14TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 6 - 7, 2020 | BOSTON, MA
presented by ESPN

NBA Lineup Analysis on Clustered Player Tendencies: A new approach to the positions of basketball & modeling lineup efficiency of soft lineup aggregates

Samuel Kalman, Jonathan Bosch

Track: Basketball
ID: 1548738

Abstract

Basketball has recently been considered more of a "position-less" sport. Traditionally, the five positions of basketball are point guard, shooting guard, small forward, power forward, and center. However, most NBA players have skills, styles, and preferences that cannot be defined by a single traditional position. With ten seasons (2009-2018) of NBA player stats [1] that account for a player's efficiency, opportunity, and tendencies, we were able to implement unsupervised machine learning techniques to create a framework for how NBA playing styles can be clustered on the court and used for strategic decision making when building rosters and creating lineup rotations. These player clusters can be considered new positions, and they give more accurate and detailed insight to what role a player possesses when on the court and how effective he could be in that role. Our unique methods give players a soft assignment for all clusters, that is, a probabilistic weighting onto each of the clusters indicating their likelihood of specific cluster fit. After analyzing the distribution of the various player stats within each new position, we were able to generate a player role for each cluster. Previous work [2,3] has also clustered NBA players into new positions. However, our methods incorporate different approaches of unsupervised machine learning, and offers an extension to the player clustering by additionally investigating the lineup efficiencies of different combinations of these new positions, playing on the basketball court together in collaboration.

Our work offers a more specific way for people to consider player types and positions in the NBA. It also provides insight into which combination of player types yield the most effective basketball performance. This can be beneficial to NBA front offices when acquiring and developing talent, as well as coaches when making in-game lineup decisions. Our models also contain a predictive component where we have the ability to predict the net rating of a potential lineup. As we have recently witnessed a massive change in playing style by most of the NBA (i.e. The Three Point Era), our work provides a more accurate approach for people to analyze and understand the roles, responsibilities, and combinations of specific groups of players in the NBA.

1. Introduction

According to Sports Reference LLC [1], LeBron James' position is listed as a power forward, point guard, small forward, and shooting guard. James Harden's position is listed as a point guard and shooting guard. These players, along with many others in the NBA, possess a role and set of skills that cannot be defined by a single traditional position. It is evident that the five traditional positions of basketball (point guard, shooting guard, small forward, power forward and center) are not optimal for defining the value, playing style, and fit of players in the NBA. Consider the example of Chris Paul and Derrick Rose. Both players are listed as point guards but bring drastically different skills and playing styles to their respective teams. Rose is a scorer, notorious for his driving and finishing ability near the basket. Paul, on the other hand, is a facilitating playmaker with prestigious passing ability. Simply categorizing the two players as point guards, does not give any insight into the differences in how they tend to play, and in what ways they bring value to a team. There must be a better way to categorize Paul with similar players that possess a pass first, playmaking role, and Rose with similar players that frequently drive and score near the basket.

The second part of the ambiguous position definition problem is the difficulty in modeling and predicting how certain player-position combinations will perform together. For example, a lineup of five ball dominant players at each of the traditional positions would lead to fit issues, even if the players fill out necessary size requirements at their position. Alternatively, as we have seen with LeBron James led lineups in the latter half of his career, surrounding James with complementary skills, specifically shooting, has led to a great deal of success. Using statistical techniques, such as hierarchical clustering, allows us to assess which mix of player characteristics complement each other, leading to successful lineups that perform better than the sum of their parts. The detailed information regarding the degree of ball dominance or the floor spacing ability in this example is lost by only defining a player with the five traditional positions.

Our contribution with this paper is to use unsupervised machine learning to cluster players together based on efficiency and playing style to better define the role a player possesses for an NBA team. Using these clusters, we tested two modeling techniques to predict the effectiveness of five-man lineup combinations of these newly defined positions.

We answer the following questions with our framework:

- Is there an underlying clustering distribution of NBA players based on skills and tendencies?
- Can we create positions that better describe what value a player brings to the game of basketball?
- Do significant relationships exist between combinations of these new player types?
- If so, what combinations are optimal for basketball effectiveness and success?

2. Data

2.1 Data Collection & Pre-Processing

We scraped, manually, ten years of NBA player statistics from 2009-2018. The data was collected from multiple tables from Sports reference LLC [1] and was joined together on the player's name and season. The dataset contained 5,512 observations of 73 variables. The variables were a combination of advanced statistics, per 100 possession statistics, and shot distribution statistics. Each row corresponds to one player's statistics for that season (a player can show up in our dataset ten times is he played in all ten seasons during the 2009-2018 time period). This allows us to identify the evolution of a player throughout their career. We have seen player evolution and development become much more apparent since the “analytics movement” [4] has shaped the NBA in recent years.

Figure 1: *Distribution of games played in the NBA from 2009-2018*

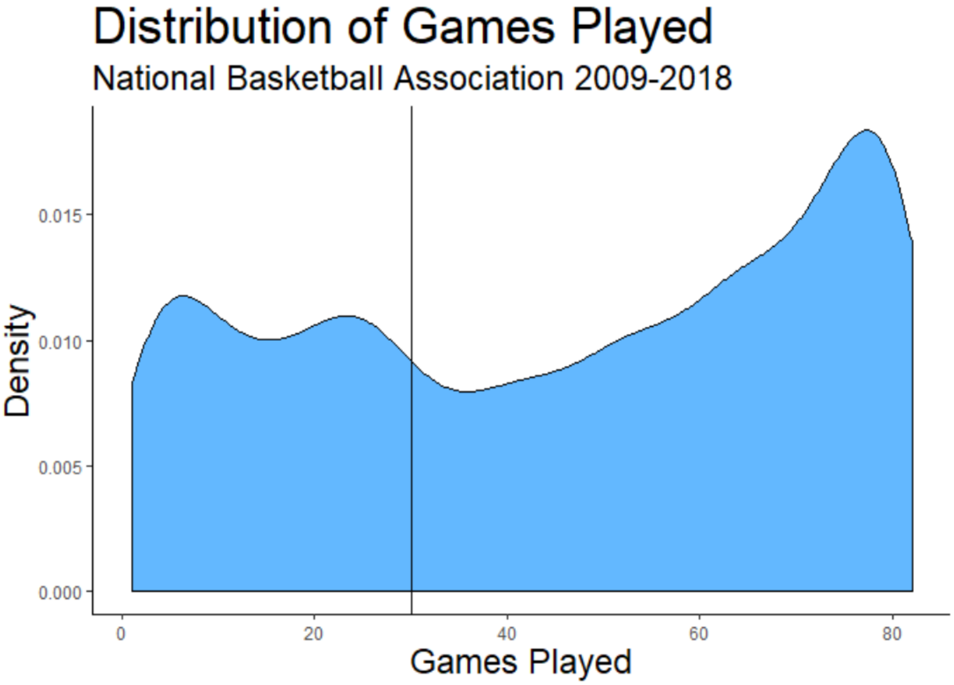


Figure 1 displays the distribution of games played in a season for each player-season observation in our dataset, indicating three modes of the games played near 5 games, 25 games, and 75 games. We arbitrarily filtered our data to only contain players that exceed thirty games played for a season, providing an adequate sample size to accurately represent how that player performed during that season. This resulted in a dataset of 3,608 player-season observations.

In order to evaluate the lineup effectiveness of the combinations of player clusters, we collected the last ten seasons (2009-2018) of NBA five-man lineup data from stats.NBA.com [5]. We had 2,000 unique lineups for each of the ten seasons in our data set, totaling 20,000 lineups with each player name, the season, the team, and other advanced lineup statistics such as net rating. Net rating, developed by Dean Oliver [6], is the lineup’s scoring differential per 100 possessions. Later, we will discuss how we used these data to evaluate player combinations.

2.2 Variable Selection

To cluster players based on their efficiency and playing styles, we needed to select statistics and measurements that accurately captured these traits. We also factored in statistics that accounted for the opportunity the player gets when on the floor, such as points and field goal attempts, both calculated on a per 100 possession scale. We collectively chose 23 of the variables that we believe best account for what a player does when he is on the court, specifically regarding their skills, habits, and opportunity. Below, the table defines each of the 23 variables. Notice, all statistics are calculated as rates, besides height. This allows us to control for the massive differences in playing time for NBA players throughout an entire NBA season (Player A may score more total points than

Player B, who does not get as much playing time as Player B. We cannot say who is the better scorer, however).

Table 1: *Brief Description of our 23 variables used for clustering*

Variable	Description
Height	Player height, in inches
Offensive Rebound Rate	% of available offensive rebounds a player gets while on the floor
Defensive Rebound Rate	% of available defensive rebounds a player gets while on the floor
Assist Rate	% of teammate field goals that a player assisted while on the floor
Steal Rate	% of opponent possessions that end with a steal by the player while on the floor
Block Rate	% of opponent field goal attempts blocked by the player while on the floor
Turnover Rate	Turnovers committed per 100 offensive possessions
Points	Points scored per 100 offensive possessions
Usage Rate	% of offensive team possessions used by the player while on the floor
Player Efficiency Rating	Per-minute production standardized such that the league average is 15
Free Throw Rate	Number of free throws made per field goals attempted
Free Throw Percentage	Number of free throws made per free throw attempt
Field Goals Attempted	Number of field goals attempted per 100 possessions
2FG%	Number of two-point field goals made per attempt
3FG%	Number of three-point field goals made per attempt
2FG Assist Rate	% of two-point field goals that are assisted
3FGA%	% of field goal attempts that are three-point attempts
Corner 3FGA%	% of three-point field goal attempts from the corner
3FG Assist Rate	% of three-point field goals that are assisted
Dunk Attempt Rate	% of all field goal attempts that are dunks
0-3 ft FGA%	% of all field goal attempts between zero and three feet from the basket
3-10 ft FGA%	% of all field goal attempts between three and ten feet from the basket
10ft-3p FGA%	% of all field goal attempts between ten feet from the basket and the three-point line

3. Clustering

3.1 Model- Based Clustering

To restructure the positions in the NBA, we used unsupervised clustering techniques, K-means clustering and model-based clustering. Unsupervised learning is a machine learning technique that attempts to group data that are unlabeled. For our purposes, the "new positions" of the NBA are unknown in the dataset. K-means clustering requires a desired number of target clusters, K, and every data point in our set is allocated to each of the clusters through reducing the within-cluster variation [7]. K means clustering can follow dimension reduction, in which we reduce our 23 variables into lower dimensional space. We attempted K-means clustering with and without a linear dimension reduction method called principal component analysis. We were not satisfied with K-means clustering due to the amount of suggested clusters, from the silhouette score [3], that were output as ideal (2, in both cases), as well as the statistical distributions of the data points, specifically the players and player stats contained within each cluster.

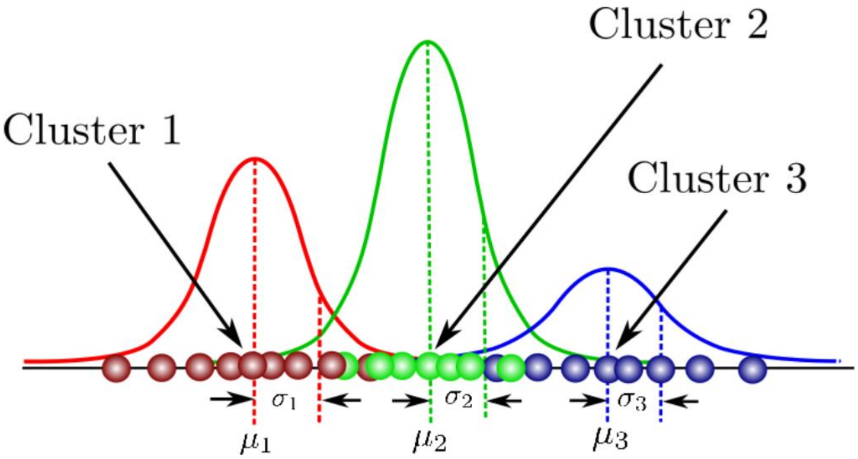
We then decided to approach our methods with model-based clustering. Unlike K-means, model-based clustering results in a soft assignment, indicating the probability each data point belongs to a cluster [8]. Model-based clustering uses an expectation-maximization (EM) algorithm to fit Gaussian finite mixture models to our 23-dimensional data frame. Unlike other clustering techniques, model-based clustering can treat choosing the number of clusters as a model-selection problem, using likelihood-based values such as the Bayesian Information Criterion (BIC) to select the optimal number of clusters. The algorithm finds the Maximum-Likelihood Estimate (MLE) of Equation 1 to find the optimal distribution underlying the unlabeled data. [9]

Equation 1: *Gaussian Mixture Model Likelihood Function*

$$L(\theta|X_1,\ldots,X_n)=\prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k N(x_i;\mu_k,\sigma_k^2)$$

In equation 1, the parameter μ_k defines the mean of the distribution, σ_k^2 is the variance, and π_k represents a mixing probability that defines how big or small the Gaussian function will be [10]. Figure 2 is a graphical look into how these parameters relate to the cluster distributions.

Figure 2: *Graphical Representation of Gaussian Mixture Model*



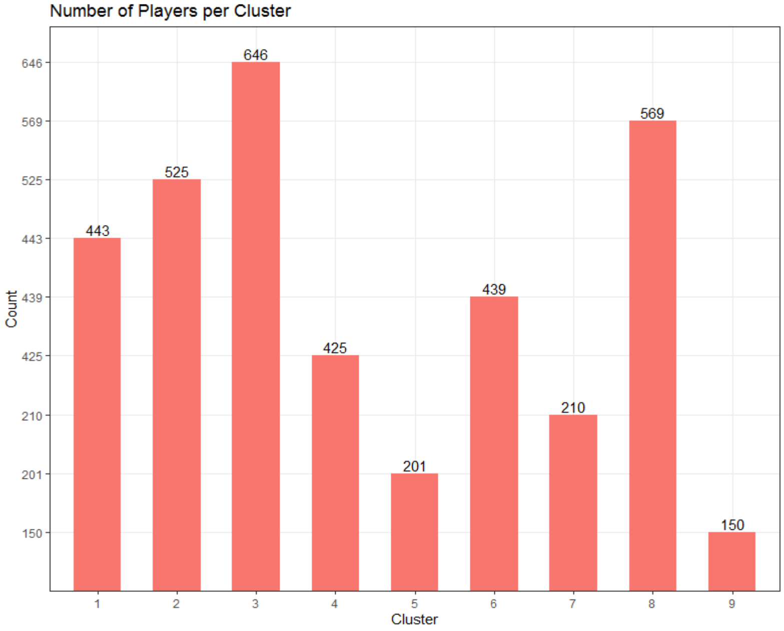
We used the “mclust” package in R [11] to implement model-based clustering via Gaussian components, providing us with the “soft” cluster probabilities based on each cluster distribution.

3.2 Restructuring Positions in the NBA

Our data contained variables of different scales. For example, rate statistics are represented as percentages, while height in inches, is represented as an integer. Since we are attempting to model the proximity of these points in high dimensional space, scaling the variables was an important step we took to ensure that each statistic would be weighted appropriately, relative to all the others.

Using the mclust () function from the mclust R package, the algorithm identified nine clusters in equal and ellipsoidal covariance structure based on BIC. This shape description is part of the output from mclust but cannot be visualized since the data is in 23-dimensional space. Figure 3 shows the distribution of player-season counts among the nine clusters.

Figure 3: *Distribution of the players among the nine clusters*



The cluster labels, numbered one through nine, do not have any inherent meaning so we performed extensive exploratory data analysis (EDA) on each cluster to understand the different types of skills and tendencies our algorithm identified. Our EDA included summary statistics, boxplots of statistics, and exploring the NBA player names within each cluster and comparing it with our prior knowledge of NBA basketball. Figure 4 is an example of one of the nine cluster breakdowns, the one we labelled "Three Point Shooting Guard".

Figure 4: *Breakdown of Cluster 6*

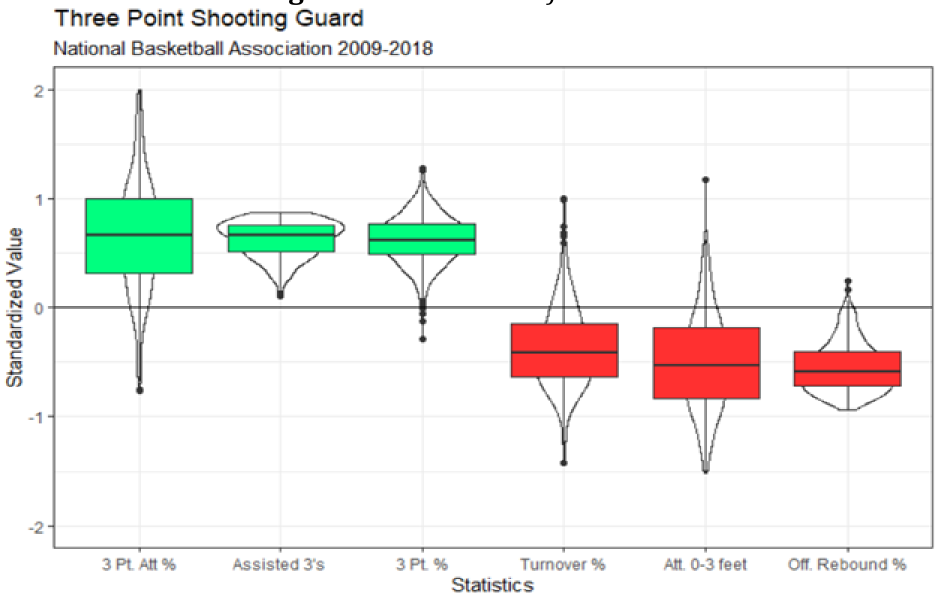


Figure 3 displays the summary of the distributions of scaled variables from the sixth cluster. From the nature of the scaled statistics, values above 0 (denoted in **green**) signifies the statistic is above average, compared to the same statistic in the other clusters. Likewise, values below 0 (denoted in **red**) are below average. In this example from cluster 6, most of the distributions of three-point attempt rate, threes made from assists, and three-point percentage lie above 0 (i.e. above average). On the other hand, cluster 6 tends to be below average on shots from 0 to 3 feet from the basket, turnover rate, and offensive rebound rate. The turnover rate being below average is a good trait, showing that these players do not turn the ball over to their opponent at a high rate, as turnovers are a loss of possession. From extensive EDA, we found that players from this cluster are predominantly catch and shoot players who rarely score around the basket, do not crash the offensive boards, and do not have the ball in their possession often enough, or for long enough to turn it over at a high rate. For these reasons, we named this cluster the "Three Point Shooting Guard". Table 2 displays a cluster name, example players, and a description for each of the nine clusters we identified.

Table 2: *Description of the new positions*

New Position	Description	High Stats	Low Stats	Example Players
High Usage Guard	A guard who operates with the ball in his hands and is a good distributer. Less efficient than a Ball Dominant Scorer and not as pass-first as a Floor General.	AST rate Usage rate	2FG AST rate Height	‘14 Lou Williams ‘14 Brandon Jennings
Stretch Forward	A player whose role is to stretch the floor and hit threes. Taller and a better rebounder than a Three Point Shooting Guard. Does not dribble as much as a Skilled Forward.	3FGA % Height	Usage rate FTr	‘12 Shane Battier ‘13 Steve Novak
Three Point Shooting Guard	Catch and shoot three-point shooter. Shorter than a Stretch Forward. Does not have the ball in his hands as much as a High Usage Guard. Role is to shoot, versus create.	3FG% 3FG AST rate	OReb rate Turnover rate	‘17 Klay Thompson ‘18 JJ Redick
Traditional Center	Plays near the rim. Does not shoot much from as far as a Mid-Range Big. Highly effective rebounder, and rim protector.	Dunk att. rate OReb rate	3FGA% 3FG%	‘15 DeAndre Jordan ‘18 Tyson Chandler
Versatile Role Player	Average in most statistics. Does not excel in anything but is not well below average at anything. Mixture of guards and forwards.	2FG AST rate OReb rate	Points FGA	‘14 Shaun Livingston ‘18 Bam Adebayo
Floor General	Guard that is pass-first. Low in height and does not shoot as often as a High Usage Guard or a Ball Dominant Scorer.	AST rate Turnover rate	Height 2FG AST rate	‘11 Jason Kidd ‘12 Rajon Rondo
Mid-Range Big	Plays at the rim but can step out and shoot a 10-16 feet jump shot. Better defensive rebounder that an offensive rebounder.	10ft - 3p FGA% DReb rate	3FGA% 3FG%	‘09 Pau Gasol ‘15 Tiago Splitter
Skilled Forward	Tall Forward who is skilled. Can drive at defenders but most three pointers are assisted. Better rebounder than a Stretch Forward but shoots less three pointers.	Dreb rate 3FG AST rate	AST rate Steal rate	‘14 Anthony Davis ‘11 Serge Ibaka
Ball Dominant Scorer	More efficient than a High Usage Guard. Looks to score first but has passing ability when necessary. Most two pointers are made from driving to the basket. Most often the teams go-to scorer.	Points Usage rate	Corner 3FGA% 2FG AST rate	‘18 James Harden ‘18 LeBron James

3.3 Example Player

As previously mentioned, model-based clustering yields probabilities for each player belonging in each cluster. To present example players (Table 2) from each cluster, we chose prototype players for each cluster with an estimated 99.9% probability of membership, meaning these player's skills and tendencies are exemplary of the identified positional group.

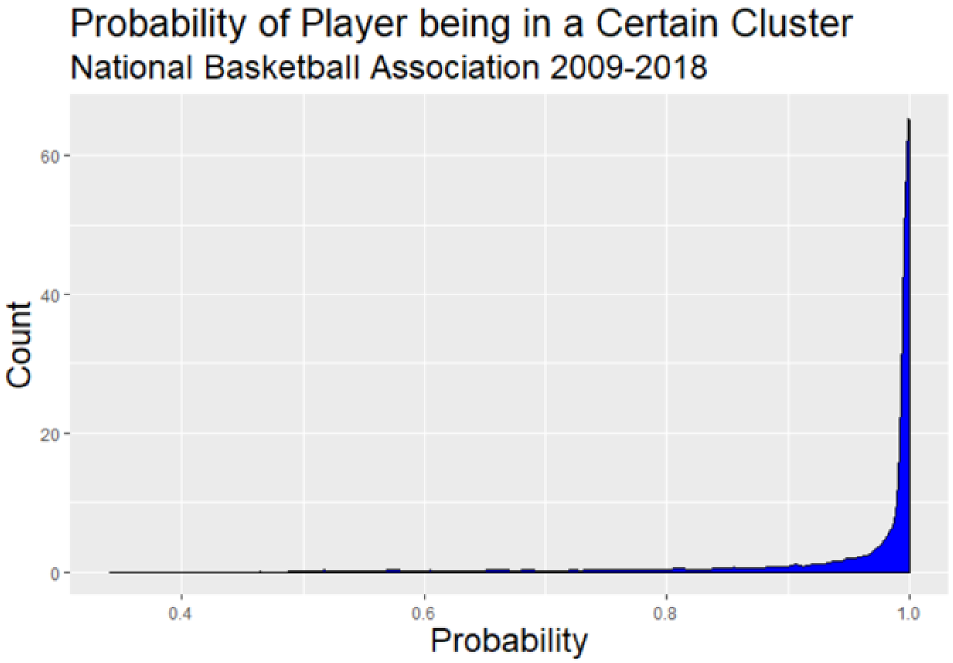
3.4 Cluster Consistencies and Changes

Given that our clustering technique is on a season level, we can recognize player evolution across their entire careers. This evolution can be as a result of developing certain aspects of their game, changing teams or coaches, and playing with different types of players in different environments. Kawhi Leonard is a player who began his career as a Stretch Forward on the San Antonio Spurs. As his career progressed and his role expanded, leading up to getting traded to the Toronto Raptors, he became a Ball Dominant Scorer. Brook Lopez is known for developing his three-point shot as a member of the Milwaukee Bucks in recent years. Our clustering model picks up on this change and shows us that he began his career as a Mid-Range Big, and evolved into a Skilled Forward, for which three-point shooting is far more relevant.

3.5 Cluster Likelihood Mapping

As discussed, one of the appealing features of model-based clustering is that we are able to calculate a probabilistic distribution of player likelihoods mapped to each of the clusters. While some players map very strongly to a single cluster, there are others who map more moderately to two or more. Figure 5 is the distribution of the probability that each player is a member of their most likely cluster.

Figure 5: *Distribution of probabilities a player belongs to their expectation-maximum cluster*



Data in Figure 5 is from the model-based clustering; each player's highest (most likely) cluster probability.

According to Figure 5, most players have a very large probability of being in one specific cluster. Based on skills and tendencies, our model found nine distinct groups of data with high group membership probability.

Some examples of players who had an even likelihood of being in multiple clusters were:

- **2017, Dirk Nowitzki** - 50% likelihood of being a Skilled Forward and a 50% likelihood of being a Stretch Forward. He was able to shoot three-pointers like a Stretch Forward as well as operate in the midrange, similar to a Skilled Forward.
- **2014, Matt Barnes** - 51% Stretch Forward and a 49% Three Point Shooting Guard. This is likely since he is 6 foot 7 inches and played as a mix between a guard and forward, with a high three-point rate.

3.6 New Positions versus Old Positions

Figure 6: *New Positions vs Old Positions*

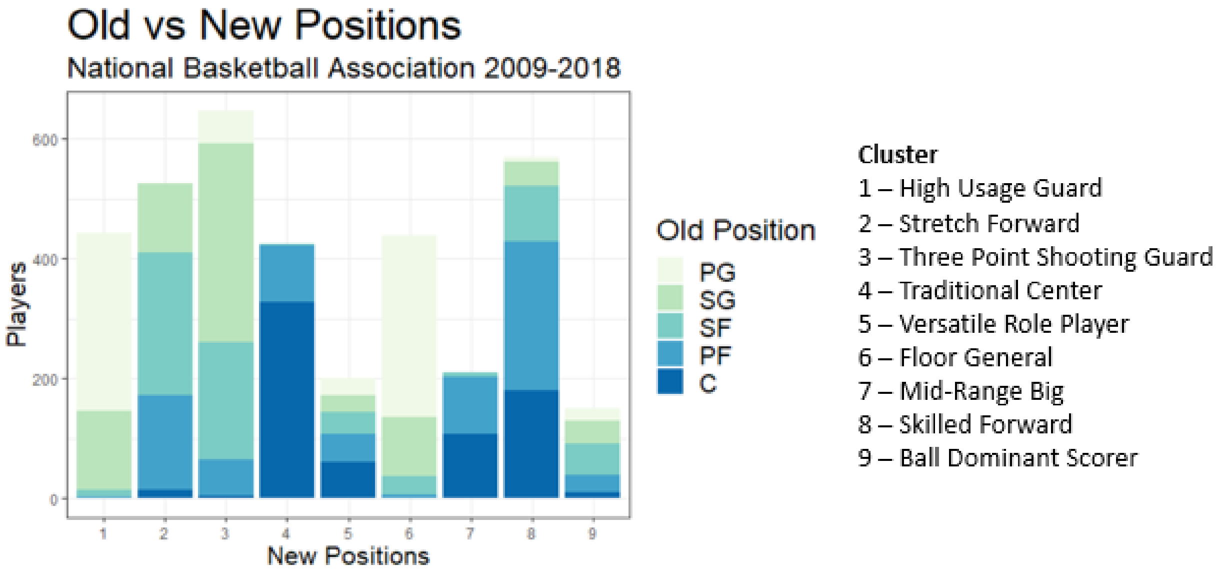


Figure 6 shows how the traditional positions compare with our nine new positions. Versatile Role Players and Ball Dominant Scorers are the two clusters that are most evenly spread among the traditional five positions. This table further illustrates the position-less aspects of basketball that have become apparent over the last few seasons. One interesting observation we saw was that one Point Guard is classified as a Traditional Center. This player is 2018 Shaun Livingston. In 2018, most of Livingston's shots came near the rim and he operated around the paint often enough for our model to classify him as a Traditional Center. Livingston was distributed between a 0.75 Traditional Center and a 0.25 Versatile Role Player. He is 6 feet 7 inches tall which is not short enough to pull him away from the Traditional Center cluster. His extremely unique playing style is difficult to categorize, even with our clustering. However, Livingston's role is better explained by a mix of a Traditional Center and a Versatile Role Player than by just as a Point Guard. Traditional Center and Versatile Role Player provide more insight into the habits and areas in which Livingston provides value to his team. He scores near the rim, grabs offensive rebounds, and does not shoot from the three-point line.

4. Lineup Analysis

4.1 Introduction

After clustering players into our 9 new positions, we were interested in identifying which combinations of players result in the most successful lineups. Given our knowledge of player likelihood mapping to each of the 9 positions, we had the ability to build fluid lineup combinations that included partial position components, specifically each player's cluster probabilities. With these lineup combinations, we built a model to project adjusted Net Rating.

4.2 Data

In order to build such a model, we collected five-man NBA lineup data from stats.NBA.com [5] over the same time span of our cluster analysis (2009-2018). Like the clustering analysis, our data was an aggregated total of that lineups season performance. The variable we were interested in predicting was Net Rating and is defined as the scoring differential when that five-player combination was on the court during a season, per 100 possessions.

Lineup data tends to be very noisy since some combinations of players play together far more often than others. If a certain combination of players only played a few minutes together the entire season, with 10 total possessions played, and went on a 10-2 scoring run in that span, then that lineup's net rating would be skewed to make them look a lot better than they might actually be. In this case, they would have a net rating of +80. NBA lineup data can be very noisy because a lot of basketball randomness can happen in small time periods without getting a true reflection on what is going on [12]. To counteract noisiness of the lineup net rating, we created an "adjusted Net Rating" using a simple empirical Bayesian component. This allowed us to maintain our sample size, while adjusting for outliers most likely present from noise. Equation 2 is how we calculated Bayesian Net Rating.

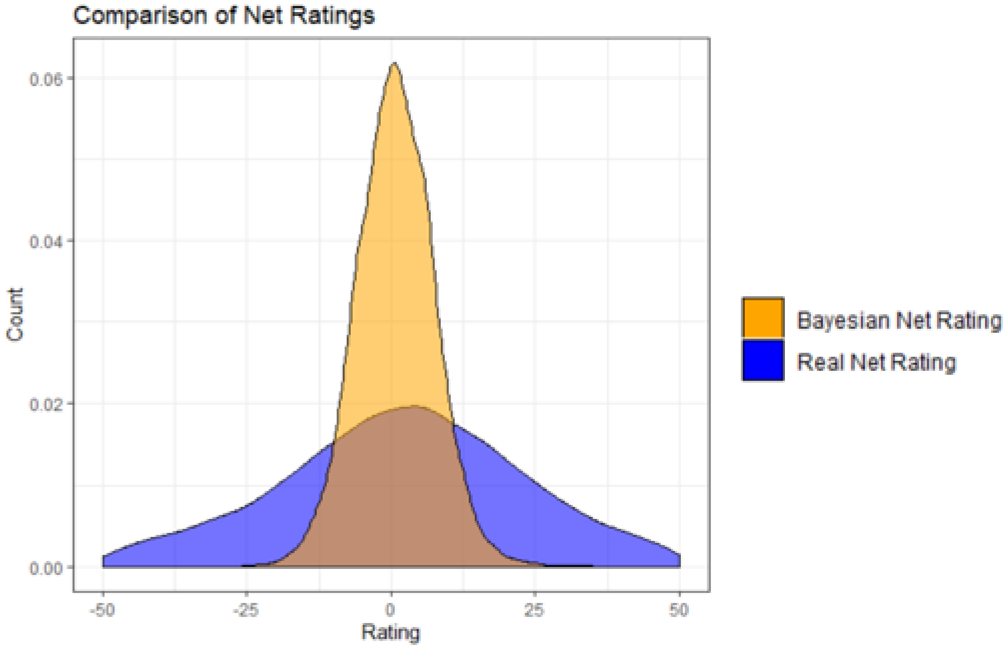
Equation 2: *Bayesian Net Rating*

$$If\ \frac{Possessions}{600} \geq 1, \qquad R = Net\ Rating$$

$$If\ \frac{Possessions}{600} < 1, \quad R = \left(\frac{Possessions}{600}\right) * Net\ Rating + \left(1 - \frac{Possessions}{600}\right) * Team\ Net\ Rating$$

Equation 2 shows that lineups with less than 600 possessions, had its Net Rating adjusted. In our case, our prior belief for how a lineup will perform is the team's net rating from that season. If lineups had over 600 possessions played together (approximately 6 games), we feel this is enough data with repetition of basketball possessions against different opponents for us to accurately evaluate the lineup success with their actual net rating. If a lineup has fewer than 600 possessions played, we weighted their net rating towards the prior (team aggregate net rating). Note, only 4.5% of the lineup combinations exceeded the 600-possession mark.

Figure 7: *Bayesian Net Rating vs Actual Net Rating*



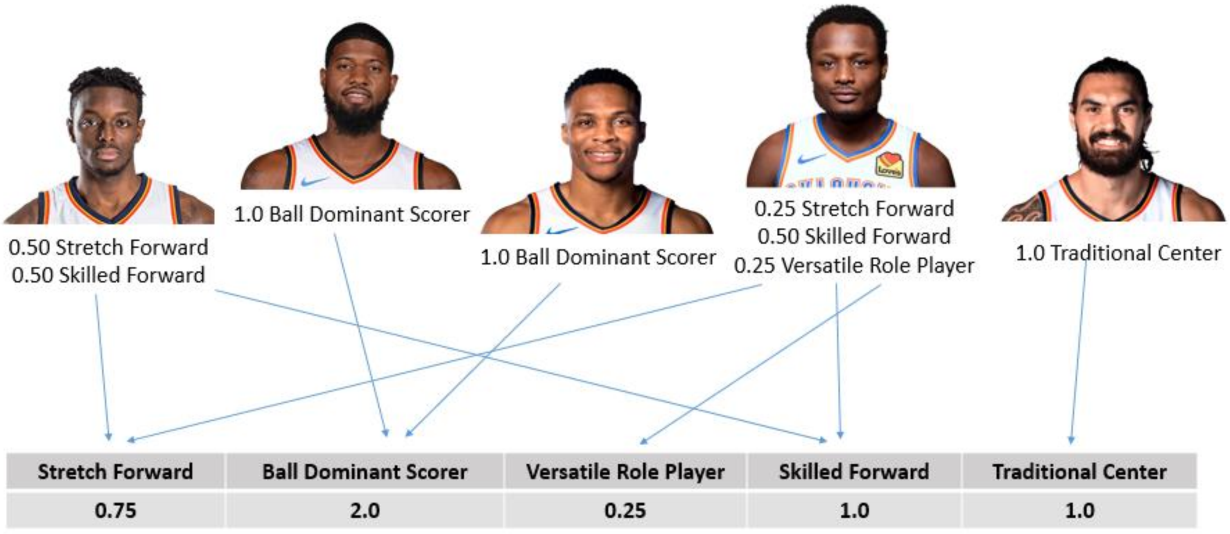
In Figure 7, the orange distribution is Bayesian net rating and the blue distribution is the actual net rating. You can see that the actual net rating's observed range is about 100, compared to the Bayesian net rating's observed range of about 50. A net rating near ± 50 is unrealistic to consist throughout an entire season, and these observations come from the lineups with much less than 600 possessions played. Consequently, the Bayesian net rating fixes this problem by weighting actual net rating towards the mean.

We began with 20,000 lineups from the ten years of data. Once we eliminated lineups that included players with less than 30 games played (our threshold for including a player in our cluster analysis), we were left with 14,234 lineups to build our model.

4.3 Creation of Predictor Variables

Our goal was to identify combinations of five-man lineups from the nine clusters that are most effective in terms of adjusted Net Rating. First, we swapped out the players name from the five-man lineup data for our cluster probabilities in the clustering data. Next, we created what we termed "soft lineups". Soft lineups take into effect that some players do not play 100% like a Ball Dominant Scorer or 100% like a Floor General. We consider the previously mentioned cluster probabilities to account for this fact. See figure 8 for how we used the cluster probabilities to build soft lineups of nine predictor variables for one of the 2018-2019 Oklahoma City Thunder lineups. Each soft lineup could potentially span all nine clusters or only one.

Figure 8: *Constructing Soft Lineups*



Stretch Forward	Ball Dominant Scorer	Versatile Role Player	Skilled Forward	Traditional Center
0.75	2.0	0.25	1.0	1.0

Figure 8 displays the technique in which we summed the cluster probabilities from all five players in the lineup. From this summation, we were able to create our nine predictor variables, representing the amount of each cluster present on the basketball court. In this example, the variables for the other four clusters not shown in the diagram would have values of 0.

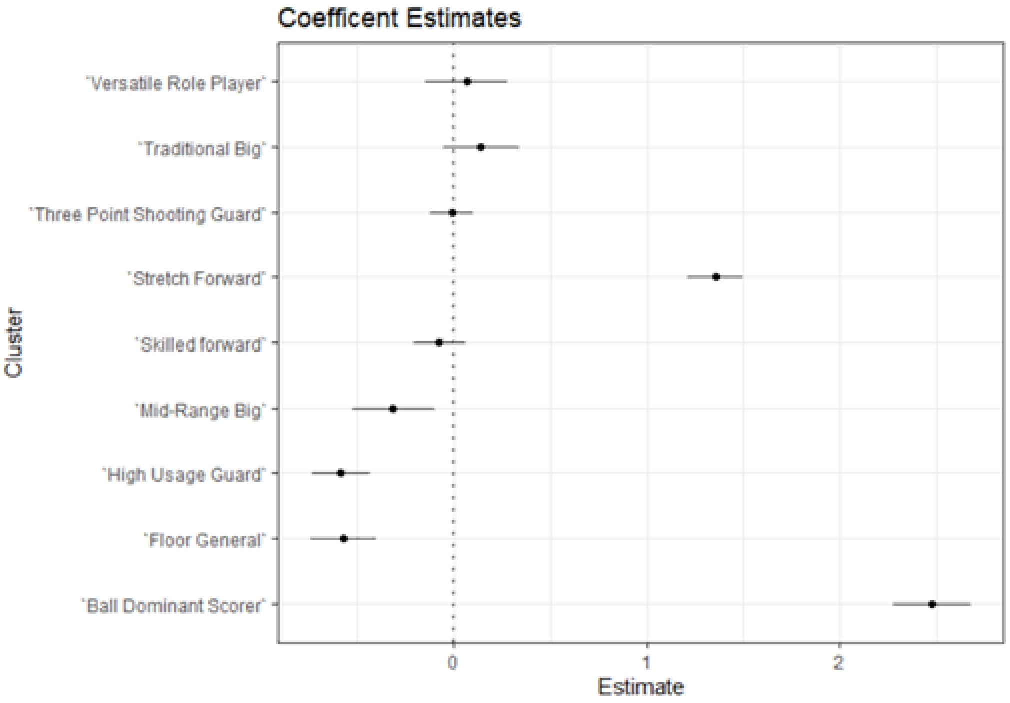
Recall (Figure 5) that our positional clustering model mapped most players very closely to just one of the positional groups identified. Because of this, many of the five-man lineup combinations contain only integer mapping onto the different clusters. For example, the 2019 champion Toronto Raptors soft lineup consisted of; 1 Floor General (Kyle Lowry), 1 Ball Dominant Scorer (Kawhi Leonard), 1 Stretch Forward (Danny Green) and 2 Skilled Forwards (Pascal Siakam and Marc Gasol).

For our lineup efficiency model, we have nine predictor variables, each representing the total amount of the cluster that is on the court from the five-man lineup combination.

4.4 Linear Regression model

Initially, we built a linear regression model to find the coefficients for the nine positional cluster predictors. This model attempts to show the linear relationship between the nine clusters and adjusted Net Rating. Figure 9 is the output from the linear model, showing the coefficients of each of the nine predictor variables, as well as the confidence intervals.

Figure 9: *Coefficients from Linear Model*



The linear regression model provides interpretability, with the tradeoff of not taking positional combinations into account (i.e. interactions between the clustered positions). According to our coefficients, if you increase the amount of Three Point Shooting Guards on the court by one, holding all else constant, the predicted adjusted Net Rating should increase by 1.36. This cannot happen, since if you add a Three Point Shooting Guard on the court, then you must also substitute another player off the court.

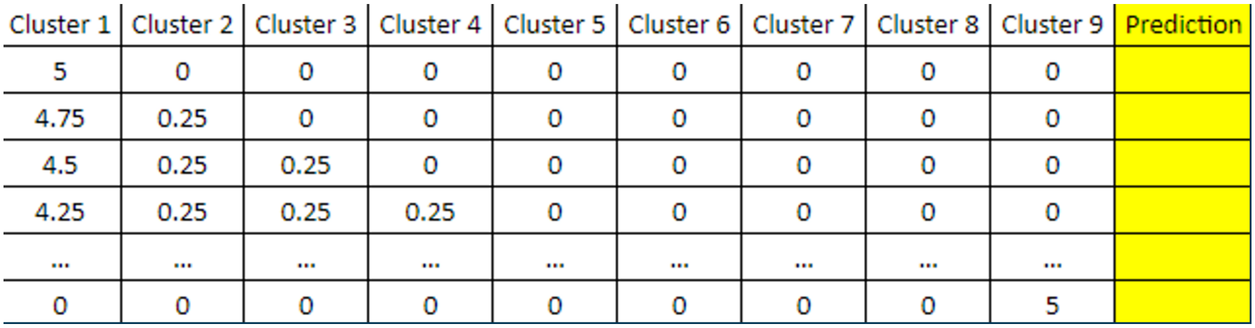
Also, our regression model suggests that if you have five Ball Dominant Scorers (the cluster with the largest coefficient), this will yield the highest adjusted Net Rating. Alternatively, if there are five Traditional Centers on the court, that lineup will have a positive efficiency rating. Both lineup constructions are highly unlikely to have the results that the linear model indicates. It is clear from this model, that the interactions between each of the clustered positions is key to building a successful model.

4.5 Random Forest Model

The next model we built was a Random Forest regression [13]. The Random Forest algorithm creates 500 decisions trees that are used to predict a continuous response. The trees are created by using ”bagging” to generate resampled versions of the dataset, fitting a decision tree at each resampled set. A random subset of variables is considered at each tree, the ensembled trees become more independent of each other, leading the decreased variance and better predictions [13]. At each branch of the trees, the algorithm is faced with a decision, in which the data itself chooses which path to go down, based on whatever requirement the branch offers. This happens repeatedly until the leaf of tree is reached where the prediction for the lineup is output. An example of a branch could be “0.99 from cluster 1”, in which the data would go to one branch if the lineup has less than 0.99 cluster 1 representation or go to the other branch if the lineup has great than 0.99 cluster 1 representation. The algorithm takes the average amongst the 500 different trees to come up with a prediction for that lineup. The trees are used to model the interaction within the clusters, because going all the way down multiple branches can tell the story of how many players you have from each cluster, or the ratio of representation between multiple clusters. The trees can analyze the effects of having specific cluster representation distributions across a specific group of clusters.

Our goal was to create predictions for all possible combinations of lineups, observed or unobserved. To do this, we need every possible lineup to be in our prediction set. Using a probability precision of 0.25, we created a matrix that contained all 3.1 million possible lineups. See figure 10.

Figure 10: *Prediction Frame*



Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Prediction
5	0	0	0	0	0	0	0	0	
4.75	0.25	0	0	0	0	0	0	0	
4.5	0.25	0.25	0	0	0	0	0	0	
4.25	0.25	0.25	0.25	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	
0	0	0	0	0	0	0	0	5	

To model the uncertainty of a predictive lineup efficiency, we bootstrapped 100 different random forest models, using all our lineup data. Bootstrapping is an effective way to model the variance in the model. This will yield us with 100 different Net Rating predictions for each of the 3.1 million possible lineups.

4.6 Modeling Results

From our random forest models, we found distinct patterns in the interactions between our clustered playing tendencies.

Figure 11: *Ball Dominant Scorer vs Stretch Forward*

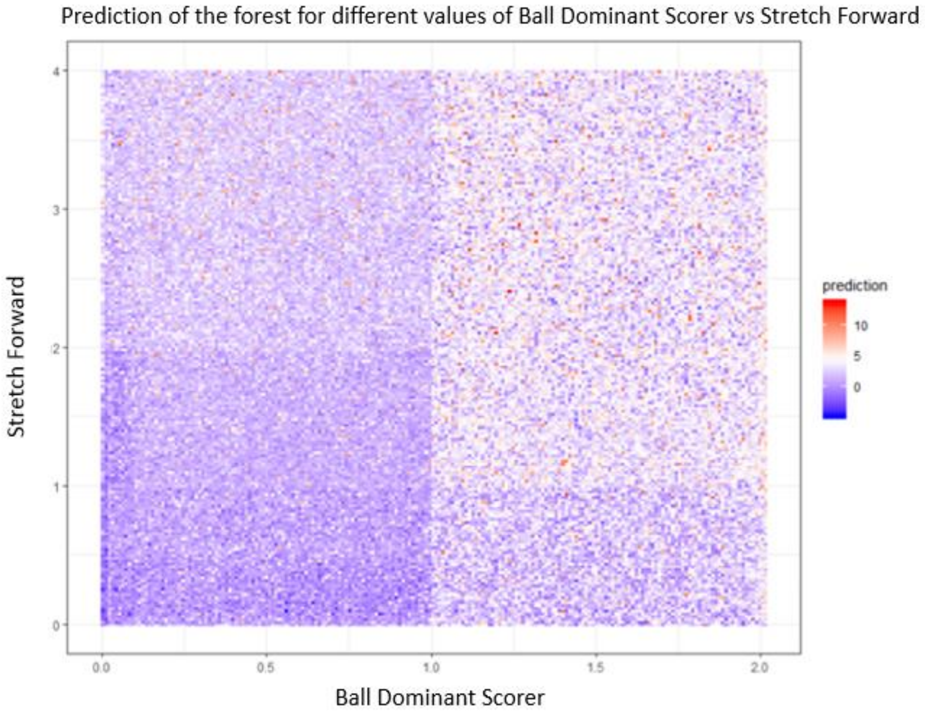


Figure 11 shows the relationship between Ball Dominant Scorers and Stretch Forwards in terms of predicted net rating. Most underperforming lineups have less than two Stretch Forwards and less than one Ball Dominant Scorer. It is very effective to have at least one Ball Dominant Scorer on the court and at least two Stretch Forwards.

What we see from our analysis is that most lineups have one high usage player, which means someone must have the responsibilities of handling the ball and getting the lineup into their sets. The high usage player will either fall into the category of Floor General, High Usage Guard or Ball Dominant Scorer. If that player can play as close to a Ball Dominant Scorer as possible (high efficiency, high assist rate), there is a greater chance of that lineup being effective. The model also suggests that spacing the court with shooters, combined with a ball dominant player leads to more effective lineups.

Figure 12: *Stretch Forward vs Versatile Role Player*

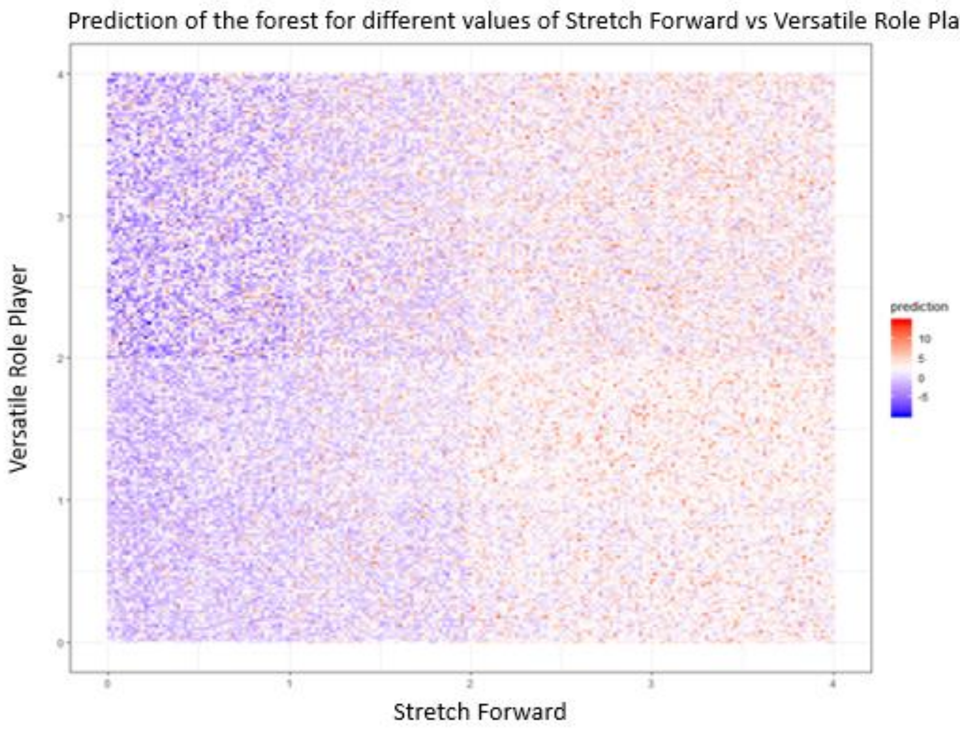


Figure 12 re-emphasizes the importance of spacing the floor with a stretch forward. The amount of role players on the court (versatile role players) are not as important if you have the ability to space the floor with stretch forward shooters.

4.7 The Best and Worst Performing 5-man Lineups

According to our 100 bootstrapped random forest models, our best performing lineup (lineup 1) has a predicted net rating of between 14.5 and 15.5. The lineup consists of 1.25 Ball Dominant Scorers, 2.25 Versatile Role Players 1 Traditional Center and 0.5 Stretch Forwards. The need to have your high usage player be efficient is evident in our highest predicted lineup. If you switch out the 1 Traditional Center and 1.5 of the Versatile Role Players to have 0.5 High Usages Guards, 1.25 Ball Dominant Scorers, 2 Stretch Forwards, 0.75 Versatile Role Players, 0.25 Three Point Shooting Guards and 0.25 Skilled Forwards (lineup 2), the net rating prediction is extremely close to the original, and still very successful. You can see that you can be successful having a guard dominated lineup if your able to stretch the floor with the two Stretch Forwards.

The lowest performing lineup (lineup 3) has 2 High Usage Guards, 2.25 Versatile Role Players, 0.25 Floor Generals and 0.5 Skilled Forwards. This lineup shows that the lack of a Ball Dominant Scorer and Stretch Forwards results in a predicted net rating between -10.7 and -10.3. We also see that both high and low performing lineups have over two Versatile Role Players. The amount of Versatile Role Players does not necessarily matter more than the amount of efficiency and floor spacing you put around those players.

Figure 13: *Prediction Distribution of Example Lineups*

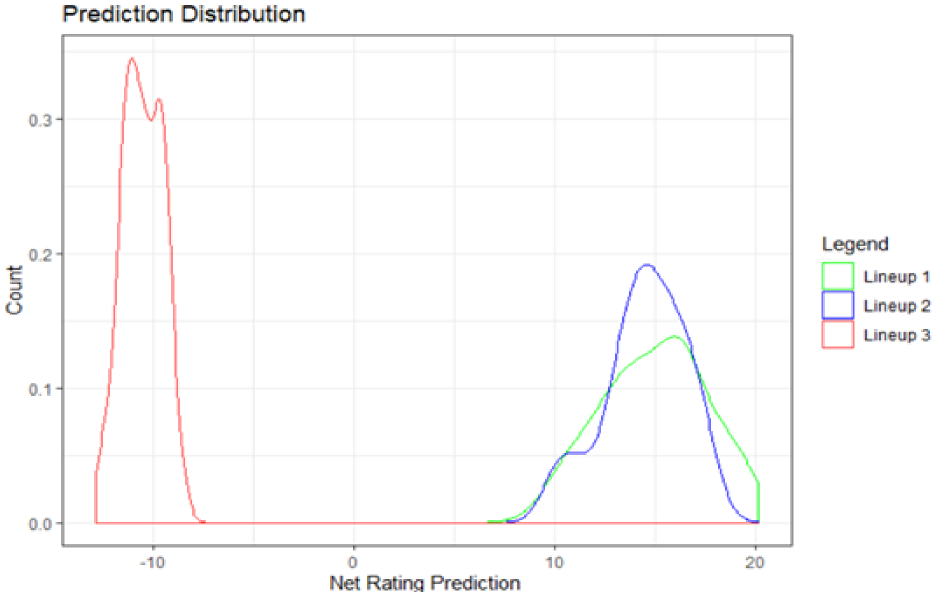


Figure 13 shows the distribution of net rating predictions for the two highest performing, and least performing lineups discussed previously. Recall that our methodology allows us to have 100 net rating predictions for all lineups. You can see that these three lineups, whether they are high or low performing, have certain performance predictions.

4.8 Warriors Death Lineup

For context, the lineup of Kevin Durant, Stephen Curry, Draymond Green, Andre Iguodala and Klay Thompson has a predicted net rating of 12.4 solely based on the five player clusters and has been termed the "death lineup". This lineup did not contain any "in between" players. The soft lineup for the death lineup consists of 1 Three Point Shooting Guard, 2 Ball Dominant Scorers and 2 Versatile Role Players, emphasizing the importance of the talent you put around your role players.

5. Conclusions

We found that NBA players can be clustered by tendencies, opportunity, and efficiency. We redefined the positions of basketball in a way that more accurately describes what role, effectiveness, and responsibilities a player provides when he is on the court. Our methods provide NBA teams, front offices, and fans in general an alternative way to position and categorize different types of basketball players.

We also found that combinations of the clustered positions do matter in terms of lineup efficiency. Playing five of the same types of players, such as Ball Dominant Scorers, will not yield as much success as putting five players on the court with complementary skills and playing styles. Specifically, we found that it is important that a lineup has an efficient high usage player. It is also key that you can create space with respectable shooters around that high usage player(s). Given that we are in the middle of the small ball and three-point revolution in the NBA, we found it particularly interesting that our model indicates that the traditional big man is not totally ineffective. The ability of centers and forwards to add opportunities through offensive rebounds, while limiting the other team's effectiveness shooting at the rim, are key to success. Where a big man is devalued is when he leaks out to shoot mid-range jumpers, thus taking on the role of a Mid-Range Big, which is not represented in any highly effective lineups.

NBA front offices can utilize our methodology in draft strategy, free agency, and trades. For example, if a team has four players they know they like on the court at the same time and NBA free agency is approaching, they can use our Random Forest model to identify which optimal cluster(s) is missing from their current four-man lineup. This could result in that team finding out what players, based on their cluster memberships, to target for acquisition in free agency.

There are many opportunities for extensions and future work of this analysis. We analyzed five-man lineup combinations, but we also can break it down even further and analyze two-man or three-man combinations in basketball. This would be applicable in the current NBA as many "super teams" build their rosters around two or three players. Additionally, we want to correctly quantify for the distributions of cluster probabilities. In other words, if a lineup has 1.0 players from cluster 3, we currently have no information regarding whether that 1.0 value is from one single player or if it summed to 1.0 from two or three players cluster probabilities. Lastly, we are going to use the 2019-2020 NBA season as a test for season long net rating predictions, making those predictions with our models, using last season's clusters for each player. We are excited to see how our predictions will hold.

In this research, we showed that machine learning techniques can be used to redefine the traditional positions of basketball in an increasingly "position-less' era of the NBA. This gave us the ability to predict high/low performing lineups based on cluster membership. Our 3.1 million lineup prediction set can be extremely useful for reasons discussed above. We hope that our findings can impact the NBA and the way in which people categorically consider a player's contributions and roles to their team, and the optimal combinations of those contributions and roles when on the court together.

6. Acknowledgments

We would like to say thank you to a few people who have inspired, aided, and help review this project. First off, thank you to Carnegie Mellon University, Rebecca Nugent, Peter Freeman, Alden Green, and Ron Yurko for their fantastic mentorship throughout the 2019 Carnegie Mellon Sports Analytics Summer Research Program, in which we were both students selected for the program. This paper is an extension of a project we began working on during the summer program and would not have happened without our acceptance to the program and their wonderful guidance. Next, thank you to Maksim Horowitz of the Atlanta Hawks, who originally inspired this project idea to us during the summer. Maksim has guided us every step of the way for this research and has really been a crucial figure in the development and strategy of this work. Lastly, we want to thank the 2019 Carnegie Mellon Sports Analytics Conference, and the 2019 Midwest Sports Analytics Meeting for hosting us and allowing us to present our work, which resulted in important, structural feedback.

References

[1] "NBA Player Stats: Advanced, Per 100, Shooting." Sports Reference LLC, www.basketball-reference.com/leagues/NBA.html.

[2] Cheng, Alex. "Using Machine Learning to Find the 8 Types of Players in the NBA." Fastbreak Data, 9 Mar. 2017, https://fastbreakdata.com/classifying-the-modern-nba-player-with-machine-learning-539da03bb824.

[3] Schoch, David. "Analyzing NBA Player Data II: Clustering Players." Schochastics, 4 Mar. 2018, http://blog.schochastics.net/post/analyzing-nba-player-data-ii-clustering/.

[4] "Analytics Movement." NBAstuffer, https://www.nbastuffer.com/analytics101/nba-analytics-movement/.

[5] "Lineups Advanced." NBA Stats, stats.nba.com/lineups/advanced/.

[6] Oliver, Dean. Basketball on Paper: Rules and Tools for Performance Analysis. Potomac Books, Inc., 2011.

[7] Garbade, Michael J. "Understanding K-Means Clustering in Machine Learning." Towards Data Science, 12 Sept. 2018, https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.

[8] "Model Based Clustering Essentials." Data Novia, https://www.datanovia.com/en/lessons/model-based-clustering-essentials/

[9] Bonakdarpour, Matt. "Introduction to EM: Gaussian Mixture Models." FiveMinuteStats, 22 Jan. 2016, stephens999.github.io/fiveMinuteStats/intro_to_em.html.

[10] Carrasco, Oscar Contreras. "Gaussian Mixture Models Explained." Towards Data Science, 3 June 2019, towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95

[11] Scrucca L., Fop M., Murphy T. B. and Raftery A. E. (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models The R Journal 8 /1, pp. 205-233

[12] "Getting the Most out of NBA Lineups with Machine Learning." ZigZag Analytics, http://www.zigzaganalytics.com/home/getting-the-most-out-of-nba-lineups-with-machine-learning.

[13] Breiman, L. Machine Learning (2001) 45: 5. https://doi.org/10.1023/A:1010933404324