensive skills, and player archetypes. We evaluate $\mathbb{E}y = \mathbb{E}[y(\ell_1,\ell_2)]$ by constructing the combined skill Z of all players of archetype a from the individual player skills X and by constructing the matchup index m from the archetypes a of all ten players on the court.

Then we define the value $\mathbf{v}(\ell_1,\ell_2)$ of lineup ℓ_1 versus an opposing lineup ℓ_2 by the difference in expected net points when ℓ_1 is on offense versus defense,

$$\mathbf{v}(\ell_1,\ell_2) := \mathbb{E}[y(\ell_1,\ell_2)] - \mathbb{E}[y(\ell_2,\ell_1)]. \tag{2.7}$$

We then define the value of lineup ℓ_1 as the average value of ℓ_1 against each of last year's playoff teams,

$$\mathbf{v}(\ell_1) := \frac{1}{16} \sum_{\ell_2 \in \substack{\text{last year's} \\ \text{playoff teams}}} \mathbf{v}(\ell_1,\ell_2). \tag{2.8}$$

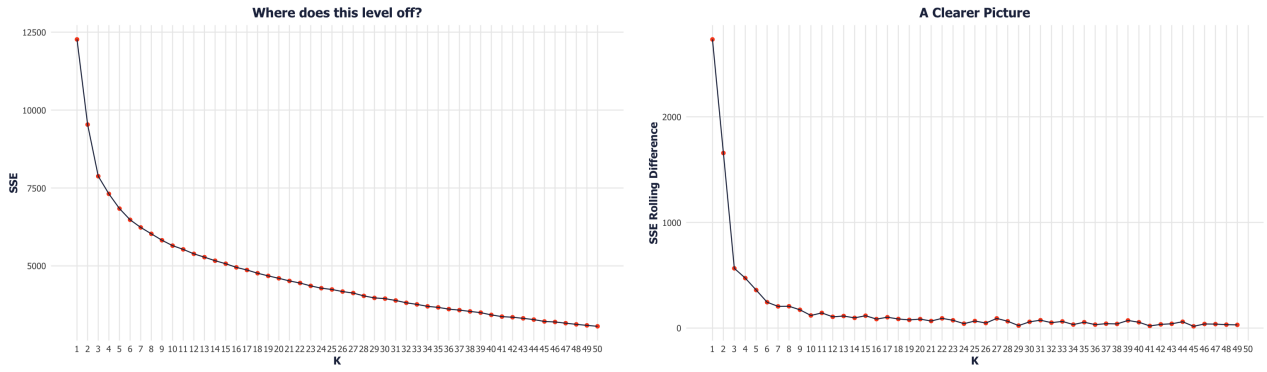
We pit the lineup ℓ_1 against just playoff team's because we want to optimize for being successful against good teams with the ultimate goal of pursuing a championship.

2.4 Fitting player archetypes

We cluster players into archetypes that describe their role within a lineup via K-means clustering.[2] We cluster on the 48 variables describing how a player plays (detailed in Section 2.1). We use

[2]We used Alex Stern's K-means clustering code from https://alexcstern.github.io/hoopDown.html.

$K = 8$ archetypes because the rolling difference in the sum of squared error for K-means levels off around there (see Figure 2b). Also, the assigned player archetypes for $K = 8$ passed the sniff test (i.e., they looked reasonable).



(a) (b)

Figure 2: On the left: the sum of squared error for K-means clustering of player archetypes as a function of K. On the right: the rolling difference in the sum of squared error for K-means clustering of player archetypes as a function of K.

We summarize the $K = 8$ player archetypes as follows,

1. Scoring Wings,
2. Non-Shooting, Defensive Minded Bigs,
3. Offensive Minded Bigs,
4. Versatile Frontcourt Players,
5. Offensive Juggernauts,
6. 3&D,
7. Defensive Minded Guards,
8. Playmaking, Initiating Guards.

We visualize the statistical makeup of each player archetype in Figure 3.

2.5 Fitting lineup superclusters

We cluster archetype-lineups, or five-man combinations of archetypes, into superclusters that describe how a five-man lineup plays together as a unit via K-means clustering. We cluster on weighted archetype-lineup data describing how a archetype-lineup plays (detailed in Section 2.1). Specifically, for each of the 182 observed unique archetype-lineups, we take a weighted average of each lineup statistic from Section 2.1, weighting by the number of minutes played by each observed five-man lineup. We use $K' = 6$ superclusters because the rolling difference in the sum

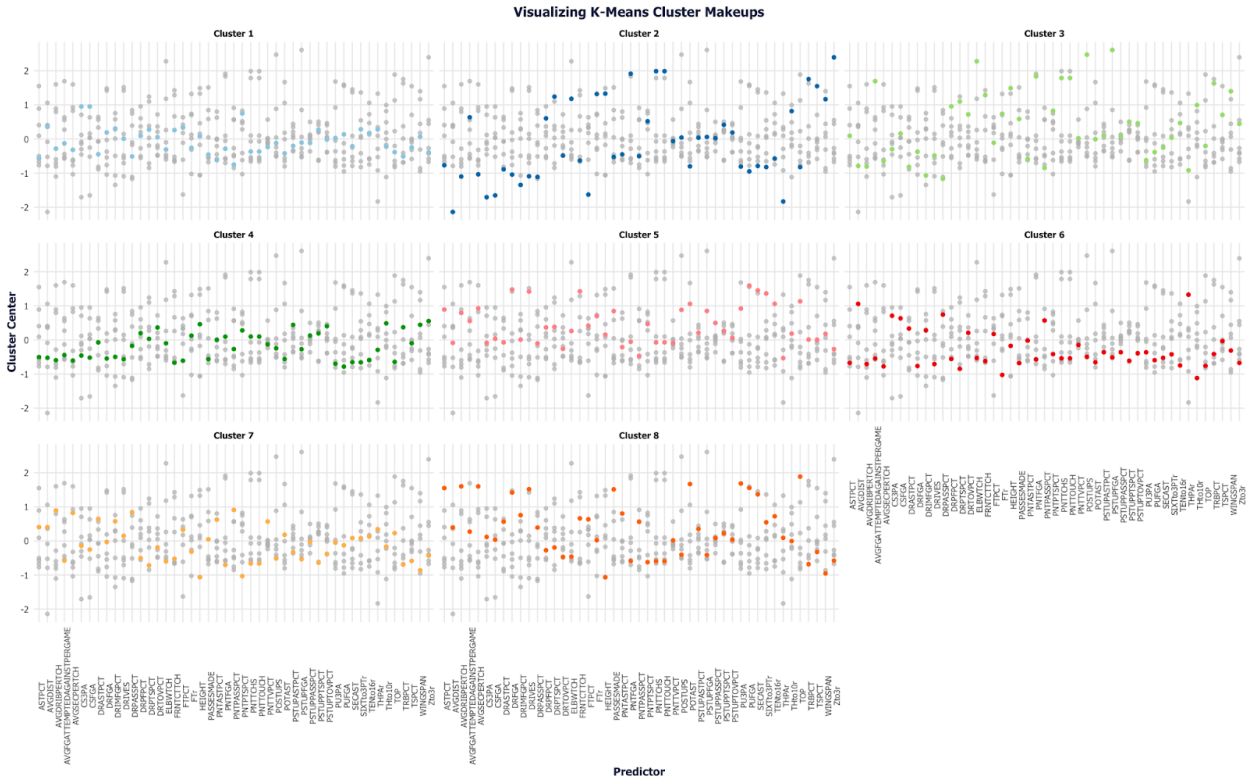
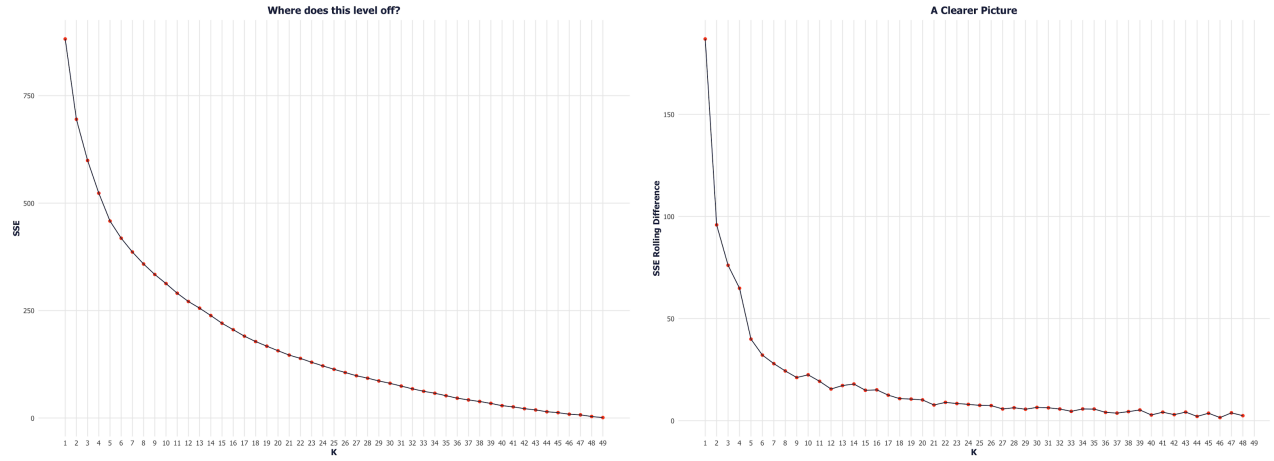


Figure 3: Visualizing our K player archetypes derived from K-means clustering.

of squared error for K-means levels off around there (see Figure 2b). Also, the assigned lineup superclusters for $K' = 6$ passed the sniff test (i.e., they looked reasonable).



(a) (b)

Figure 4: On the left: the sum of squared error for K-means clustering of lineup superclusters as a function of K. On the right: the rolling difference in the sum of squared error for K-means clustering of lineup superclusters as a function of K.

We summarize the $K = 6$ lineup superclusters as follows,

1. Three-Point Symphony,
2. Half-Court Individual Shot Creators,
3. Slashing Offenses,
4. All-Around with Midrange,
5. Chaos Instigators,
6. Up-Tempo Distributors.

We visualize the statistical makeup of each lineup supercluster in Figure 5.

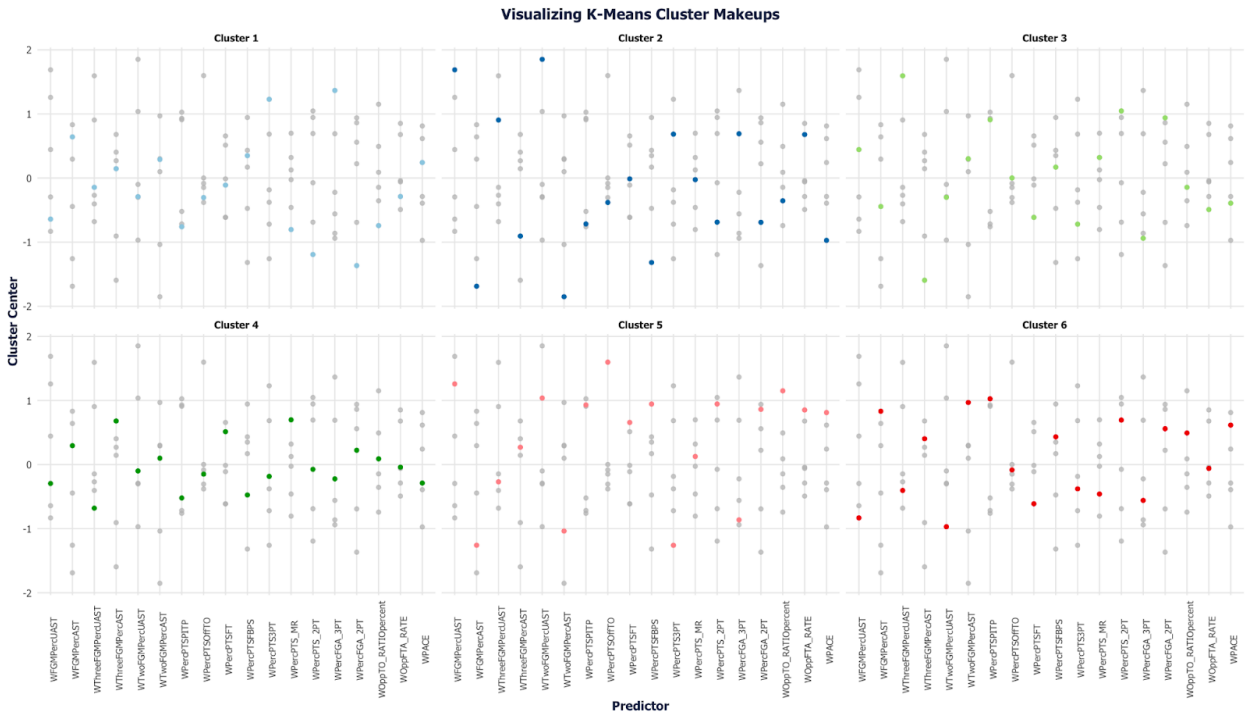


Figure 5: Visualizing our K lineup superclusters derived from K-means clustering.

3 Results

To obtain our posterior samples, we run one MCMC chain for 10,000 iterations. We implement our sampler in Stan (Carpenter et al., 2017) and perform our MCMC simulation using the **rstan** package (Stan Development Team, 2022). The Gelman-Rubin $\hat{R}$ statistic is less than 1.1, suggesting convergence (Gelman and Rubin, 1992). It took about 18 hours to run the chain.

Now we apply our model for the effectiveness of a lineup from Formula (2.8) to conduct player acquisition. We focus on a specific scenario: given four solidifed starters, which fifth player should

we add to the lineup? We search over all players in our dataset who made less than $25 million last season.

First we consider the Los Angeles Lakers. Near the 2022-2023 trade deadline, they traded for Rui Hachimura to join their core of LeBron James, Anthony Davis, and Austin Reaves. The Lakers had a solid four-man lineup but still needed a fifth starter. An analysis considering traditional positions in the NBA would suggest that the Lakers need a point guard. The Lakers followed suit and traded for D'Angelo Russell. We classify LeBron James as an "offensive juggernaut", corresponding to offensive versatility and ball-dominance, making the ball-handling skills of a point guard redundant. According to our model, 3&D and defensive-minded guards are predicted to fit better alongside the Lakers core. In Figure 6 we highlight three such players who are highly recommended by our model.

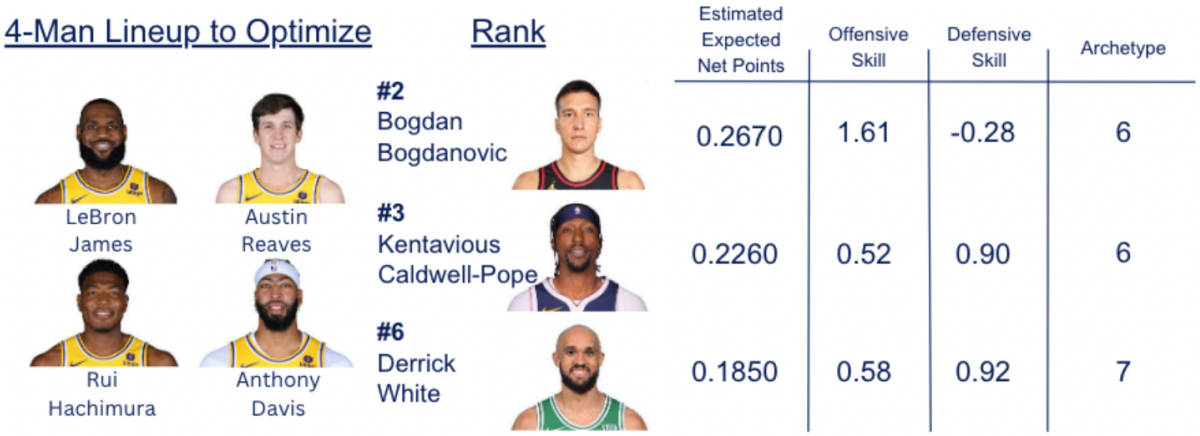
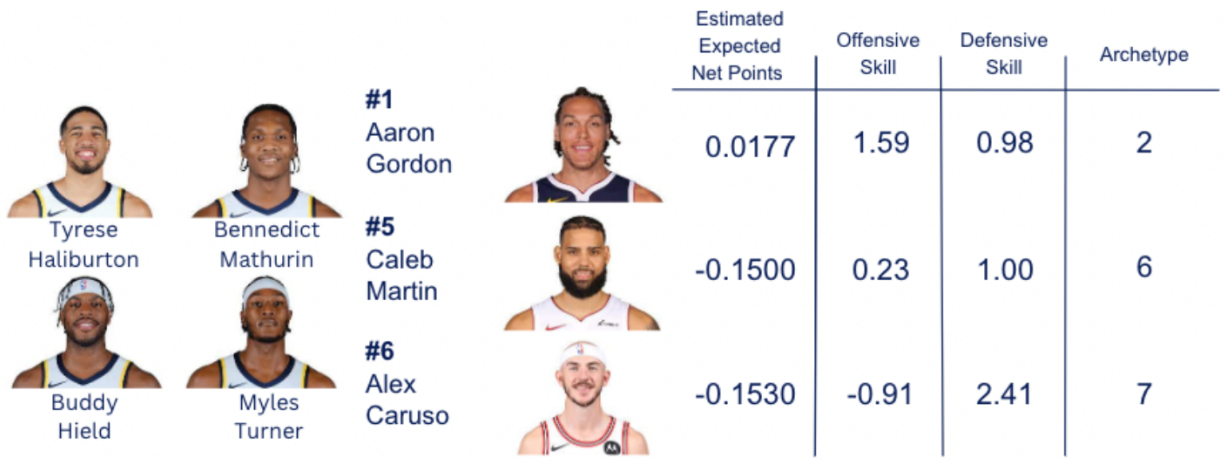


Figure 6: Players recommended to join the Lakers core.

Next we consider the Indiana Pacers. Near the 2022-2023 trade deadline the Pacers were a middle of the pack team that needed a spark. Their core four consisted of young star point guard Tyrese Haliburton, rookie Bennedict Mathurin, solid veteran shooting guard Buddy Hield, and solid veteran big man Myles Turner. Conventional wisdom regarding traditional positions would suggest that the Pacers need a power forward. But, given that three of the four Pacers had negative defense skill ratings, our model recommends acquiring any defensive-minded archetype, inlcuding a defensive big, 3&D, or defensive guard. Simply put, we believe the Pacers needed to prioritize defense. In Figure 7 we highlight three such players who are highly recommended by our model.

Finally we consider the Phoenix Suns prior to the start of the 2023-2024 season. The Suns created an unprecedented pairing of three offensive juggernauts: Bradley Beal, Devin Booker, and Kevin Durant. Over the summer, the Suns front office debated whether to keep or trade offensive big Deandre Ayton. The Suns ended up trading Ayton for fellow offensive big Jusuf Nurkic. Our



	Estimated Expected Net Points	Offensive Skill	Defensive Skill	Archetype
#1 Aaron Gordon	0.0177	1.59	0.98	2
#5 Caleb Martin	-0.1500	0.23	1.00	6
#6 Alex Caruso	-0.1530	-0.91	2.41	7

Figure 7: Players recommended to join the Pacers core.

model believes that a defensive big would have been a better fit, shown in Figure 8. Interestingly, according to our model Nurkic was the best available offensive big (but he was still predicted to be a worse fit than any of these defensive bigs.)

	Estimated Expected Net Points	Offensive Skill	Defensive Skill	Archetype
#3 Nic Claxton	0.0751	0.51	1.16	2
#6 Kevon Looney	0.0732	0.24	1.58	2
#11 Onyeka Okongwu	0.0567	0.57	0.72	2

Figure 8: Players recommended to join the Suns core.

4 Discussion

Player acquisition is one of the fundamental problems of basektball analytics. An analyst may be tempted to recommend simply acquiring the best available player, where best is defined by an all-encompassing skill metric. How a player fits with his teammates, however, is also important in determining the effectiveness of a lineup. Thus in this paper we measure the effectiveness of a lineup as a function of the offensive skill, defensive skill, and player archetype of all ten players on the court. We create a novel model which captures interactions between the skills and roles of

all ten players. We find that fit is just as essential as skill in crafting a lineup.

Although we improve upon the state-of-the-art, our analysis is not without limitations. First, the impact of skill may be nonlinear. Perhaps the impact of skill is much heavier in the tails than in the fat of the distribution. Using logit-transformed skill should remedy this. Additionally, in this work we hard-cluster players into archetypes and hard-cluster lineups into superclusters using K-means clustering. In reality, while certain players are traditional single-position players (e.g., Rudy Gobert is a quintessential defensive minded big), other players fall near the border of two or even three positions. In future work we recommend exploring soft-clustering players and lineups (i.e., fitting the probability that each player belongs to each cluster) using, say, a Gaussian mixture model. In that scenario it is straightforward to modify Formula (2.5) to incorporate soft-clusters.

References

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1):1–32.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472.

Medvedovsky, K. and Patton, A. (2022). *Daily Adjusted and Regressed Kalman Optimized projections — DARKO*. https://apanalytics.shinyapps.io/DARKO/.

Stan Development Team (2022). *RStan: the R interaface for Stan*.