USING CLUSTERING METHODS TO REDEFINE POSITIONS AND ANALYZE EFFECTIVE LINEUPS IN THE NBA

Sebastien Amato / Gabriel Herrera-Lim / Yixiao Zhang

Abstract

The game of basketball has evolved dramatically since the day it was first played in 1891 by Dr. James Naismith. In the NBA, the game used to be played around the biggest player on the floor - the center. This was one of the five traditional positions established in the NBA, with the point guard, shooting guard, small forward, and power forward. These dominant big men towered over less athletic players, which made teams craft their playbooks around them, both on offense and on defense. As the game progressed, teams started shifting their focus to faster, more athletic scorers playing the shooting guard and small forward positions. In the modern game, players have begun diversifying their skills and become more athletic. This has resulted in a more "positionless" game, with players no longer conforming to the siloed description of their listed positions.

With more than a decade of NBA player stats, revolving around efficiency, production, relative value added, and tendencies, we were able to use unsupervised machine learning methods to cluster players into different modern playing styles and use those clusters to identify possible replacement players for each team's "weakest link". Upon generating each player cluster, we viewed their distribution of stats along a scaled spectrum to understand and determine an appropriate role for that cluster. We then used a linear regression and random forest model to identify cluster interaction and impact on Adjusted Net Rating from lineup data retrieved from NBA Stats. For team recommendations, we identified the weakest link (using the RAPTOR metric) of the most used lineup of 6 NBA teams and created a short list using a Comparison Index calculated with salary and value added differences.

This project was inspired by previous work from the MIT Sloan Sports Analytics Conference and Medium.com. However, with the addition of the Comparison Index metric and replacement analyses, we evolved the study further for team-building applications. With this project, teams and fans can identify and consider player types and modern-day positions in the NBA. It also provides insight for teams to identify players based on their relative value, shedding light on the many undervalued players around the league. These players often go unnoticed because of the market they play in, or their lack of opportunity in their current team. Lastly, the project's output on clusters will help up-and-coming players identify valuable skills and tendencies in the modern game and model their development towards those.

Today's NBA

The National Basketball Association is one of the epicenters of sports analytics. Basketball has evolved through its history, because of the players, the financial implications and the popularity of the sport. Lately, however, the game has partly evolved with the help of analytics.

The NBA collects a great deal of data thanks to which analysis can be ran. The latest changes have been interesting in such a way that traditional lineups are no longer in the books and traditional positions (point guard, shooting guard, power forward, center) do no longer define players. Players with the same listed position evolve differently on the court and throughout their career. The skillset and characteristics of players is what really define their playing style rather than the position listed on the NBA website.

The Golden State death lineup from 2014 to 2019 that dominated the league for multiple consecutive seasons was everything but a traditional lineup. Klay Thompson, Stephen Curry, Kevin Durant, Andre Iguodala and Draymond Green have several listed positions, and some overlap each other. Stephen Curry was a point guard, Klay Thompson was considered a shooting guard, and all three of Andre Iguodala, Draymond Green, and Kevin Durant were traditionally listed as small forwards. Yet, the role of these players and their contribution to the team is different. The Golden State death lineup is the origin of a trend. We have witnessed situations where two players at the same position adapt their playing style to be effective together, we saw the emergence of new roles and players evolving throughout their career.

Problem Faced and Intended Goal of the Project

With this, creating an effective lineup becomes a complicated task for NBA front offices where they no longer try to put the best player at each given position. Teams have rather been emphasizing chemistry between players according to their skills and characteristics. A player can contribute differently to a team than what his listed position connotes. This project intends to understand what the new positions and roles in the NBA are and to be able to assign players in these new roles. This project can also be used to help teams assess player performances, improve team building, create optimal lineups and replace the weakest link of each by the best possible player with respect to rating and budget constraints.

Questions Answered

- What new roles are there in the NBA, and how do they compare to the traditional listed positions?
- Which players compose these siloed roles, and what statistics characterize these clusters?
- Which roles are most impactful on a successful team and lineup? What combinations of these roles are most often successful in today's NBA?
- Are there ways to replace a lineup's weakest link - how do we quantify that?

Limitations of the Project

- Injured players and free agents, due to missing data, were not included in lineup and replacement analyses. Ideally, free agents would be top targets for teams as they do not have to give up value in players to sign these players. However, we could not value these players based on a projected salary.
- In addition, we were not able to fully create an in-depth replacement parameter, which could contain many factors that teams may look for such as: Impact-Salary relationship, Point of Contract.
- Defensive statistics and tendencies were not as utilized in clustering as offensive statistics and tendencies because of the availability of data for the former. Therefore, most of the clusters were based on offensive production and tendencies.
- The time frame of the project limited the amount of time and effort placed in parsing and expanding available data for compatibility. Many manual changes had to be made in matching player codes to join tables on, especially since they come from differing sources.
- The capability of our machines for analysis limited some aspects of the machine learning techniques that we employed, as well as the size of data we were able to process.

Data Collection and Pre-Processing

We were able to manually scrape stats from Basketball-Reference, including players that have played in the last 10 seasons (2010-2011 to 2019-2020). We also included their prior stats to track their development and "evolution" throughout their careers. We also excluded seasons where players played less than 30 games to provide a substantial sample of their performance. In the current season, Stephen Curry suffered a hand injury that led him to play all but 5 games. Thus, the season was not included in the final dataset, among others. This led us to a dataset with 6,724 rows and 75 columns, each being a specific player's season played, arranged alphabetically and by season. For example, row 1 in the dataset is Alex Abrines' rookie season, 2016-17.

Variable Selection

In order to cluster players based on their production, efficiency, and tendencies (essentially, their playing style), we chose specific stats that allowed us to capture playing styles and maximize separation between each. This led us to use 24 variables, outlined in Figure 1.

We believe that each of these 24 variables captures what a player does and is given the opportunity to do on the court, given their strengths and styles. Each stat used, to standardize and scale respective to player opportunity and playing time, was projected as a rate, with the exception of height. Height was also used to distinguish between different players, as height most often dictates their role and ability on the floor. These statistics will be used in dimension reduction and, eventually, in understanding our player clusters.

Variable	Description
Height	Player height (in inches)
Points per 100 Poss.	Points scored per 100 offensive possessions
Field Goal Attempts per 100 Poss.	Number of field goals attempted per 100 offensive possessions
Offensive Rebound Rate	% of available offensive rebounds a player gets during playing time
Defensive Rebound Rate	% of available defensive rebounds a player gets during playing time
Assist Rate	% of teammate FGs a player assisted during playing time
Steal Rate	% of opponent possessions ended with a player's steal during playing time
Block Rate	% of opponent FGAs blocked by a player during playing time
Turnover Rate	Number of turnovers committed per 100 offensive possessions
Player Efficiency Rating	Standardized per-minute production rate
Usage Rate	% of offensive possessions used by a player during playing time
Free Throw Rate	Number of free throws per field goals attempted
Free Throw Percentage	Number of free throws made per free throw attempted
2P Field Goal Percentage	Number of 2-point field goals made per attempt
2P Field Goal Assisted Rate	% of 2-point field goals that are assisted
3P Field Goal Percentage	Number of 3-point field goals made per attempt
3P Field Goal Assisted Rate	% of 3-point field goals that are assisted
Corner 3P Field Goal Attempt Rate	% of all 3P field goal attempts from the corner spots
Dunk Attempt Rate	% of all field goal attempts that are dunks
0-3 feet Field Goal Attempt Rate	% of all field goal attempts from 0-3 feet
3-10 feet Field Goal Attempt Rate	% of all field goal attempts from 3-10 feet
10-16 feet Field Goal Attempt Rate	% of all field goal attempts from 10-16 feet
16 feet-3Pt Field Goal Attempt Rate	% of all field goal attempts from 16 feet to before the three-point line
3P Field Goal Attempt Rate	% of all field goal attempts from beyond the 3P line

Dimension Reduction

To effectively cluster our players, we decided to reduce our 24 variables to a lower dimensional space. In this case, we used Linear Discriminant Analysis (LDA). LDA uses supervised learning (with result groups being the traditional positions) to reduce variables while maximizing retained information. In this case, each variable was scaled before applying reduction. After applying LDA with the MASS package on R, we ended up with 4 distinct dimensions and 75% accuracy when applied to the full dataset.

Clustering Methods Used

Once dimensions were reduced and separation was maximized, we decided to test the k-means clustering technique as our potential approach to clustering players. K-means clustering uses the elbow method to recommend an ideal number of clusters and every data point is allocated closest to the mean of each determined cluster thereby reducing intra-cluster variation. However, from the silhouette score and the number of clusters recommended (k = 2 or 3), we decided that the k-means clustering method was not satisfactory and moved on to another clustering technique, model-based clustering.

Model-based clustering is another technique that attempts to fit the given data into clusters retroactively with an expectation-maximization algorithm and the Bayesian Information Criterion (BIC) score, which is likelihood-based. This allowed us to see which models and cluster numbers best maximized separation between clusters. The technique, which uses the mclust package on R, allows us to compare different models and their respective performance metrics. Using model-based clustering, we ended up with 7 different clusters. Unlike k-means clustering, model-based clustering also provided us with z-scores to represent the likelihood of a certain data point (in this case, a player's specific season) to be part of a certain cluster. This information will be used to (1) identify players with near-even likelihood of being in multiple clusters, and (2) create probability-based "soft" lineups to aid in our ideal lineup analysis.

Clustering Results

While former projects have identified eight or nine clusters, our model-based clustering results yielded seven clusters in which players can be categorized: [1] three-point shooting swingman, [2] stretch forward, [3] rim-running athletic big, [4] high usage guard, [5] traditional big, [6] versatile role player, and [7] ball dominant playmaker. Their top and bottom 4 stats are visualized in Appendix A.

The three-point shooting swingman (Cluster 1), is a catch and shoot three player. His role does not involve creating or playmaking but rather finding an open spot to shoot behind the three-point line. His three points statistics are therefore stunning, and his defensive statistics are above average. Klay Thompson, Joe Ingles, and Doug McDermott in several seasons have perfectly demonstrated how three-point shooting swingmen play even though they do not have the same listed position. Klay Thompson is listed as a shooting guard; Joe Ingles is listed as shooting guard and as a small forward; and Doug McDermott is listed as both a small and power forward. Regardless of positions, they have similar playing style and their main contribution in terms of scoring comes from three pointers.

The stretch forward (Cluster 2), plays both forward positions as his role is to stretch the opponent's defense. Shooting threes is also an important part of his playing style. In terms of physique, the stretch forward is, on average, taller than the three-point shooting swingman, which makes him a better rebounder. His usage rate and defensive statistics, however, are lower than average. Rashard Lewis in 2012-2013, Nikola Mirotic in 2017-2018 and Markieff Morris in 2014-2015 embodied the stretch forward role in those seasons and for much of their careers.

Rim running athletic bigs (Cluster 3) are fast and towering figures, which gives them high rebounding statistics. What makes them different from traditional centers are their offensive abilities. They are decent ball handlers and are able to drive the ball. Their usage rate is higher than most big players in the league. Elite players from this cluster are also able to distribute the ball and create assists. Bam Adebayo, Giannis Antetokounmpo and Anthony Davis best represent this role. Although their physiques are like traditional bigs, they are well-rounded and able to make a difference on their own offensively.

Clustering Results

The fourth category, high usage guards, are elite creators. They have good ball-handling and shooting skills. They distribute the ball and organize the team on the floor. They have high assist rates, field goal percentages, and free throw percentages. In terms of size, they are shorter than most players in the league. Their defensive and rebounding stats are consequently low. This position is mostly filled by point guards such as Stephen Curry, Damian Lillard and Deron Williams.

Traditional bigs (Cluster 5), are much taller than league average. Their playing styles are mostly defined by their size. They mainly defend and rebound and, therefore, are primarily positioned inside. Their perimeter scoring statistics are close to inexistent and their usage rate is low. Steven Adams, Roy Hibbert and DeAndre Jordan are known traditional bigs. Compared to rim-running athletic bigs, their offensive usage and tendencies are low.

The versatile role player (Cluster 6), is a well-rounded player. He has decent statistics in many domains but does not excel or is far below in any. As a group, versatile role players have wide ranges in most statistics and cannot be defined by them. A common trait between all of them is their almost nonexistent three-point statistics. Shaun Livingston (in his years in Golden State), Damian Jones and Montrezl Harrel (in his rookie year with the Rockets) are some players who represent this category. All three players have different listed positions which shows the variety of versatile role players in the league both in their role and in their physique.

Finally, we have the ball dominant playmaker (Cluster 7). He is a well-rounded player overall. He distributes and handles the ball well with a good assist rate. He is an above-average shooter and his size (taller than average) makes him a good inside player as well. Compared with high-usage guards, he has both good offensive and defensive ratings, and has a similarly high usage rate. He is usually unassisted when he scores as he mainly scores off the dribble. LeBron James, James Harden and Ben Simmons are great examples of ball dominant playmakers. Although their listed position goes from point guard to power or small forward, they contribute to their team in similar ways.

Leader Clusters per Stat

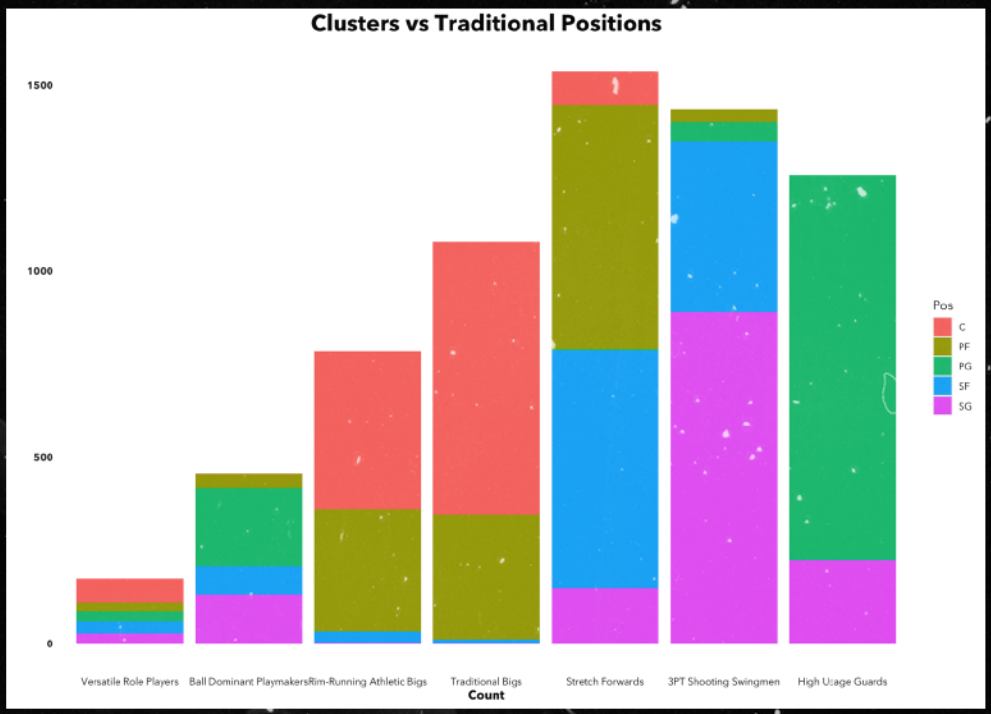
In terms of leader cluster per stat used, the results are in line with players' playing style, characteristics, tendencies, and overall skills. One main takeaway from this analysis is that versatile role players are not leaders in any of the statistics.

Statistics	Leader Role
Height	Traditional Bigs
PER	Rim-Running Athletic Bigs
DRB Rate	Rim-Running Athletic Bigs
ORB Rate	Traditional Bigs
AST Rate	High Usage Guards
STL Rate	High Usage Guards
BLK Rate	Traditional Bigs
Point per 100	Ball Dominant Playmakers
Usage Rate	Ball Dominant Playmakers
Free Throw Rate	Traditional Bigs
Free Throw %	High Usage Guards

Statistics	Leader Role
FGA per 100	Ball Dominant Playmakers
2PT FG%	Traditional Bigs
2PT FG %assisted	Traditional Bigs
3PT FG%	Three-Point Shooting Swingmen
3PT FG %assisted	Stretch Forwards
Corner 3 Attempt Rate	Stretch Forwards
Dunk Attempt Rate	Traditional Bigs
%FGA 0-16ft	Traditional Bigs
%FGA 16ft-3P	High Usage Guards
%FGA 3P	Three-Point Shooting Swingmen

Distribution of Clusters and Traditional Players

Due to the versatility of certain players that have been discussed previously, as we tried to realize our project potential which shall provide a fresh perspective of traditional positions and our clusters, we also counted on this visualization to leave us a better idea of how traditional positions may fit in different clusters. In other words, when we analyzed players, instead of utilizing traditional positions, we categorized them into clusters that measure their actual functionality on the team.



The biggest players on the team are supposed to dominate the rim. Accordingly, most centers are categorized as Traditional Bigs or Rim-Running Athletic Bigs, while a few are Stretch Forwards or Versatile Role Players. This can be an illustration of the fact that the game of basketball is gradually switching from its original, physical-oriented game, to a fast-paced game that requires more agility even for the big guys. Today, Kevin Love is a great example of Centers with such flexibility.

Players with a certain physique who play offensively with their backs toward the basket and who play defensively under the rim are Power Forwards. They, however, sometimes are used interchangeably with the Centers or even as shooters. Based on our classification model, they are mostly Stretch Forwards, Traditional Bigs, or Rim-Running Athletic Bigs.

Smaller than Power Forwards and Centers with enhanced agility and shooting skills, while bigger than the Guards, Small Forwards are considered the most versatile positions, as well proven in our model that Small Forwards, instead of High Usage Guards, are in every other cluster. LeBron James, who takes the ball, dominates the rim, facilitates the offensive line, shoots deep threes, guards the Centers and the Guards, is traditionally considered as a Small Forward.

Exceptional shooting skills and comparatively weaker body type properly describe Shooting Guards. This description is perfectly reflected in our model since they are mostly 3PT Shooting Swingmen or a few smaller positions. Legendary Shooting Guard Ray Allen's extraordinary shooting skill, to some, saved the career of a notable legend by making a corner buzzer-beater game-winning three against San Antonio Spurs in 2014.

Backcourt players who mostly facilitate the team's offensive plays are Point Guards. These players stand shorter than the rest of the lineup, have great ball control, and implement a greater vision of the game as a whole. In our model, they are High Usage Guards or Ball Dominant Playmakers. Russell Westbrook, who dominates the ball almost half of the game and with perfectly-rounded stats, is the best reflection of such descriptions.

Player Development

Today, players may switch roles in one single game. The utilization of clusters plays an important role in defining the league. With that being said, it is not uncommon for a superstar to utilize different functionalities throughout their career. Before joining the Miami Heat in 2010, Chris Bosh played predominantly as a Rim-Running Athletic Big at Toronto Raptors that in need of absolute height and rim-attacking, pick-and-roll plays. However, at Miami, LeBron James destroyed the rim every night with a very high level of field-goal performance; the original role of Chris Bosh then seemed duplicated. Bosh, instead, transformed himself and made certain game winning threes by running in and out of the three point line as a Stretch Forward. Blake Griffin, will always be remembered as the one who brings damage to the rim, as he has appeared in almost every compilation of best dunks. Some violent, in-your-face dunks infer his superior athleticism as a Rim-Running Athletic Big. Later on, as his athleticism faded, he developed a perimeter shot, becoming a Stretch Forward. What happened next - signing with the Pistons - really promoted him to the core of the team, and he then has become a Ball Dominant Playmaker. All of the classifications based on our model are proved by these stories that made these players stars as we see today, leading to another fact that understanding the actual utilization, or clusters, of players on a lineup, other than understanding traditional positions, is imperative in today's basketball philosophy.

Likelihood Mapping

From the model-based clustering method, we were able to retrieve z-scores that represented each player's per season probability of being part of a certain cluster. For example, Danilo Gallinari of the Oklahoma City Thunder (in the 2019-20 season) had a z-score of 0.339 for Three-Point Shooting Swingmen, and a z-score of 0.639 for Stretch Forward (with less than 0.01 for other clusters). This means, more or less, that he had a 34% likelihood of being a Three-Point Shooting Swingman and a 64% likelihood of being a Stretch Forward given the stats that we applied in the clustering technique. Below is a sample of 3 players in the current season and their respective z-scores for each cluster.

Player	TPSS	SFOR	RRAB	HUG	TDB	VRP	BDP
Kawhi Leonard	0.162	0.103	0.001	0	0	0.031	0.703
Jamal Murray	0.27	0	0	0.198	0	0.006	0.526
Daniel Theis	0	0	0.552	0	0.438	0.009	0

This allowed us to dig deeper and identify certain players that had close likelihoods of being part of certain clusters. Majority of players map very strongly to one single cluster, while some map nearly evenly with two or more clusters.

Players Mapping Closely With 2 Clusters

Jae Crowder: He can stretch the defense by making three pointers, like a shooting swingman, and guard multiple positions on the other end like a stretch forward. He has better defense, but stands shorter than most stretch forwards.

Andre Drummond: What makes this big guy valuable is his flexibility and some agility that makes him a player who fits in both clusters of the big men: he can be either a Rim-Running Athletic Big or a Traditional Big.

Mike Bibby: As a current coach of a professional basketball team, it is not hard to imagine his vision, rather than an exceptional shooting skill, of the game as a whole when he played at Sacramento Kings where he could be either a Ball Dominant Playmaker or a High Usage Guard.

Data Collection and Pre-Processing

After clustering and identifying player roles, we wanted to move on and understand which lineup compositions and cluster combinations were most effective in achieving success on the court. In order to effectively evaluate lineup effectiveness with player cluster composition, we collected data from the most recent season (2019-20) from stats.NBA.com. We manually scraped and parsed through each lineup to retrieve each player part of that specific lineup, their games, minutes, and possessions played, and their net rating (among other statistics not used in this analysis). We decided just to use the most recent season's lineup data because of compatibility issues with our player data set. Given more time, it is ideal to include lineup data from the last 10 seasons as well. Using the player likelihood ratings from our clustering model, we were able to join the probabilities of each player within a lineup and aggregate each lineup's composition within each cluster. For example, the 2019-20 Boston Celtics' most used lineup composed of Kemba Walker, Jaylen Brown, Gordon Hayward, Jayson Tatum, and Daniel Theis. The group played in 17 games for a total of 188 minutes, 396 possessions, and a 10.71 Net Rating. Given each player's likelihood of mapping into each cluster, we got a "soft" aggregate lineup probability composition (rounded up to the nearest .25/.50/.75/.00 for easier analysis), which you can see in Appendix B.

From the table, we can see that this Celtics lineup is composed of, on aggregate, one Three-Point Shooting Swingman, one and ¼ Stretch Forward, 0.500 Rim-Running Athletic Big, 0.500 High Usage Guard, 0.500 Traditional Big, 0.000 Versatile Role Player, and 1.250 Ball Dominant Playmaker. This can be interpreted as having approximately two shooters to stretch the floor, 1 athletic big man to roam the inside, and 2 high usage players to facilitate the offense. This lineup's net rating is 10.71, which is a great rating, considering that this rating is the point differential per 100 possessions. This is an example of an effective lineup.

Linear Regression Model

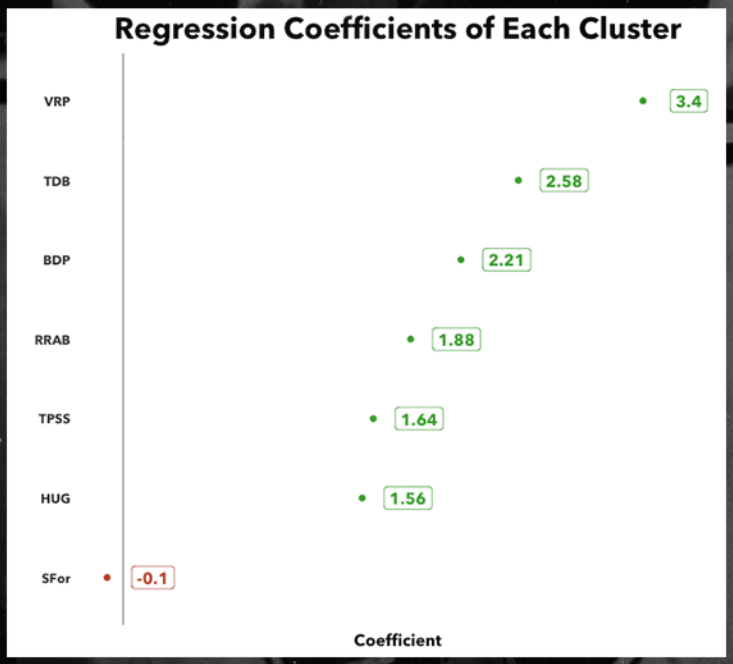
To see which specific clusters had the highest individual impact on performance, we decided to first perform a linear regression analysis. The dependent variable that we used was Adjusted Net Rating, which we computed in the following way, based on a computation that was determined by Samuel Kalman and Jonathan Bosch, the writers of the MIT Sloan Sports Analytics Conference study.

If Possessions >= 550, Adj. Net Rating = Net Rating

If Possessions < 550, Adj. Net Rating =

$$\left(\frac{\text{Possessions}}{550} * \text{Net Rating}\right) + \left(1 - \frac{\text{Possessions}}{550}\right) * \text{Team Net Rating}$$

The independent variables we used were the aggregated player cluster probabilities for each lineup. We got the following results:



In this case, the Versatile Role Player actually had the highest coefficient, and the Stretch Forward had the lowest. However, this does not show us the interaction between each cluster within a five-player lineup. Obviously, having five versatile role players on the floor the entire game does not give you a Net Rating of 17.0. We then decided to tackle this problem using a more in-depth technique, the random forest model.

Random Forest Model and Predictions

The Random Forest regression model creates decision trees used to predict a continuous response. The decision trees that make up the forest are created by "bagging" methods to generate resampled versions of the given dataset, fitting a decision tree at each set. Ensembled trees become more independent, which leads to better predictions in different situations. The algorithm takes the average amongst the different decision trees in the forest to come up with an aggregate prediction for that lineup. This model can give us a better view and understanding of interaction between clusters. We used our existing lineup data to train the model and used it to create a prediction for every possible lineup in the NBA, which was made as a matrix similar to the figure shown below. Since the creation of a database in 0.25 intervals was ideal but could not be done given the processing capabilities of our computers, we decided to use 0.33 intervals.

Clust. 1	2	3	4	5	6	7	Pred
5	0	0	0	0	0	0	
4.67	0.33	0	0	0	0	0	
...	...	...	...	...	...	...	
0	0	0	0	0	0	5	

Random Forest Results

As a result of our random forest model predictions, we got the following combination from the comparisons of frequency of these players with Adjusted Net Rating. First, we compared high usage guards and ball dominant playmakers and found that they usually work better with just 1 of either cluster in the lineup as high usage players. Next, we compared those high usage players with three point shooting swingmen and stretch forwards and found that three point shooting swingmen were more valuable than stretch forwards, due to their increased defensive capabilities, and that 1 or 2 of these players were optimal for a lineup to space the floor for the high usage player. Lastly, we found that rim-running big men were slightly more valuable alone than traditional big men if having more than 1 type of big man on the floor. For rim-running big men, which means having two-way skilled big men on the floor, having two patrolling the floor and setting screens allows greater movement and spacing for the ball handler and shooters. This is mostly consistent with many successful lineups today. One example is the most used lineup of the Los Angeles Lakers (12.6 Adj. Net Rating), composed of Avery Bradley, Danny Green, LeBron James, Anthony Davis, and Javale McGee. James acts as the primary ball handler in these situations, with Bradley and Green spacing the floor and Davis and McGee roaming inside to set screens or finish at the rim. Additionally, having Davis (an elite big with the ability to finish off the dribble, shoot off the catch, or distribute on all three levels) on the floor increases the strength of this lineup. Details of the comparisons are in Appendix C.

Weakest Link Analysis

In order to better illustrate the utilization of our clustering model, we chose six teams based on their actual performance in this season to do a further analysis: recommending optimal lineups for managers when they consider replacing certain players on their teams. These teams are selected based on current ranking:

MIL 53-12 | LAL 49-14 | BKN 30-34 | ORL 30-35 | CLE 19-46 | GSW 15-50

Based on our model, we were able to choose the player with the weakest link to the team's most used lineup. But before showing the results, we would like to discuss the manipulation process.

Weakest Link Analysis

After selecting the teams, the process starts from selecting the lineup that we are going to analyze. Since starting lineups may change throughout the season, we chose the most used lineup for each team (i.e. lineup with the most possessions). The original lineup for each team is as follows:

BKN	CLE	GSW	LAL	MIL	ORL
Garrett Temple	Kevin Love	Draymond Green	LeBron James	Brook Lopez	Nikola Vucevic
Spencer Dinwiddie	Tristan Thompson	Glenn Robinson	JaVale McGee	Wesley Matthews	Evan Fournier
Joe Harris	Cedi Osman	D'Angelo Russell	Danny Green	Eric Bledsoe	Aaron Gordon
Taurean Waller-Prince	Collin Sexton	Willie Cauley-Stein	Avery Bradley	Khris Middleton	Markelle Fultz
Jarrett Allen	Darius Garland	Damion Lee	Anthony Davis	Giannis Antetokounmpo	Jonathan Isaac

As having mentioned, there is a column in our dataset that measures the effectiveness of having a specific player on the lineup. "RAPTOR" measures the linkage between the player himself and the others on the same team. The weakest link on a lineup, then, is the player with the lowest RAPTOR Score. After obtaining the most used lineups, we created shortlists for each player with the weakest link on each lineup by comparing our articulated "Comparison Index". When considering the determinants in the comparison index, we took several calculations of different variables, and eventually we decided to use the commensurate addition of the difference of salary, and the difference of raptor scores between the replacing target and the options available. Furthermore, we considered options as available based on the following criteria. The replacing players must: (1) have a lower salary than the original player, (2) have a higher RAPTOR score than the original player, and (3) be in the same cluster as the original player.

$$Comparison\ Index = \left|\frac{Salary_{new} - Salary_{orig}}{Salary_{new}}\right| + \left|\frac{Raptor_{new} - Raptor_{orig}}{Raptor_{new}}\right|$$

With this comparison index, we were able to create lists of replacing targets for each player on the original lineup. For example, for Avery Bradley at the Los Angeles Lakers, we were able to obtain multiple options, where Terence Davis would be an optimal choice as a replacement with a 2.00 index, since he has a significantly higher RAPTOR rating at 2.6 (compared to Bradley at -1.0) and a rookie salary of $898,310 as an undrafted rookie (compared to Bradley at $8.8M).

The bigger the index is, the better the option is. Indices greater than or equal to 2 are considered high-value replacements, with those between 2 and 1 being moderate-value replacements, and those below 1 being low-value replacements. To illustrate how this meets our expectation of accuracy, we were interested in seeing if there's anyone who may replace LeBron James practically. In other words, who is cheaper but had a greater influence on the team. Then, we got Kawhi Leonard, which was not overwhelmingly surprising. Technically, Leonard receives a lower payroll when having a better RAPTOR score.

Team Recommendations

We would make recommendations based on our manipulation: replacing the players with the weakest link to their team. So, the six players we are going to replace are: Taurean Waller-Prince (BKN), Darius Garland (CLE), Glenn Robinson (GSW), Avery Bradley (LAL), Wesley Matthews (MIL), and Markelle Fultz (ORL).

BKN	CLE	GSW	LAL	MIL	ORL
Garrett Temple	Kevin Love	Draymond Green	LeBron James	Brook Lopez	Nikola Vucevic
Spencer Dinwiddie	Tristan Thompson	Terence Davis	JaVale McGee	Terence Davis	Evan Fournier
Joe Harris	Cedi Osman	D'Angelo Russell	Danny Green	Eric Bledsoe	Aaron Gordon
Derrick Jones, Jr.	Collin Sexton	Willie Cauley-Stein	Terence Davis	Khris Middleton	Jalen Brunson
Jarrett Allen	Jalen Brunson	Damion Lee	Anthony Davis	Giannis Antetokounmpo	Jonathan Isaac

Team Recommendations

The new lineups, after replacing players with the weakest raptor score by the options with the highest comparison index is shown above. Of course, there are more than one options for each player, so that we put the top three options for these players in Appendix D.

As readers may have noticed, Terence Davis appeared in three of the four extremes: once in the worst team, and twice in the top two teams. One conclusion we may draw is that teams are always in pursuit of brilliant shooters and guards, since in our model, Terence Davis is categorized as a 3PT Shooting Swingman. This also puts Davis in the spotlight as a very valuable rookie, despite playing limited minutes on the Toronto Raptors. The fact that in the past several years, championship teams were composed of two stars, a great defender, and exceptional shooters is also redeemed by our model. The game is changing, so is the demand. However, better teams, in terms of net rating and current standing, are always in less need of replacements as proven by a lower number of replacing options available from our model.

Conclusion

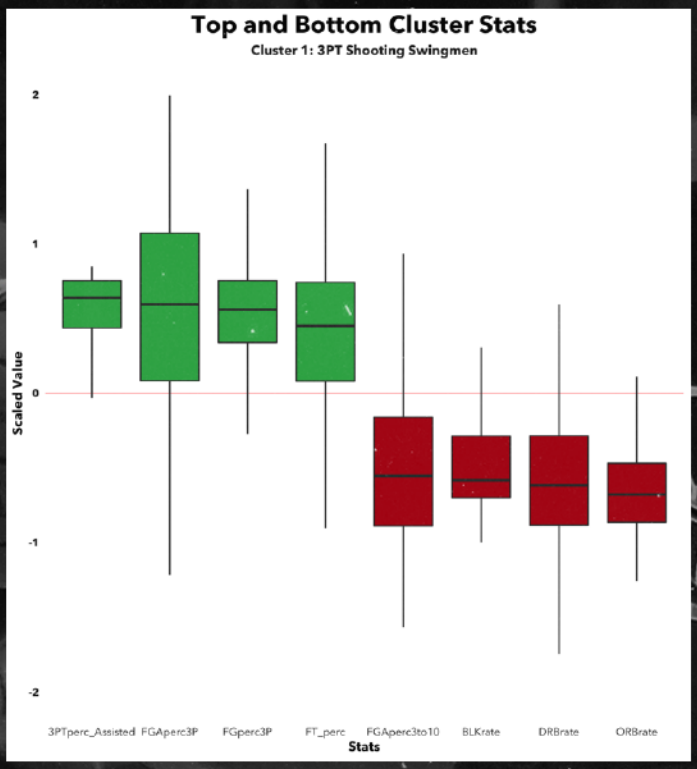
At the end of the project, we found that there are ways to statistically cluster NBA players based on their production, efficiency, and tendencies. We were able to determine modern-day player "positions" and more accurately describe different playing styles. Of course, we acknowledge that there is no 100%-foolproof way to place these players in silos, since players from each cluster aren't fully confined to the strengths of their clusters. Moving forward, we also foresee the NBA game changing once again, and this will lead to different clusters of players forming and lineup efficiencies changing.

We were also able to get a general sense of an effective lineup composition, as teams that performed substantially better had primary players from 3 clusters: (1) an elite high usage player (either a high usage guard or a ball dominant playmaker), (2) a skilled athletic big man, and (3) three-point shooting swingmen who double as 3-and-D players. NBA front offices and other basketball leagues can use this method and data to identify opportunities for roster construction and rebuilding, as well as identifying optimal lineups and key matchups. This would also be a good opportunity for teams to identify low-interest players and capitalize on their understated value, much like what Billy Beane did in Moneyball. Conversely, it sheds light on often under-valued players who go under the radar in smaller markets, or have less playing time on various teams.

If this study were to be taken further, the comparison and replacement analysis could even be fortified with a formalized statistical optimization as a foundation for targeting players - with more constraints and value identifiers. This could also be done in context of contracts, with long-term contracts harder to trade and injured/aging stars seen as trade risks. Different combinations could also be analyzed, for key player combinations. Seeing as how the game has transitioned within the decade from three-player superteams to this year's plethora of competitive duos, this may also be a point of interest for general managers. Also, with increased computing power, we can appropriately predict lineups with more yearly data and expanded permutations. Lastly, for players on the up-and-coming as high school, college, and international recruits, they are able to identify which skills and tendencies are valuable in today's game and model their development towards these playing styles.

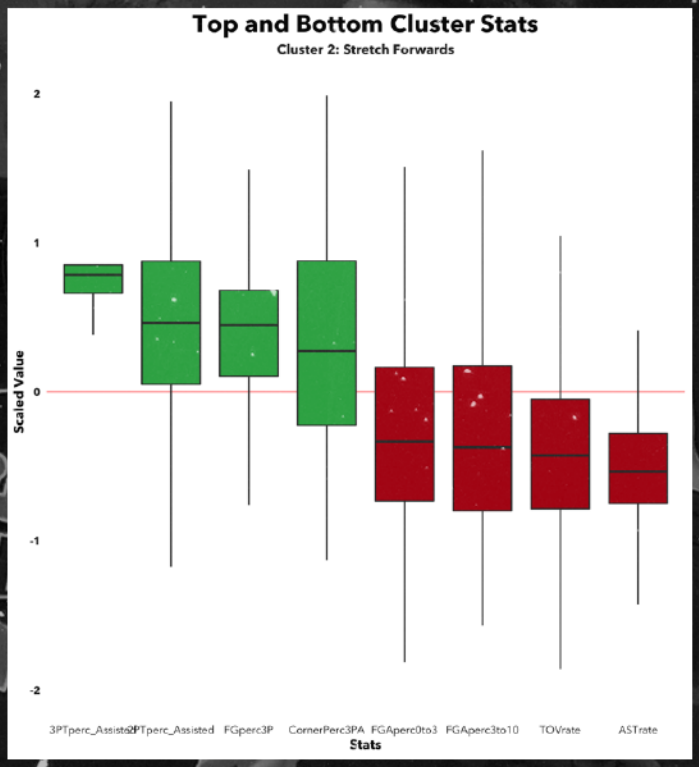
This research project was able to show the power of machine learning in adapting player role identification and team management to the modern era of the NBA. Hopefully, this project can also be applied to different leagues (such as lower-level leagues and university athletics leagues) and develop the understanding of team and lineup building in those contexts.

APPENDIX A - Top and Bottom Stats per Cluster



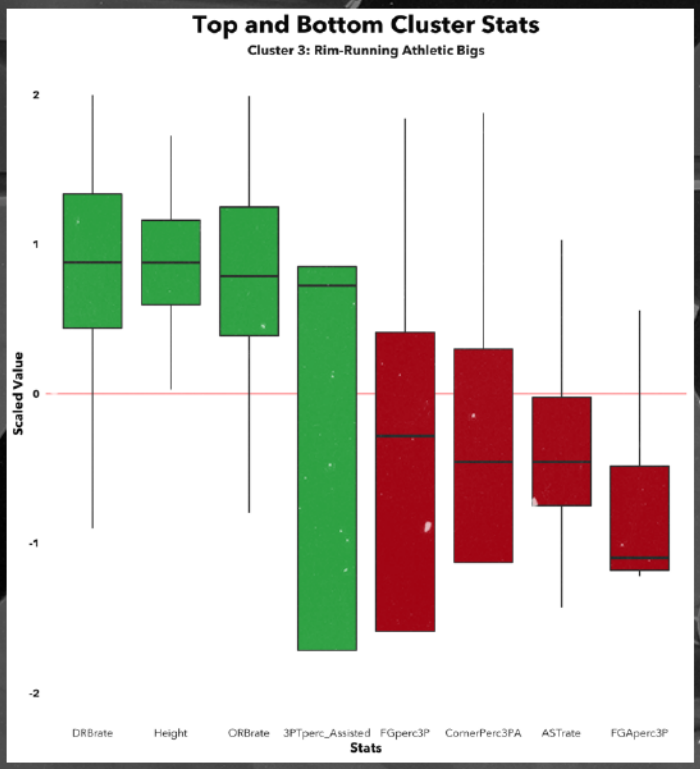
Notable Top Stats: 3PT Assisted, 3PT FG Attempt Rate

Notable Bottom Stats: DRB/ORB Rate, 3-10Ft FG Attempt Rate



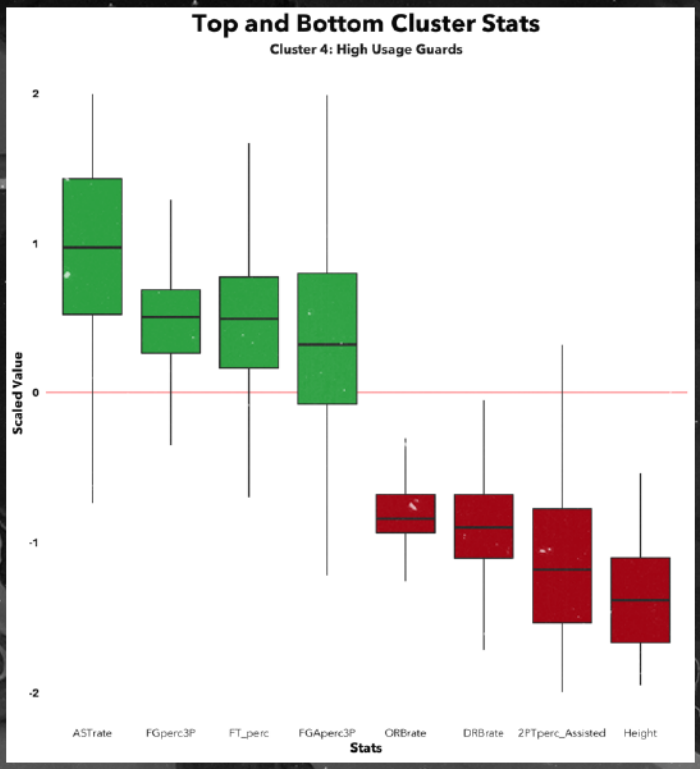
Notable Top Stats: 3PT/2PT Assisted, Corner 3 Attempt Rate

Notable Bottom Stats: Assist Rate, TOV Rate, Inside FG Attempt Rate



Notable Top Stats: DRB/ORB Rate, Height

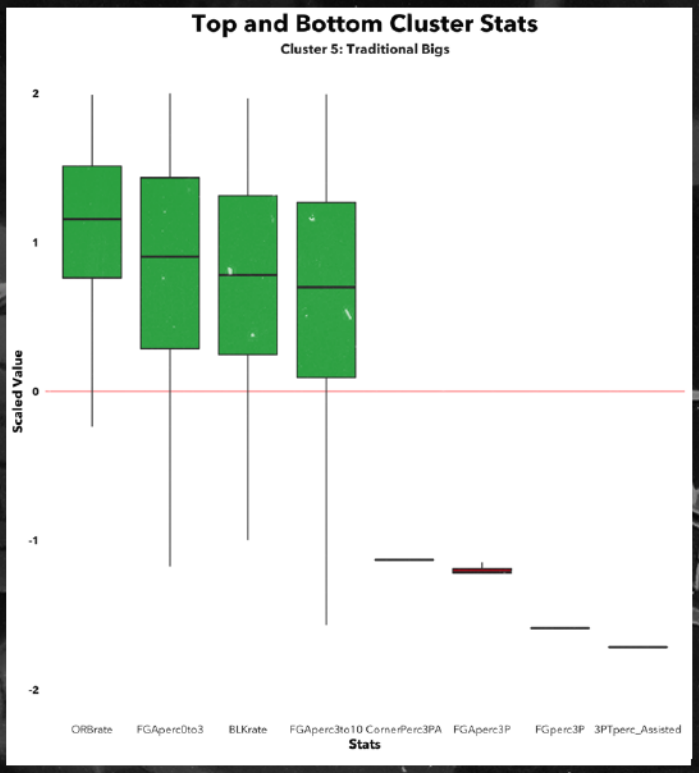
Notable Bottom Stats: 3P FG Attempt Rate



Notable Top Stats: Assist Rate, 3P FG%, FT%

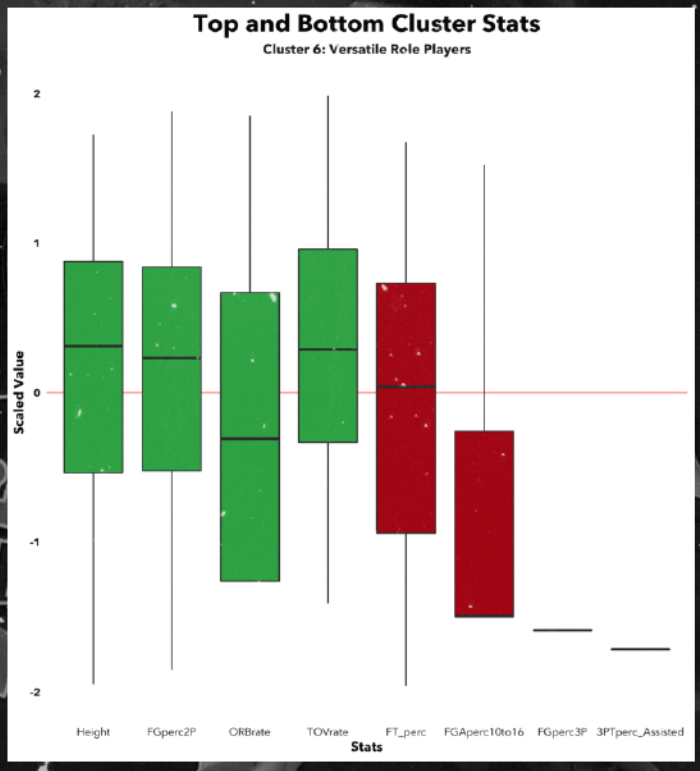
Notable Bottom Stats: Height, ORB/DRB Rate

APPENDIX A - Top and Bottom Stats per Cluster



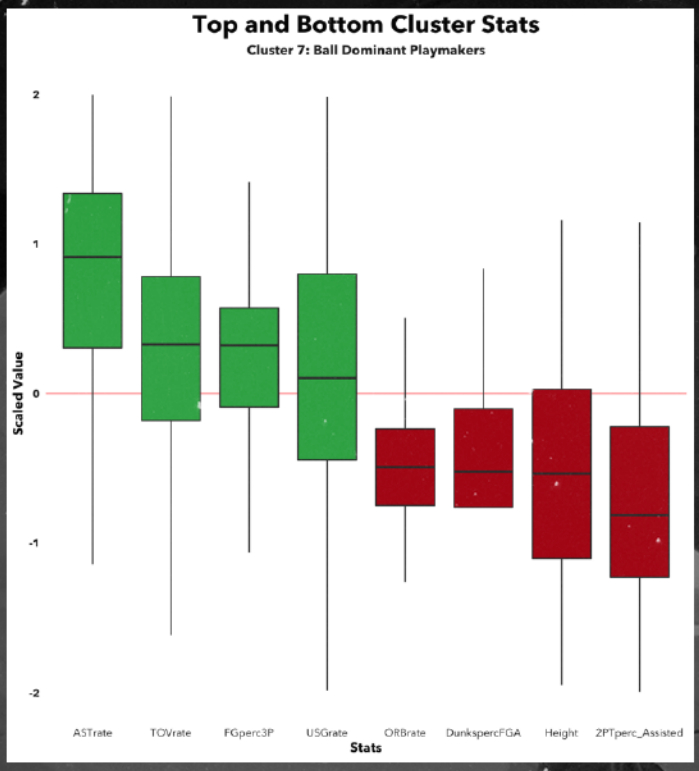
Notable Top Stats: ORB/BLK Rate, 0-10Ft FG Attempt Rate

Notable Bottom Stats: 3PT Shooting Stats



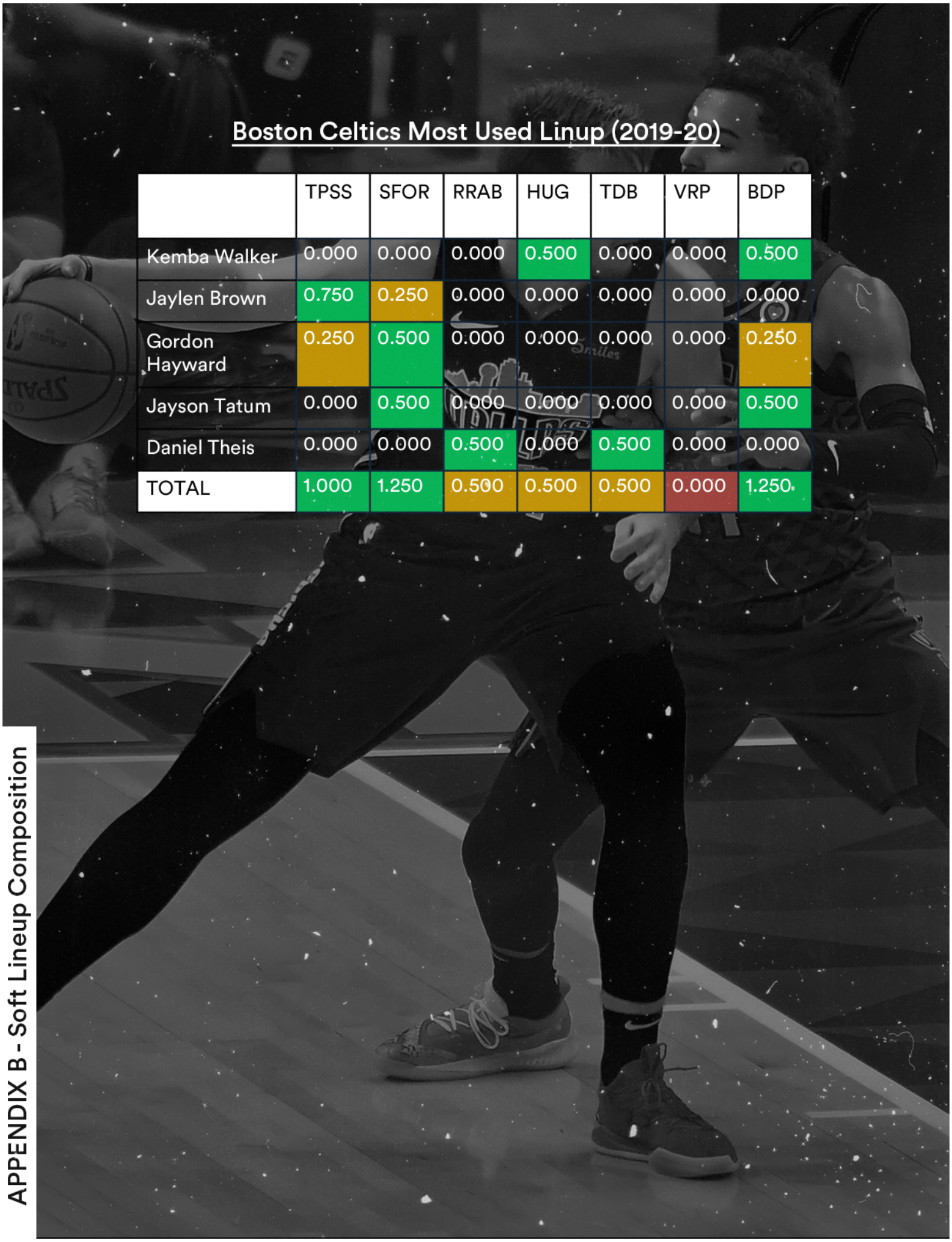
Notable Top Stats: Height, TOV Rate

Notable Bottom Stats: Shooting (3PT, Mid, FT) Stats



Notable Top Stats: AST Rate, TOV Rate, Usage Rate

Notable Bottom Stats: 2PT %Assisted, Height, ORB Rate



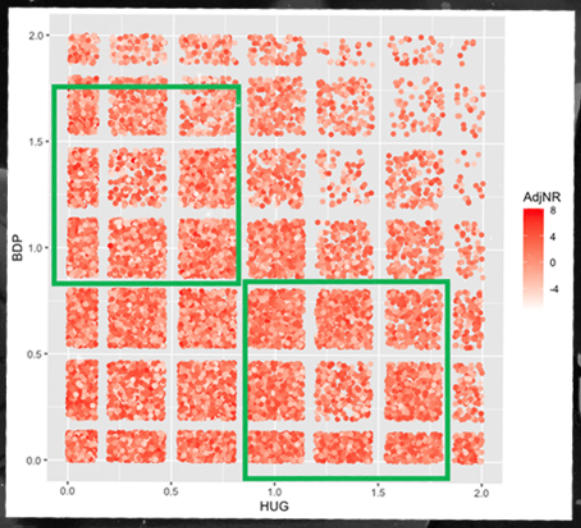
Boston Celtics Most Used Linup (2019-20)

	TPSS	SFOR	RRAB	HUG	TDB	VRP	BDP
Kemba Walker	0.000	0.000	0.000	0.500	0.000	0.000	0.500
Jaylen Brown	0.750	0.250	0.000	0.000	0.000	0.000	0.000
Gordon Hayward	0.250	0.500	0.000	0.000	0.000	0.000	0.250
Jayson Tatum	0.000	0.500	0.000	0.000	0.000	0.000	0.500
Daniel Theis	0.000	0.000	0.500	0.000	0.500	0.000	0.000
TOTAL	1.000	1.250	0.500	0.500	0.500	0.000	1.250

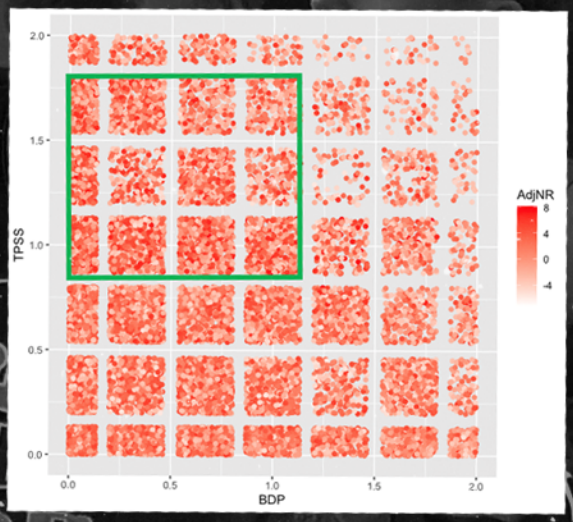
APPENDIX B - Soft Lineup Composition

Cluster Random Forest Prediction Analysis

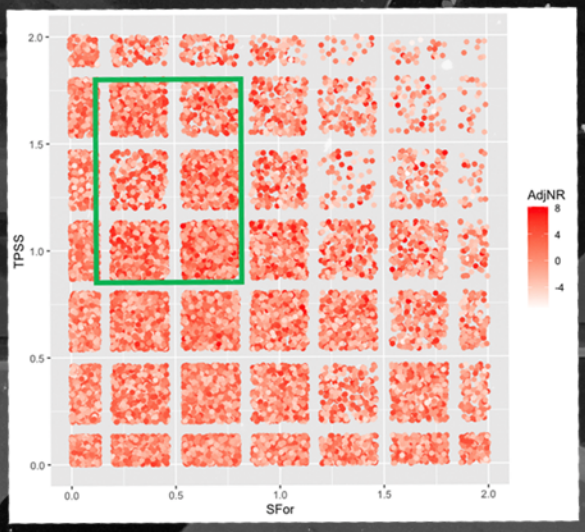
High Usage Guards and
Ball Dominant Playmakers



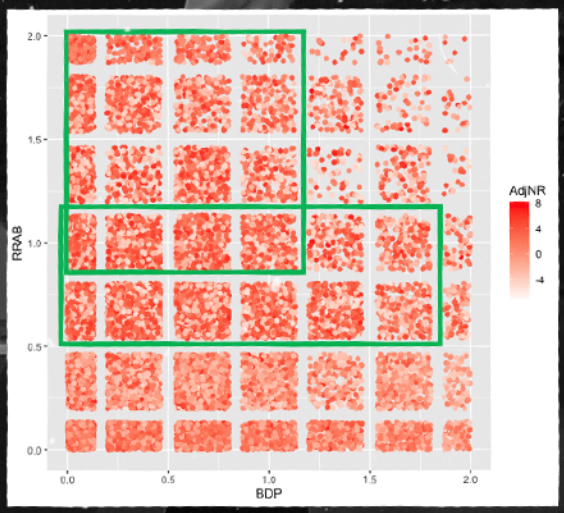
Ball Dominant Playmakers and
Three Point Shooting Swingmen



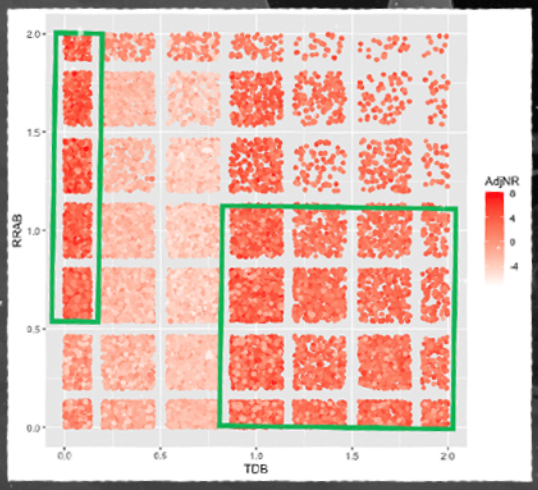
Stretch Forwards and
Three Point Shooting Swingmen



Ball Dominant Playmakers and
Rim-Running Athletic Big Men



Traditional Big and
Rim-Running Athletic Big Men



APPENDIX C - RF Prediction Comparisons

APPENDIX D - Team Recommendations

Milwaukee Bucks

Record: 53-12 | Adjusted Team Net Rating: 9.91

Most used lineup

936 Possessions
408 Minutes Played
19.1 Adjusted Net Rating

Targets

TERENCE DAVIS
Toronto Raptors
2.58 RAPTOR
$898,310
CI: 1.70

CODY MARTIN
Charlotte Hornets
2.44 RAPTOR
$1,173,310
CI: 0.46

DUNCAN ROBINSON
Miami Heat
2.73 RAPTOR
$1,416,852
CI: 0.37

Los Angeles Lakers

Record: 49-14 | Adjusted Team Net Rating: 7.57

Most used lineup

829 Possessions
389 Minutes Played
12.6 Adjusted Net Rating

Targets

TERENCE DAVIS
Toronto Raptors
2.58 RAPTOR
$898,310
CI: 2.00

CODY MARTIN
Charlotte Hornets
2.44 RAPTOR
$1,173,310
CI: 1.30

DUNCAN ROBINSON
Miami Heat
2.73 RAPTOR
$1,416,852
CI: 0.99

Brooklyn Nets

Record: 30-34 | Adjusted Team Net Rating: -1.22

Most used lineup

1032 Possessions
491 Minutes Played
-0.1 Adjusted Net Rating

Targets

DERRICK JONES, JR.
Miami Heat
2.14 RAPTOR
$1,645,357
CI: 2.12

RONDAE HOLLIS-JEFFERSON
Toronto Raptors
3.14 RAPTOR
$2,500,000
CI: 1.89

ISAAC BONGA
Washington Wizards
1.33 RAPTOR
$1,416,852
CI: 1.78

APPENDIX D - Team Recommendations

Orlando Magic

Record: 30-35 | Adjusted Team Net Rating: -1.02

Most used lineup





508 Possessions
240 Minutes Played
1.66 Adjusted Net Rating

Targets

JALEN BRUNSON
Dallas Mavericks
2.10 RAPTOR
$1,416,582
CI: 1.96

RAUL NETO
Philadelphia 76ers
1.51 RAPTOR
$1,737,145
CI: 1.63

DEVONTE GRAHAM
Charlotte Hornets
1.51 RAPTOR
$1,416,582
CI: 1.58

Cleveland Cavaliers

Record: 19-46 | Adjusted Team Net Rating: -7.73

Most used lineup





1432 Possessions
679 Minutes Played
-8.6 Adjusted Net Rating

Targets

JALEN BRUNSON
Dallas Mavericks
2.10 RAPTOR
$1,416,582
CI: 1.86

DEVONTE GRAHAM
Charlotte Hornets
1.51 RAPTOR
$1,416,582
CI: 1.63

RAUL NETO
Philadelphia 76ers
1.51 RAPTOR
$1,737,145
CI: 1.55

Golden State Warriors

Record: 15-50 | Adjusted Team Net Rating: -7.94

Most used lineup





260 Possessions
123 Minutes Played
-6.66 Adjusted Net Rating

Targets

TERENCE DAVIS
Toronto Raptors
2.58 RAPTOR
$898,310
CI: 2.49

PATRICK PATTERSON
Los Angeles Clippers
0.25 RAPTOR
$737,067
CI: 1.76

CODY MARTIN
Charlotte Hornets
2.44 RAPTOR
$1,173,310
CI: 1.70