

DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY


Evaluating Lineups and Complementary Play Styles in the NBA

Link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:38811515

Terms of use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at
https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility
https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story
The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

Contents

1 Introduction 1

2 Data 13

3 Methods 20
- 3.1 Model Setup . . . 21
- 3.2 Building Player Profiles Representative of Play Style . . . 24
- 3.3 Finding Latent Features via Dimensionality Reduction . . . 30
- 3.4 Predicting Point Differential Based on Lineup Composition . . . 32
- 3.5 Model Selection . . . 34

4 Results 36
- 4.1 Exploring the Data: Cluster Analysis . . . 36
- 4.2 Cross-Validation Results . . . 42
- 4.3 Comparison to Baseline Model . . . 44
- 4.4 Player Ratings . . . 46
- 4.5 Lineup Ratings . . . 51
- 4.6 Matchups Between Starting Lineups . . . 54

5 Conclusion 58

Appendix A Code 62

References 65

Acknowledgments

As I complete this thesis, I cannot imagine having completed it without the guidance of my thesis advisor, Kevin Rader; I am very lucky to have had a thesis advisor who is as interested and knowledgable in the field of sports analytics as he is. Additionally, I sincerely thank my family, friends, and roommates, whose love and support throughout my thesis-writing experience have kept me going.

Analytics don't work at all. It's just some crap that people who were really smart made up to try to get in the game because they had no talent. Because they had no talent to be able to play, so smart guys wanted to fit in, so they made up a term called analytics. Analytics don't work.

Charles Barkley

1

Introduction

Measuring an individual player's contribution to his team's success is a central question in basketball analytics. National Basketball Association (NBA) coaches and executives are tasked with building rosters and lineups that maximize their chances of winning a championship, so understanding which players make the biggest positive contri-

butions towards that goal is critical. In a team game such as basketball, however, it can be difficult to attribute team success to the individual players on a roster; indeed, confounding can be a big problem when five players play at a time, and up to twelve players can play for a team in any given game. A common solution to this issue is to look at per-game box score statistics such as points per game, rebounds per game, and assists per game. While this is a serviceable starting point to distinguish the league's superstars from its benchwarmers, these summary statistics are unable to capture many of the events that happen on the court; for example, they are unable to measure many of a player's contributions on the defensive end. Moreover, they are frequently misleading due to confounding factors like playing time, teammates, and opponents.

In an attempt to more completely summarize the impact of a player, basketball analysts turned to the concept of plus/minus. Plus/minus statistics begin from the fact that basketball is ultimately a game of points: the goal is to outscore one's opponent, or in other words, to achieve a positive point differential. The plus/minus framework extends this fact to the player level; fundamentally, plus/minus attempts to measure the point differential attributable to a given player. The most basic form of plus/minus is raw plus/minus, which is simply the difference between a player's team's points and the opposing team's points while the player is on the floor. Raw plus/minus was first used in hockey in the mid-20th century and became an official National Hockey League statistic in 1968; many years later,

the statistic was introduced in basketball and was ultimately adopted by the NBA in the mid-2000s[18].

Although this metric has the correct intention, it also has several flaws. First, it fails to consider the relative strength of the team on which a player plays. For example, Anthony Davis is widely considered a superstar for the New Orleans Pelicans but had a raw plus/minus of -217 in the 2015-16 season[2]. It would be misguided to conclude that Anthony Davis is a negative influence on the Pelicans; in fact, all but two players on the Pelicans had a negative raw plus/minus, and none had a raw plus/minus greater than 20, as the team finished with a poor record of 30 wins and 52 losses. A second, related flaw of raw plus/minus is that it fails to consider the teammates with which and the opponents against which a player frequently plays. This can be an issue when considering lineups containing particularly good or bad players; for example, Udonis Haslem was able to accumulate a raw plus/minus of +258 with the 2012-13 Miami Heat because he started the majority of the season's games with stars like LeBron James, Dwyane Wade, and Chris Bosh[2]. Again, it would be foolish to conclude that Haslem's 2012-13 season was superior to Davis's 2015-16 season based on their raw plus/minus numbers. Moreover, raw plus/minus fails to consider the quality of one's opponents, so a player who frequently plays against another team's starters will be judged the same as a player who typically plays against another team's backups.

Many extensions and modifications of raw plus/minus have been made to overcome these flaws. The first such metric is net plus/minus, which is defined for a player as a team's point differential when the player is on the court minus the team's point differential when the player is not on the court[3]. This metric alleviates the first concern about raw plus/minus by comparing the player's raw plus/minus to a baseline of the team's point differential without the player, rather than raw plus/minus's implicit baseline of 0 (note that when a team's point differential without a player is zero, that player's net plus/minus is equivalent to his raw plus/minus). Therefore, when a team is particularly bad, their players who contribute relatively positively are still able to achieve a positive net plus/minus, as their team is better with them than without them. Likewise, a poor player who achieves a high raw plus/minus by simply playing with a great team that accumulates a high overall point differential is not rewarded the same way under net plus/minus. Unfortunately, although net plus/minus does address the first issue with raw plus/minus, it fails to account for a player's teammate and opponent quality, so players who frequently play with or against particularly great or mediocre players will be judged by net plus/minus in part by with whom and against whom they play, rather than by their own individual contributions.

To explicitly control for a player's teammates and opponents, analysts have extended the plus/minus idea even further to create a metric called adjusted plus/minus (APM)[14]. APM accounts for the shortcomings of raw and net plus/minus by explicitly controlling

for a player's teammates and opponents on each possession in order to estimate the player's effect on the team's point differential, compared to an average player. The APM framework marks a distinctive shift from simpler statistics like net plus/minus in that its computation is not as straightforward as simply aggregating point differentials; instead, it is computed using a linear regression where the unit of observation is a possession, the predictors are indicators representing whether a given player is on the floor, and the response variable is the points scored on the possession. Specifically, this thesis will assume the following model for adjusted plus/minus, taken from Ilardi & Barzilai[10]:

$$y_i = \sum_{o \in P} \beta_{\text{off},o} x_{\text{off},io} - \sum_{d \in P} \beta_{\text{def},d} x_{\text{def},id} + \beta_{\text{HCA}} x_{\text{HmOff},i} + \beta_{\text{const}} + \epsilon_i \tag{1.1}$$

where y_i represents the points scored by the offense on possession i, P is the set of all players, $x_{\text{off},ip}$ is 1 if player p is in the game on offense for possession i and 0 otherwise, $x_{\text{def},ip}$ is 1 if player p is in the game on defense for possession i and 0 otherwise, $x_{\text{HmOff},i}$ is 1 if the home team is on offense for possession i and 0 otherwise, and ϵ_i is a Gaussian error term with zero mean. The coefficients $\beta_{\text{off},p}$ and $\beta_{\text{def},p}$ are the offensive and defensive APM ratings respectively to be estimated for player p, β_{HCA} is a coefficient to estimate home-court advantage, and β_{const} is the intercept included so the other coefficients are measured in points per possession relative to the league-wide average. The coefficients are estimated simultaneously by fitting the model using ordinary least-squares, which results in coefficient

estimates for each player in the data set, where the sum of one lineup's coefficients minus the sum of the opposing team's lineup's coefficients is the model's estimate of the expected point differential between the two lineups (typically prorated to 100 possessions). By including predictors for every player on the court, the model is able to "adjust" each player's plus/minus for the quality of the players he plays with which and against which he plays, thereby statistically isolating the player's contribution. This is much more effective in capturing the contributions a player can make that do not appear in the box score's summary statistics.

While APM provides a great improvement over raw and net plus/minus, it can have some problems in its basic form. One issue is multicollinearity, or high correlation between the predictors; when certain players are very frequently or very rarely on the court at the same time, it can lead to numerical instability and thus can cause problems during estimation. Another problem with APM is that it tends to overfit the data; when the number of training observations (in this case, possessions) is relatively low compared to the number of parameters being estimated (the number of players involved in the sample), simple linear regression tends to fit the data too closely. This results in a failure to generalize when making out-of-sample predictions, which is critical if APM is to be trusted as a predictive model for future performance. Finally, adjusted plus/minus assumes a player's level of play is constant over time for the duration of time used in the training data. This assumption is

inherent to the model, and while in reality, players' skill levels likely fluctuate in the short-term and certainly vary over the course of a career, this tends to be a reasonable simplifying assumption for the problem at hand because APM and its variants are generally used to estimate a player's contribution for a single year, during which a player's skill level is unlikely to fluctuate significantly. However, some research has been done to allow changes over time in APM estimates[6].

To address the issues of multicollinearity and overfitting, Sill[16] introduces the concept of regularization to the adjusted plus/minus framework, leading to the creation of Regularized Adjusted Plus/Minus (RAPM). Each player's RAPM is estimated by the same linear regression model used for APM, given in (1.1), except instead of estimating each player's coefficients using ordinary least squares (OLS), the coefficients are estimated using ridge regression. Whereas the OLS estimator minimizes the loss function

$$\sum_{i=1}^{n} \left(y_i - \boldsymbol{x_i}^T \boldsymbol{\beta}\right)^2 \tag{1.2}$$

where y_i is the response variable, $\boldsymbol{x_i}$ is the vector of predictors, and $\boldsymbol{\beta}$ is the vector of coefficients, ridge regression minimizes

$$\sum_{i=1}^{n} \left(y_i - \boldsymbol{x_i}^T \boldsymbol{\beta}\right)^2 + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta} \tag{1.3}$$

where λ is a regularization hyperparameter that controls the strength of regularization. The additional term introduced in (1.3) serves to penalize coefficients as they stray farther in either direction from 0; this constraints β and creates a tendency for coefficient estimates to be "shrunk" towards zero. Ordinary least squares is an unbiased estimator, meaning that $\mathbb{E}\left(\hat{\beta}_{\text{OLS}}\right) = \beta$; in other words, the OLS estimates $\hat{\beta}$ are equal to the true values of β in expectation. However, unbiasedness comes at the cost of higher variance due to the bias-variance tradeoff, and as a result ordinary least squares is prone to failure in situations with high multicollinearity or high dimensionality relative to the size of the training data[4,11]. Regulariation solves this problem by introducing some bias to the OLS estimator while reducing its variance. It has been shown that introducing regularization significantly improved the accuracy of the APM model (as measured in root mean squared error) when predicting out-of-sample point differentials[16].

These plus/minus metrics measure an individual player's contribution with varying degrees of success. However, we return the question that all NBA coaches and general managers face: how does a team best maximize its chances of winning a championship? As discussed previously, winning in basketball requires outscoring one's opponent, so this question reduces to one of team point differential. For a coach, this in essence reduces to deciding which five players should take the court at any given time in order to maximize one's chances of outscoring his opponent. For a general manager, it boils down to deciding how

to construct a roster that maximizes one's chance of winning. Before, we went on to answer these questions by examining each individual player's contribution to the team's overall point differential. However, if the reason we care about an individual's value is because it is integral to team success, then we ought to view the player in the context of his team. After all, basketball is a team sport, and coaches and executives in the NBA do not make their decisions about players in a vacuum; rather, they make decisions based on a player's value in the context of the team they already have. Indeed, chemistry is often critical to success in the NBA, so there is a strong incentive to ensure that players on a team fit well together. Therefore, it is crucial to analyze not only how much an individual player contributes to team plus/minus, but also how interactions between a team's players impact the team's overall point differential.

It is problematic, then, that none of today's player evaluation metrics consider interaction effects between players with different skill sets. Raw and net plus/minus do not directly consider one's teammates in computing a player's rating, so they do nothing to evaluate how well a player plays with different types of players; on the other hand, APM and RAPM make the explicit assumption that there is no interaction between player qualities. That is, these models assume that, given a home lineup with players $h_1, \ldots, h_5$ and an away lineup with players $a_1, \ldots, a_5$, holding nine players fixed and substituting h_1 with a new player h_6 will result in a change in expected point differential per possession equal

to the difference between h_6's rating and h_1's rating. Of course, the styles of play of h_1 and h_6 will often play a large part in determining the resulting change in point differential, as the way a lineup plays together can change completely when a substitution is made. While RAPM is the most prominent method of player evaluation in basketball analytics today, there has been some research concerning how players interact to produce outcomes that sets the groundwork for the approach I will take in this thesis.

The research that most influenced my approach was Maymin et al.[13] The authors introduce a framework they call "Skills Plus/Minus" (SPM), wherein a probit model is used to estimate the probability of a play ending with a specific outcome given which players are on the court and a few other predictors (e.g., which team is at home). Each player has an offensive and defensive rating associated with each possible play outcome. The authors then use these ratings and models of possession outcomes to simulate games by estimating the probabilities of various outcomes based on the lineups and then drawing from the corresponding multinomial distribution. Using this simulation framework, the authors were able to determine each skill's worth, measured in points over replacement player; which skills combined to increase overall point differential more (or less) than the sum of the individual skills; how each team's starting lineup is affected by synergies between players; and finally, which trades would be mutually beneficial. The main difference between this approach and the one I will present is that their approach explicitly models how synergies effect outcomes by mod-

eling each possible outcome of a play, rather than simply predicting the point differential on a given play; this granular modeling approach results in an increase in interpretability, but a decrease in expressiveness. The benefit is that it allows simulation based on SPM, which allows one to interpret exactly how different skills interact with one another. The downside is that, while each outcome is dependent on all ten players on the court, the probability models make some simplifying assumptions and do not allow for interactions or nonlinearities between players due to the model specification.

A very similar analysis was done by Kuehn [12] in which the author attempts to answer the question of how much a player contributes to a lineup by considering the probability of a possession ending with a specific outcome as conditional on the players involved in the game and their propensities to commit certain actions. Conceptually, this is very similar to the Maymin et al. approach, with two main differences. First, Kuehn considers the substitutability of different possession-ending outcomes. For example, adding a talented three-point shooter to a lineup makes it less likely that his teammates will take three-point shots. Second, while Maymin et al. used simulation to analyze different lineups, Kuehn used the possession model, the estimated propensities of each player to commit each possible action, and the estimated substitutability of actions to analyze lineups. With this approach, Kuehn reported which players helped and hurt their lineups the most, and whether NBA teams award positive complementarities when determining how much to pay a player.

Finally, the research of Arcidiacono et al.[1] created player ratings by modeling the probability of a given player scoring as a function of his ability to score, his teammates' ability to help him score, and his opponents' abilities to defend. Fitting this probability model results in three ratings for each player, which the authors used evaluate players and to estimate any theoretical lineup's expected point differential per possession. This analysis was compelling because the probability model was nonlinear and because it explicitly models how much each player's offensive contribution was divided between scoring itself and facilitating on offense.

My approach, while conceptually similar to all three of these analyses, will be slightly different. Roughly, the approach involves summarizing each player based on statistics describing his play and then using these profiles to predict expected per-possession point differential based on the play styles of the players on the court; for more detail, see Chapter 3. The most notable divergence of my approach is that it does not involve per-player estimands; rather, players are represented quantitatively by statistics summarizing their talents and play styles. Indeed, many summary statistics are used in order to fully capture a player's tendencies and play style. Moreover, whereas the other models assume additive relationships between player ratings, my approach leverages interactions between play styles to directly model the expected point differential per possession on the team level.

2

Data

In order to carry out the analysis herein, it was necessary to gather play-by-play data with complete lineups. To do this, I collected play-by-play data from the past eleven completed NBA seasons, starting with the 2005-06 season and ending with the 2015-16 season. The data was scraped from Basketball Reference [2]; this site was chosen for several

reasons. First, the website makes its data freely available and is organized simply and effectively, simplifying the scraping process. The second and more important reason the site was chosen is because it assigns unique identifiers to each entity for which it has data, including players, teams, seasons, coaches, and referees. Using these unique identifiers, it is very easy to cross-reference and join different data sources from the site; this makes complex analyses simpler because there it eliminates guesswork and ambiguities in data wrangling. For example, while other data sources may have ambiguities arising from players with similar names, Basketball Reference's unique identifiers make it easy to distinguish between, for example, Tim Hardaway Jr. and Tim Hardaway Sr. or Isiah Thomas and Isaiah Thomas.

The data scraping was done in Python using the sportsref library, an open-source Python library for scraping sports data from Basketball Reference and its sister sites, which include Pro Football Reference and Baseball Reference, among others. I am the sole creator, developer, and maintainer of the library, which was created to make sports analytics easier for those who understand sports and basic data analysis but do not have the programming knowledge to scrape data themselves. The library is still in the early stages of development; its code can be found on GitHub[8].

More specifically, the data used in this thesis was scraped from the play-by-play section that Basketball Reference provides for virtually every game since the beginning of the 2000-2001 NBA season. For an example of what this looks like, see figure 2.1. This page gives

Play-By-Play Jump to: 1st | 2nd | 3rd | 4th

scoring play | tie | lead change

1st Q				
Time	**Boston**		**Score**	**Golden State**
12:00.0	Start of 1st quarter			
12:00.0	Jump ball: A. Bogut vs. A. Johnson (S. Curry gains possession)			
11:39.0			0-3	+3 D. Green makes 3-pt shot from 25 ft
11:14.0	J. Sullinger misses 2-pt shot from 18 ft		0-3	
11:13.0			0-3	Defensive rebound by S. Curry
11:09.0			0-3	Turnover by S. Curry (bad pass)
11:00.0	A. Bradley misses 2-pt shot from 16 ft		0-3	
10:59.0			0-3	Defensive rebound by D. Green
10:48.0			0-6	+3 D. Green makes 3-pt shot from 23 ft (assist by A. Bogut)
10:35.0	E. Turner makes 2-pt shot from 18 ft (assist by J. Sullinger)	+2	2-6	
10:25.0			2-6	D. Green misses 3-pt shot from 28 ft
10:24.0	Defensive rebound by J. Sullinger		2-6	
10:16.0	A. Bradley misses 2-pt shot from 6 ft (block by D. Green)		2-6	
10:15.0	Offensive rebound by A. Bradley		2-6	
10:10.0	E. Turner misses 2-pt shot from 12 ft		2-6	
10:09.0	Offensive rebound by J. Sullinger		2-6	
10:07.0	Turnover by J. Sullinger (lost ball; steal by K. Thompson)		2-6	
9:54.0			2-6	Turnover by S. Curry (lost ball; steal by E. Turner)
9:48.0	A. Johnson makes 2-pt shot from 1 ft (assist by E. Turner)	+2	4-6	
9:24.0			4-6	A. Bogut misses 2-pt shot from 12 ft
9:20.0	Defensive rebound by A. Johnson		4-6	
9:13.0	I. Thomas misses 3-pt shot from 24 ft		4-6	
9:12.0			4-6	Defensive rebound by A. Bogut

Figure 2.1: An example of the Basketball Reference play-by-play source for a Warriors-Celtics game in April 2016.

detailed text descriptions of each event throughout the game, but does not contain structured data to identify, for example, whether an event is a rebound or who the shooter is on a field goal attempt. Therefore, regular expressions were constructed for each type of event in the dataset in order to create structured data from these descriptions; altogether, there are about twenty different types of events.

Moreover, several issues exist in the raw data that required cleaning. Perhaps the most difficult problem with the raw data is that some events were ordered incorrectly. Examples include rebounds being listed prior to their corresponding missed shots and inconsistencies

surrounding when substitutions and timeouts occur during free throws; these problems were remedied by a complicated sorting scheme within each possession so that events were listed in a logical order. Other, slightly simpler problems included tracking which team was on offense or defense at any given time and consolidating redundant substutitions in which a player would sub in and immediately sub out.

Once the raw data was cleaned, additional features were added to the data. For instance, simple new features include the home and away teams' points scored on each event, which can be used to easily compute point differential for any subset of events. One very important and more complicated feature was identifying "plays," a possession-like concept introduced by Maymin et al. In particular, the definition of a play is identical to the definition of a possession, except that an offensive rebound extends a possession but begins a new play. Distinguishing plays from possessions was important because it provides more context for the profile statistics described in section 3.2, as each play allows another opportunity to log events such as field goal attempts, rebounds, assists, and personal fouls. Therefore, when computing rate statistics for players, it is more revealing to use per-play statistics than per-possession statistics. Differentiating plays was done by labeling events that end a play with an indicator variable and computing the cumulative sum of these variables to create a play identification number for each event. Specifically, play-ending events include turnovers, field goal attempts on which fouls were not committed, free throws that are not succeeded

by another free throw, and the end of each quarter or overtime period. Possessions were identified using a similar procedure; while plays are more indicative of play style for most of the profile statistics in section 3.2, others (in particular, RAPM) require possession data. Moreover, throughout the analysis herein, technical fouls and technical free throws were discarded from the play-by-play data set, as they sometimes caused issues when identifying plays and possessions and are generally rare enough that they can be safely ignored for the sake of this analysis.

Next, analyzing lineups requires data that indicates which players are on the court for any given event. While the data provided by basketball-reference.com does not explicitly give this information, it does include substitution data, where each substitution is logged as an independent event. However, substitution events are only logged for substitutions in the middle of a quarter or overtime period; substitutions between periods were not indicated in the raw data. Therefore, to determine which players start each period, it was necessary to comb through each period and note which players are involved in events (including being substituted out) prior to being substituted in. When a player who started a period was never substituted out and was never involved in an event, the player can be determined by computing each player's minutes played based on the play-by-play data and comparing this value to his minutes played as listed in the box score. Once the starters of each period were determined, it is relatively straightforward to use substitution data to determine each

boxscore_id	secs_elapsed	poss_id	off_team	def_team	detail
201604050SAC	2416.0	222933	SAC	POR	couside01 misses free throw 1 of 1
201602280DAL	2440.0	168496	DAL	MIN	bareajo01 misses 2-pt shot from 4 ft
201512050TOR	1358.0	57123	TOR	GSW	Personal take foul by ezelife01 (drawn by jose...
201511290NYK	918.0	48221	NYK	HOU	willide02 makes 2-pt shot at rim (assist by ga...
201511070SAS	2229.0	17897	CHO	SAS	Defensive rebound by millspa02
201601040PHI	1807.0	99151	MIN	PHI	Defensive rebound by noelne01
201512090BOS	1534.0	61355	CHI	BOS	gibsota01 misses 2-pt shot from 1 ft (block by...
201512100CHI	1108.0	63424	LAC	CHI	redicjj01 makes free throw 2 of 2
201601290MIL	453.0	133985	MIA	MIL	bayleje01 enters the game for antetgi01
201511110SAC	2003.0	23768	DET	SAC	johnsst04 misses 3-pt shot from 24 ft
201512050PHI	2794.0	56839	PHI	DEN	stausni01 enters the game for mccontj01
201603120PHI	2122.0	187975	PHI	DET	stausni01 makes free throw 1 of 2
201603040BOS	440.0	174876	BOS	NYK	Defensive rebound by thomala01
201601300NOP	1384.0	135784	BRK	NOP	lopezbr01 misses free throw 1 of 2
201510300ORL	774.0	5375	ORL	OKC	fournev01 makes free throw 1 of 1
201511110ATL	1271.0	21868	ATL	NOP	New Orleans full timeout
201601030WAS	474.0	97574	WAS	MIA	Offensive rebound by gortama01
201511190CLE	1789.0	33784	MIL	CLE	antetgi01 makes 2-pt shot at rim (assist by pa...
201603150IND	388.0	191471	IND	BOS	Turnover by stuckro01 (traveling)
201602010SAC	2153.0	138904	SAC	MIL	Offensive rebound by acyqu01

Table 2.1: An example of events logged in the play-by-play data.

team's lineup for any given event.

It is important to note that Basketball Reference's play-by-play is not official, and in some cases there was missing data. For the seasons being used here, play-by-play was available for every game, but there were errors in parsing the raw data in four games out of the 13,289 games (0.03%) played during the eleven-season span; all of these errors were caused by inconsistent or insufficient substitution data in the play-by-play logs, resulting in an inability to determine the players on the court for at least one event in the play-by-play. These

games were simply ignored for the sake of this analysis. Moreover, in order to ensure that the data was by-and-large accurate, each player's raw plus/minus was calculated in each remaining game based on the lineup data inferred from the play-by-play and compared to the raw plus/minus listed in the box score summary provided by Basketball Reference. For 91.6% of the games for which there is complete data, there was no difference between any player's computed plus/minus and his official plus/minus. In 99.4% of the games, there is no absolute difference between the plus/minus figures that exceeds two points. Finally, in 99.91% of the games, the sum of all players' computed plus/minus was equal to zero, which is a necessary condition of correct raw plus/minus data within a game. These figures suggest that the vast majority of the data is correct and can be trusted with reasonably high confidence.

3

Methods

At a high level, the goal of this analysis is to determine how effective a lineup will be, taking into account the talents and tendencies of all ten players on the court. This involves two main steps: first, there must be a quantitative representation of each player's abilities and play style. Then, these "player profiles" of the players on the court can be used

to predict the outcome of each possession, measured in terms of points scored. Therefore, the problem can be stated as a regression with each possession as a unit of observation, points scored as the response variable, and features describing the talents and play styles of the players on the court during the possession as predictors.

A more in-depth discussion of the setup of the model is described in section 3.1. Then, the process of building player profiles to represent each player-season is described in section 3.2. Next, the process of reducing the dimensionality of these profiles is described in section 3.3 and the regression techniques for predicting points in a possession are described in section 3.4. Finally, the manner in which a model was selected and trained is described in section 3.5.

3.1 Model Setup

The goal of building player profiles for each player is to quantitatively represent a player's play style in a given season. Play styles can change from season to season due to factors like age and changes within a team, such as coaching changes or roster moves through trades or free agency. Therefore, in order to quantify a player's play style for season t, using statistics from within season t is essential. However, it is nonsensical to use play-by-play data from season t to construct player profiles if we are to use these player profiles to predict point outcomes in season t because the purpose of the model from the perspective of a coach is

to accurately predict point differential prior to observing these outcomes. Therefore, one possible solution is to use play-by-play data from the first half of season t to build player profiles and then to use these player profiles to predict point outcomes for each possession in the second half of season t. There are a few advantages to this setup; first, it ensures that the player's current play style is represented by using data from within the same season. Second, rookies can often become important contributors to a team, so it is advantageous that this model that can include such first-year players.

A second aspect of the model setup is that not every player in the league is given a representation; instead, the number of plays (defined in chapter 2) for which each player was on the court was computed, and the top 360 players were given unique player profiles. The remaining players were binned into a "replacement players" category, for which only one player profile was created using statistics from all such players. The cutoff value of 360 was determined by using the fact that there are 30 teams in the NBA and each team is allowed twelve active players on their roster in any given game; this approach was borrowed from the Maymin et al. analysis. This aspect of the model is both a computational convenience and a helpful addition to the model, for it allows comparison to a baseline replacement player. For example, we can evaluate a player by predicting the player's impact on the expected point differential between two teams of replacement players; similarly, lineups can be evaluated by predicting the point differential between the lineup and a lineup of replace-

ment players. See chapter 4 for such analyses.

One issue with the setup so far is that a player that would otherwise be in the top 360 players by number of plays may be treated as a replacement player if they are injured for the majority of the first half of the season. To avoid this issue, data from season $t-1$ was included in addition to the first half of season t in order to generate a player profile for season t. In order to give more weight to more recent data, statistics recorded in the first half of season t were weighted by a factor of 6. Because only half of season t is used, this weighting scheme results in roughly a 75%/25% split between statistics recorded in the first half of season t and statistics recorded in season $t-1$. As a result of this weighting scheme, players who played very significant time in season $t-1$ would still be given a unique player profile for possessions in the second half of season t, even if they suffer injury for all of the first half of season t. To generate rate statistics, this weighting scheme was used to find a weighted value for both the numerator and the denominator of the rate statistic, rather than weighting the rates from each season; this was done in order to avoid giving too much weight to either component of the blend due to a disparity in sample sizes. For example, a weighted version of free throw percentage for a player in season t was computed as $(\text{FTM}_{t-1}+6\text{FTM}_t)/(\text{FTA}_{t-1}+6\text{FTA}_t)$, where FTA_t is the number of free throws a player attempted in season t and FTM_t is the number of free throws a player made in season t.

3.2 Building Player Profiles Representative of Play Style

In order to quantify a player's play style, I computed many rate statistics from the play-by-play data described in chapter 2. It is important to note that all statistics are rate statistics, as the goal is to capture the manner in which a player plays, rather than the amount he plays. In order to take playing time out of the equation, statistics were computed per play, where a play is essentially any opportunity a team has to score (see chapter 2). While sample size can normally be an issue with rate statistics, this concern is alleviated in this case by the fact that all players not in the set of the 360 players with the most plays were binned together when computing statistics such as plays, field goals, and personal fouls.

In defining a player's play style, it is helpful to conceptualize the game of basketball as consisting of three components: offense, defense, and rebounding. Of course, each of these has its own subcomponents, but this division is still helpful for breaking down the game. For each of these aspects, I computed several rate statistics that describe a player's play. In addition to these traditionally-defined rate statistics in which a raw total of a certain event was divided by another, such as field goal attempts per play, I compute offensive and defensive regularized adjusted plus/minus ratings (ORAPM and DRAPM) for each player and for the replacement player category using the model setup described in equation 1.1 and estimating coefficients using a weighted version of the ridge regression described in 1.3. Specifically, possessions from the first half of season t are given six times the weight of possessions

from season $t-1$; these weights were applied by adding weights to the loss function for each training sample, so coefficients were found by minimizing

$$\sum_{i=1}^{n} w_i \left(y_i - \boldsymbol{x_i}^T \boldsymbol{\beta}\right)^2 + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta} \tag{3.1}$$

where $w_i = 6$ if possession i occurred in the first half of season t, and $w_i = 1$ if possession i occurred in season $t-1$. These ORAPM and DRAPM ratings were meant to indicate general overall skill on each end of the floor, and to capture the contributions a player can make that aren't captured by play-by-play data.

Finally, after the player profiles were computed, they were standardized within each year by subtracting the mean of a given profile feature within a given year and dividing by the standard deviation. To clarify, a player profile for season t includes data from season $t-1$ and the first half of season t; a weighted version of each rate statistic is computed first, and standardization only comes after all raw player profiles are computed. This has two justifications: first, it puts each feature on the same scale; second, it controls for the changing landscape of the NBA by comparing players to other players within a given year, rather than comparing player-seasons that took place in very different contexts.

Metric	Numerator	Denominator	Aspect of Play Style
FGA/play	field goal attempts	offensive plays	scoring burden
PF drawn/play	personal fouls	offensive plays	ability to draw fouls
FTA/FGA	free throw attempts	field goal attempts	ability to get to the line
FT%	free throws made	free throw attempts	ability to convert shooting fouls
% of 2-pt FGM assisted	assisted 2-pt FGM	2-pt FGM	tendency to create 2-pt shots
% of 3-pt FGM assisted	assisted 3-pt FGM	3-pt FGM	tendency to create 3-pt shots
% of FGA by region	FGA from region	total FGA	distribution of FGA distances
FG% by region	FGM from region	FGA from region	efficiency from different distances
AST/teammate FGM	assists	teammate FGM	tendency to facilitate
Lost balls/play	lost balls	offensive plays	tendency to lose the ball
Bad passes/play	bad passes	offensive plays	tendency to throw bad passes
Travels/play	travels	offensive plays	tendency to travel
Offensive fouls/play	offensive fouls	offensive plays	tendency to commit offensive fouls
Block rate	blocks	opponent 2-pt FGA	tendency to protect the rim
Steals/play	steals	defensive plays	ability to generate steals
PF/play	personal fouls	defensive plays	tendency to commit defensive fouls
Opponent TO/play	opponent turnovers	defensive plays	ability to generate turnovers
Offensive rebounding rate	offensive rebounds	offensive rebound opportunities	ability to grab offensive boards
Defensive rebounding rate	defensive rebounds	defensive rebound opportunities	ability to grab defensive boards

Table 3.1: Simple rate statistics collected to capture a player's play style.

3.2.1 Offensive Features

A player's offensive game can be broken down into two main aspects: the manner in which they score and the manner in which they handle the ball and avoid turnovers. While there are other aspects of a player's offensive game, such as movement without the ball, these aspects are difficult to measure using play-by-play data. In order to measure a player's scoring play style, the following rate statistics were collected: field goal attempts per play, free throw attempts per field goal attempt, free throw percentage, percentage of made two-point field goals that were assisted, percentage of made three-point field goals that were assisted, personal fouls drawn per play, percentage of field goal attempts that come from various distances from the basket, and field goal percentages from various distances from the basket. Each of these is meant to indicate a slightly different aspect of the manner in which a player scores; for example, field goal attempts per play is meant to measure the degree to which a player shoulders the scoring burden for his team, and free throw percentage is often used to measure a player's pure shooting ability in addition to his ability to convert shooting fouls into points. For a description of each of these statistics and what they intend to measure, see table 3.1.

The statistics describing how often and how efficiently a player shoots from various distances require more explanation. The regions chosen, as well as justifications for these choices, follow:

- 2-point shots that are 0-4 feet from the basket. This roughly represents shots from the restricted area, which tend to be layups or shots out of the post.
- 2-point shots that are 5-8 feet from the basket. This region roughly represents the rest of the paint, as the paint extends eight feet from the basket towards the sideline in each direction. These shots tend to be floaters or shots from the post.
- 2-point shots 9-16 feet from the basket. This represents the traditional mid-range shot; note for context that the free throw line is 15 feet from the basket.
- 2-point shots that are 17 or more feet from the basket. These are longest two-point shots, and are generally considered the least efficient shots in basketball.
- 3-point shots that are 23 or fewer feet from the basket. These represent three-point shots from the corner, as the three-point line is 22 feet from the basket in the corners, but 23.75 feet from the basket as it arcs from one corner to the other.
- 3-point shots that are 24-26 feet from the basket. This range represents non-corner three-point shots.
- 3-point shots that are 27-30 feet from the basket. This range represents deep three-point attempts, a shot that is typically taken only when the shot or game clock is near zero or when a shooter has exceptional range. Shots longer than 30 feet are rare and are typically desperation heaves at the end of a quarter, which are not indicative of play style.

In order to measure a player's tendency to handle the ball and to avoid or cause turnovers, the following rate statistics were collected: assists per teammate field goal made while the player was on the court, lost balls per play, bad passes per play, travels per play, and offensive fouls per play. The first statistic is a measure of how much of a team's offense relies on a player as a facilitator when he is on the court; it is important to note that teammate field goals made were only counted when the player was on the court. The final four statistics

are each a disjoint component of turnovers per play; indeed, lost balls, bad passes, travels, and offensive fouls are by far the four most frequent types of turnovers. This gives a more granular look into how a player turns the ball over.

Finally, as described previously, a player's offensive RAPM was also used to quantify a player's overall contribution to his team's offensive points per possesion, in addition to the rate statistics quantifying one's scoring and ball-handling.

3.2.2 Defensive Features

A common problem in quantifying a player's game is the lack of defensive statistics that are recorded; this is mostly due to the fact that defense has relatively few events that can be recorded. The main defense-related events that occur in the play-by-play dataset are blocks, steals, personal fouls, and turnovers. Therefore, the following rate statistics were collected to quantify defensive play styles: blocks per opponent 2-point field goal attempt (also known as block rate), steals per play, personal fouls per play, and opponent turnovers per play. It is important to note that for block rate, opponent 2-point field goal attempts were only counted if the player was on the court for the shot and thus had the opportunity to block it; also, block rate is traditionally computed using only 2-point field goal attempts because three-point shots are very rarely blocked. Likewise, opponent turnovers were only considered for plays where the player was on the court. These statistics are summarized in table 3.1.

3.2.3 Rebounding Features

Rebounding is a critical and often overlooked part of basketball that can make or break a team. However, there are few statistics that effectively describe a player's ability to rebound; here, only two rate statistics are gathered: offensive rebounding rate and defensive rebounding rate. These are defined as the number of offensive (or defensive) rebounds a player grabs per opportunity the player has to grab an offensive (or defensive) rebound. In other words, defensive rebounding rate is the total number of defensive rebounds a player grabs over the number of defensive rebounds his team grabs while he's on the floor plus the number of offensive rebounds the opposing team grabs while he's on the floor; offensive rebounding rate is defined similarly. For completeness, these are also described in table 3.1.

3.3 Finding Latent Features via Dimensionality Reduction

Player profiles comprised of the statistics described above were then passed through various dimensionality reduction algorithms in order to find latent features that describe a player's play style. This was done in order to make the features more succinct, in that dimensionality would be reduced, and to reduce multicollinearity, as many of the statistics above are correlated with one another. Several dimensionality reduction techniques, both linear and nonlinear, were tested in order to minimize cross-validation error on held-out data in the subsequent regression (see section 3.4). Specifically, principal component analysis (PCA),

Isomap, and locally linear embedding (LLE) were all used to reduce the dimensionality of the player profiles.

First, PCA is a well-known linear dimensionality reduction technique that generates linearly independent variables, called principal components, from the correlated player profiles. The first principal component is chosen such that it maximizes the amount of variance in the data that it explains; then, each subsequent principal component is chosen to maximize the amount of variance explained under the condition that it is orthogonal to each of the previous principal components[9].

Second, Isomap is a nonlinear dimensionality reduction technique that seeks to respect geodesic distances between points. The algorithm connects points that are near one another with edges and then computes the shortest path between any two points over these edges using a shortest-path algorithm such as Dijkstra's algorithm or the Floyd-Warshall algorithm. These geodesic distances are then used to form a distance matrix; the n eigenvectors of this distance matrix with the greatest eigenvalues are then used as the coordinates for the new n-dimensional space[19].

Finally, LLE seeks to reduce dimension while preserving local distances within neighborhoods of points. The LLE algorithm is similar to Isomap; like Isomap, it constructs a graph of nearest neighbors by connecting neighbors with edges. Then, a weight matrix is constructed by solving a linear system for the local neighborhood defined for each point as

the neighbors to which it is connected. Finally, like Isomap, the top n eigenvectors of this weight matrix are the coordinates for the new n-dimensional space[15].

All three of these techniques have different strengths, and it is almost impossible to identify when one technique will work and the other will not without very good knowledge of the structure and geometry of the data. Without such information about the player profiles, the best way to select a dimensionality reduction technique is to test each one and choose the model that gives the best performance on out-of-sample data; this is done via cross-validation and is described in section 3.5.

3.4 Predicting Point Differential Based on Lineup Composition

Once player profiles were passed through a dimensionality reduction algorithm, they were used to construct the design matrix for the regression predicting point outcomes on each possession. Specifically, the player profiles were incorporated into the regression via the following model:

$$y_i = f\left(a_{o_{1,i}}, \ldots, a_{o_{5,i}}, b_{d_{1,i}}, \ldots, b_{d_{5,i}}, \boldsymbol{x_{o_{1,i}}}, \ldots, \boldsymbol{x_{o_{5,i}}}, \boldsymbol{x_{d_{1,i}}}, \ldots, \boldsymbol{x_{d_{5,i}}}, h_i\right) + \epsilon_i \quad (3.2)$$

where y_i is points scored by the offense on possession i, $o_{j,i}$ is the jth best offensive player on the floor for possession i when ranked by total RAPM, $d_{j,i}$ is the jth best defen-

sive player on the floor for possession i when ranked by total RAPM, a_k is the ORAPM rating for player k, b_k is the DRAPM rating for player k, and $\boldsymbol{x_k}$ is the latent player profile for player k. Finally, h_i is an indicator for whether the home team is on offense on possession i and ϵ_i is a zero-mean error term. The function f represents some regression function that takes its inputs as predictors and models the response variable y_i. In other words, the predictors are the ORAPM ratings of the five offensive players, the DRAPM ratings of the five players on defense, and the dimensionality-reduced player profiles for each of the ten players on the court, in addition to an indicator for home court advantage. In order to impose an ordering on the players, both the offensive and defensive lineups were ordered by total RAPM.

Several possible choices for the regression model f were considered, including both linear and nonlinear regression techniques. Ultimately, linear regression, random forests[5], and gradient boosting[7] were all tested as regressors during cross-validation. Note that, although linear regression is, of course, a linear model, it does account for interactions between play styles to an extent due to the way the model is formulated. Because players are ordered by total RAPM, substituting one player for another is not as simple as the RAPM approach of adding the difference between the players' ratings; this is because a substitution could result in a different order after being sorted by RAPM, which would result in different predictors being matched with different coefficients. The final model was selected based on

performance in cross-validation, as explained in section 3.5.

3.5 Model Selection

The second half of each season contains roughly 114,000 possessions. Because of the sheer quantity of data, cross-validation over the entire dataset would be very time- and memory-consuming. Because we wish to test each possible combination of the dimensionality reduction techniques mentioned in section 3.3 and the regression techniques mentioned in section 3.4, in addition to several different hyperparameter settings for each of these techniques that can drastically change performance, it would be infeasible to do cross-validation on the entire ten-year dataset. In order to select the best model more efficiently, cross-validation was done on three seasons of data, starting with the 2011-12 NBA season and ending with the 2013-14 NBA season. Three-fold cross-validation was performed for each combination of dimensionality reduction technique and regression technique, as well as for various different hyperparameter settings for each; in particular, cross-validation was done so that in each fold, data from two of the three seasons were used to train the regression model and the third year was held out as a validation set on which to evaluate the model. Models were evaluated using root-mean-square error (RMSE), and the model was selected by choosing the model with the lowest cross-validation RMSE score. Once this model was selected, it was trained on data starting with the 2006-07 season and ending with

the 2013-14 season; the last two seasons in the dataset were held out in order to determine the test RMSE on unseen data. The results of this process are summarized in chapter 4.

4

Results

4.1 Exploring the Data: Cluster Analysis

In order to better understand the player profiles, I used the k-means clustering algorithm to separate player-seasons into clusters based on their tendencies. This algorithm works by adjusting cluster centers in order to minimize inertia, also known as

within-cluster sum of squares:

$$\sum_{i=1}^{n} \min_{c \in C} \|x_i - \mu_c\|^2$$

where C is the set of clusters and μ_c is the center of cluster c. The k-means algorithm was chosen over other clustering algorithms for a few reasons; first, it is a simple and easy-to-understand algorithm that often gives intuitive results and has few parameters to tune other than the number of clusters. Second, using inertia as the loss function for clustering assumes that clusters are convex and isotropic, which is a reasonable assumption for player types; in other words, the shapes of a cluster of players in player profile space should be roughly spherical, rather than elongated or oddly-shaped.

In order to apply k-means to the player profiles, I used player profile data from every season from 2006-07 to 2015-16, ignoring the "replacement player" category used for the regression analysis so that the 360 players involved in the most plays in each season were considered. As noted in section 3.2, the profiles were normalized within each year by subtracting a given season's average for each feature and dividing by its standard deviation; this is necessary because k-means treats each datum as a point in space, so it is necessary for each feature to be on the same scale.

Next, because k-means weights each dimension equally in computing distances, steps must be taken to ensure that the many offensive features do not dominate the relatively few defensive and rebounding features when distances are computed in k-means. To remedy

this problem, the variances of the five defensive and two rebounding features were inflated by factors of 2 and 3 respectively so that they are not drowned out by variance from the 25 offensive features. While these adjustments were ad-hoc, they yielded intuitive results; moreover, the fact that defensive and rebounding features are correlated with some offensive features ensures that some of this signal is encapsulated in the offensive features. For example, both rebounding features were heavily correlated with the proportion of a player's shots taken from within four feet of the basket, as was block rate; similarly, steal rate is correlated with a player's field goal percentage on corner three-point attempts.

In order to determine the optimal number of clusters k, the goodness of fit for a given value of k was determined by the average silhouette score over all data points. The silhouette score for a data point x_i assigned to cluster c of N_c points is given by

$$\frac{b_i - a_i}{\max(a_i, b_i)}$$

where a_i is the average distance between x_i and all other points in cluster c and b_i is the smallest average distance between x_i and the nearest cluster to which x_i does not belong:

$$a_i = \frac{1}{N_c} \sum_{x_j \in c} \|x_i - x_j\| \qquad b_i = \min_{c' \neq c} \frac{1}{N_{c'}} \sum_{x_j \in c'} \|x_i - x_j\|$$

Therefore, a silhouette score close to one indicates that the point is well-clustered, whereas

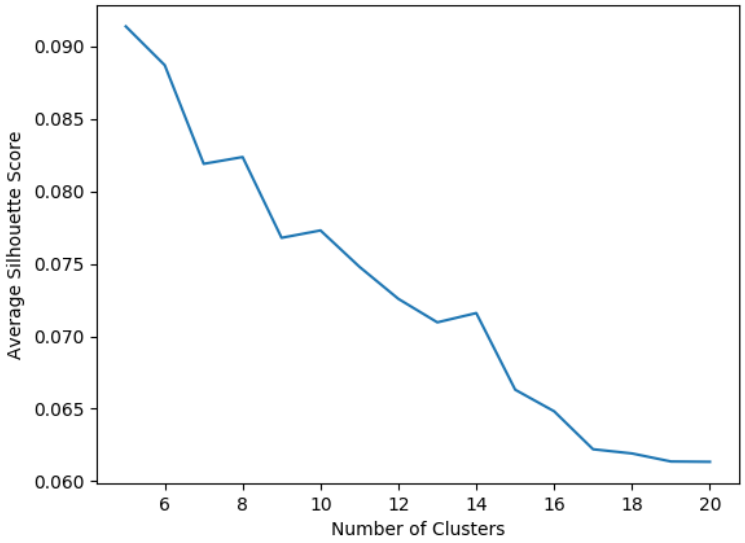


Figure 4.1: Average silhouette scores for various values of k.

a score close to zero indicates that a point is near the border between clusters. Thus, a good clustering will have a higher average silhouette score than a worse clustering. The average silhouette scores for various values of k are shown in figure 4.1.

From this figure, $k = 8$ was chosen to be the best choice for the number of clusters. This is because there is a relatively large dropoff after $k = 8$, whereas there is actually an increase in average silhouette scores from $k = 7$ to $k = 8$; moreover, while there are greater average silhouette scores for lower values of k, these values are so low that the clusters would not be as interesting. For example, if $k = 5$ was chosen, then the clusters would probably have a high correlation with the five traditional positions in basketball and thus the clustering would provide little insight.

After re-applying the k-means algorithm with $k = 8$, we arrive at cluster means as shown in table 4.1. Note that the means shown are the means of the data that is after standardization by year but prior to inflating the variance of the defense and rebounding features.

Most of these clusters can be mapped intuitively to a known player type in the NBA. For example, cluster 1 seems to represent three-point specialists like Kyle Korver or Courtney Lee, who don't shoot often but shoot most of their shots from three-point range, and are successful from that range; these players typically are assisted on their shots rather than creating their own shots off the dribble. Players in cluster 2 are capable of rebounding and protecting the rim, but are relatively unskilled offensively other than on shots within the paint; examples include Timofey Mozgov and Taj Gibson. The cluster mean for cluster 3 has a positive field goal percentage from every distance on the court and takes an above average number of shots per play; these are typically scorers who play on the wing, like Carmelo Anthony or Tobias Harris. The fourth cluster comprises players who facilitate their team's offense through assists, shoot well, and are very effective on defense by forcing steals and turnovers. These are players that tend to be point guards like Chris Paul and Rajon Rondo, but they can also come from other positions; for example, Paul George and Kawhi Leonard belong to this cluster. Cluster 5 contains skilled offensive big men who are able to make mid-range shots and stretch the floor, such as Kevin Love or Tim Duncan. Members of cluster 6 tend to be very good facilitators, as they average the highest rate of assists per team-

Feature	Cluster 1	2	3	4	5	6	7	8
FGA per play	-0.28	-0.65	0.38	0.22	0.59	0.53	-0.29	-1.03
AST per teammate FGM	-0.40	-0.83	-0.34	0.98	-0.36	1.09	-0.50	-0.82
FT %	0.36	-0.67	0.12	0.20	-0.04	0.62	-0.37	-1.42
FTA per FGA	-0.76	0.55	-0.14	-0.03	0.19	-0.23	0.10	1.47
PF drawn per play	-0.89	0.12	0.03	0.21	0.66	0.11	-0.08	0.55
Lost balls per play	-0.77	-0.03	-0.00	0.36	0.49	0.40	-0.21	-0.08
Bad passes per play	-0.39	-0.79	-0.29	0.96	-0.32	0.88	-0.33	-0.71
Travels per play	-0.41	0.31	0.13	-0.13	0.52	-0.16	0.24	0.12
Offensive fouls per play	-0.66	1.06	0.01	-0.24	0.35	-0.43	0.29	0.85
Percent of 2-pt FGM assisted	0.32	0.79	0.31	-0.86	0.56	-1.12	0.67	0.56
Percent of 3-pt FGM assisted	0.65	-1.42	0.59	0.27	-0.40	0.21	0.33	-1.56
Percent of FGA, 2-pt $\leq$ 4 ft	-0.86	0.89	-0.23	-0.19	0.35	-0.52	0.29	1.95
Percent of FGA, 2-pt 5-8 ft	-0.74	1.05	-0.08	-0.34	0.91	-0.30	-0.02	0.86
Percent of FGA, 2-pt 9-16 ft	-0.54	0.53	0.08	-0.15	0.95	0.14	0.00	-0.56
Percent of FGA, 2-pt 17+ ft	0.08	-0.40	0.20	0.09	0.07	0.34	0.29	-1.29
Percent of FGA, 3-pt $\leq$ 23 ft	1.33	-0.82	0.02	0.04	-0.76	-0.06	-0.23	-0.86
Percent of FGA, 3-pt 24-26 ft	1.06	-1.09	0.13	0.30	-0.95	0.37	-0.41	-1.13
Percent of FGA, 3-pt 27-30 ft	0.54	-0.82	-0.02	0.30	-0.72	0.56	-0.35	-0.82
FG%, 2-pt $\leq$ 4 ft	-0.20	0.30	0.20	-0.20	0.47	-0.38	0.14	0.31
FG%, 2-pt 5-8 ft	-0.18	0.18	0.15	-0.12	0.37	0.10	-0.23	-0.16
FG%, 2-pt 9-16 ft	0.04	-0.04	0.02	-0.06	0.25	0.22	-0.16	-0.53
FG%, 2-pt 17+ ft	0.13	-0.10	0.14	0.07	0.21	0.24	-0.02	-1.21
FG%, 3-pt $\leq$ 23 ft	0.51	-1.08	0.32	0.31	-0.69	0.47	-0.06	-1.16
FG%, 3-pt 24-26 ft	0.53	-1.32	0.33	0.33	-0.48	0.50	-0.05	-1.25
FG%, 3-pt 27-30 ft	0.39	-0.76	0.25	0.26	-0.40	0.35	-0.22	-0.90
Blocks per opponent 2-pt FGA	-0.51	1.26	-0.13	-0.52	0.68	-0.75	0.29	1.37
Steals per play	-0.24	-0.79	-0.46	1.48	-0.51	-0.02	0.31	-0.29
Personal fouls per play	-0.34	1.39	-0.28	-0.37	-0.03	-0.70	0.61	1.03
Opponent TO per play	-0.07	-0.30	-0.46	0.81	-0.46	-0.48	0.94	0.18
Offensive rebounding rate	-0.75	1.11	-0.08	-0.64	1.08	-0.88	0.49	1.84
Defensive rebounding rate	-0.65	0.47	0.24	-0.58	1.59	-0.95	0.42	1.48
Defensive RAPM	0.03	-0.10	-0.43	0.51	-0.01	-0.63	0.60	0.50
Offensive RAPM	-0.03	-0.48	0.12	0.17	0.18	0.18	-0.21	-0.16

Table 4.1: Player profile for the average member of each cluster, measured in standard deviations from the year's mean in a given feature.

Cluster	Description	Example Player
1	Three-Point Specialist	Kyle Korver
2	Big men with poor offensive skill	Timofey Mozgov
3	Scoring wings	Carmelo Anthony
4	Facilitator with strong defense	LeBron James
5	Floor-stretching big men	Kevin Love
6	Facilitator with poor defense	Tony Parker
7	Strong defensive wings and bigs	Ersan Ilyasova
8	Elite rebounders and rim-protectors	Dwight Howard

Table 4.2: A summary of the interpretations of the eight clusters.

mate field goal made, while also being capable shooters and subpar defenders; examples include Tony Parker, Isaiah Thomas, and Kyrie Irving. Cluster 7 contains bigger wings and big men who are able to defend at a high level; the cluster includes Ersan Ilyasova and P.J. Tucker. Finally, cluster 8 mostly contains elite rebounders and rim protectors; these are the stereotypical centers, such as Dwight Howard and DeAndre Jordan. Understanding these clusters can be helpful in interpreting some of the results below, where clusters allow an easy way to distill a player's profile into a single player type; a summary of the interpretation of these clusters is given in table 4.2.

4.2 Cross-Validation Results

As explained in section 3.5, cross-validation was performed using data from the 2011-12 season through the 2013-14 season; many different cross-validation runs were done in order to tune model hyperparameters; the results are summarized in table 4.3.

Optimal Hyperparameters	Optimal Hyperparameters	Linear Regression	Random Forests	Gradient-Boosting
		N/A	300 estimators, max depth 3	300 estimators, max depth 3, learning rate 0.1
PCA	3 components	1.263605*	1.264270	1.264506
Isomap	5 components, 5 neighbors	1.266315	1.266019	1.267266
LLE	5 components, 5 neighbors	1.266406	1.266016	1.267420

Table 4.3: A summary of the results of hyperparameter tuning and cross-validation; each entry in the table is the cross-validation root mean squared error for a given combination of dimensionality reduction and regression techniques.

Somewhat surprisingly, the regression model with the best performance in cross-validation is the simple linear regression model. Moreover, less surprisingly, principal components analysis beat out the nonlinear dimensionality reduction techniques, as models trained using PCA outperformed Isomap and LLE regardless of the regression model used. This is unsurprising, as there was no known non-flat manifold geometry known to exist in the data. As noted in section 3.4, the linear regression model does encapsulate interactions between players, to an extent; this is because the players are sorted in order of total RAPM, so unlike the baseline RAPM model discussed in the introduction and in section 4.3, substituting player 1 for player 2 is not as simple as adding player 1's ratings and subtracting player 2's ratings because the substitution can result in a shuffling of the order of the players in the regression, which in turn means that a player's feature values may be matched with different coefficients if his position in the sorted lineup changes as a result of the substitution. There-

fore, while it is not a nonlinear model, this linear regression model still functions differently than the adjusted plus/minus model in that the effect of a substitution is not necessarily the same regardless of the other players on the court as it is in the APM model. Therefore, the model selected via cross-validation was to first use PCA with 3 components to reduce the dimensionality of the player profiles, followed by a simple linear regression.

4.3 Comparison to Baseline Model

The motivation for this model is to incorporate play styles in order to improve upon the regularized adjusted plus/minus model when predicting the relative performance of two lineups. However, the (R)APM framework is used as a method for evaluating individual players and is not used to produce predictions for two arbitrary lineups; therefore, an extension of the model is necessary in order to adapt it to predict the outcome of a possession. If one is to accept the RAPM model's assumption that the point outcome of a possession can be linearly allocated to the ten players on the court as in (1.1), then it naturally follows from the model to model the points per possession of one lineup against another as a linear combination of the players' RAPM ratings with a home court adjustment:

$$y_i = \sum_{p \in OP_i} x_{\text{off},p} - \sum_{p \in DP_i} x_{\text{def},p} + \beta_{\text{HCA}} x_{\text{HmOff},i} + \beta_{\text{const}} + \epsilon_i \tag{4.1}$$

where y_i represents points on a possession i, OP_i and DP_i are the sets of offensive and defensive players on the court for possession i respectively, $x_{\text{off},p}$ and $x_{\text{def},p}$ are player p's computed ORAPM and DRAPM ratings respectively, $x_{\text{HmOff},i}$ is an indicator for whether the home team is on offense for possession i, and ϵ_i is a normally-distributed error term with zero mean. Then, the coefficients β_{HCA} and β_{const} are optimized via ordinary least squares. Note that the ORAPM and DRAPM predictors are not given coefficients; this is because the RAPM model inherently assumes that these ratings combine linearly with equal weights for each player. This assumption is made by the APM model in (1.1) wherein points on a possession is modeled as equal to the sum of the offensive players' coefficients minus the sum of the defensive players' coefficients, using an indicator to control for home court advantage and an intercept term to control for league-wide average efficiency. In other words, this is essentially the same prediction that a fitted RAPM model would make, were one to use it to predict the outcome of a possession. Therefore, this model for predicting points per possession follows logically from the RAPM framework for player evaluation.

Both the baseline RAPM model and the full model including the PCA-transformed player profiles were trained on second-half possessions from the 2006-07 season through the 2013-14 season as described in chapter 3. A comparison of these models on a variety of metrics measuring performance on out-of-sample data from the 2014-15 and 2015-16

Metric	RAPM Model	Player Profile Model
R^2	0.0001632	0.001962
Root Mean Squared Error	1.29454	1.29440
Median Absolute Error	1.03275	1.03038

Table 4.4: Comparing the out-of-sample performance of the regression on player profile data to the baseline RAPM model.

seasons is given in table 4.4. Clearly, the model including the transformed player profile features as predictors outperformed the baseline model. Although the differences are small, this is largely due to the level at which the regression is being done; that is, the upper bound on the performance of a model that uses only play-by-play data to predict point outcomes on each possession is quite low due to the noise and variability in possession outcomes. For instance, both models tested here will always predict the same point outcome given offensive and defensive lineups; however, if these two lineups play for several possessions, they are likely to have several different point outcomes over the course of these possessions. This error is unavoidable given the model specifications of the RAPM model and of the player profile model. Therefore, the model is able to improve prediction accuracy on out-of-sample data compared to the RAPM model.

4.4 Player Ratings

Using the player profile model, a given player p can be evaluated by computing the expected point differential per possession that a team of p and four replacement players would achieve

against a team of five replacement players; this assigns a "points above replacement" (PAR) rating based on how much a player contributes to an otherwise replacement-level team. Expected point differential per possession is computed by using the model to predict the points per possession when the player's lineup is on defense and subtracting this value from the predicted points per possession when the player's lineup is on offense; for simplicity's sake, the offense is always assumed to be the home team, so that the home court advantage cancels out during subtraction. Using the model trained on data from the 2006-07 NBA season through the 2013-14 NBA season, ratings were calculated for all 297 players who played at least 820 minutes in the 2015-16 season and were not considered replacement-level players for the 2016 season in the player profile analysis; the minute cutoff was chosen because it corresponds to an average of 10 minutes per game over the 82-game season; moreover, because we are evaluating players for the 2016 season, the nine replacement players will be given the player profile for the 2016 replacement-level player. The best and worst of these ratings are shown in table 4.5, which includes each player's "points above replacement" over 100 possessions as well as their ORAPM, DRAPM, and total RAPM, also prorated to 100 possessions.

There are a few things to note when interpreting these ratings. First, recall that player profiles are built from data from both the prior season and the first half of the season in question; therefore, these 2015-16 ratings are based partially on the 2015-16 season itself and

Player	Cluster	PAR Rating	ORAPM	DRAPM	RAPM
Russell Westbrook	4	6.356	3.231	0.182	3.413
Stephen Curry	4	5.633	4.124	1.390	5.515
LeBron James	4	5.041	3.371	1.429	4.800
Chris Paul	4	4.652	3.227	0.228	3.455
James Harden	4	3.799	1.814	-1.090	0.724
Kevin Durant	3	3.645	3.152	0.924	4.076
Kyrie Irving	6	2.841	2.565	-0.848	1.717
Isaiah Thomas	6	2.746	1.952	-0.965	0.986
Damian Lillard	6	2.225	1.808	-1.463	0.345
Kyle Lowry	4	2.071	1.778	0.471	2.248
Draymond Green	7	2.070	3.045	2.056	5.101
Reggie Jackson	6	1.957	1.310	0.194	1.504
John Wall	4	1.893	0.848	0.229	1.077
Klay Thompson	1	1.793	3.367	0.198	3.565
Manu Ginobili	4	1.730	1.998	1.278	3.276
⋮	⋮	⋮	⋮	⋮	⋮
Jerami Grant	7	-5.327	-1.389	0.066	-1.323
Josh Smith	7	-5.349	-2.324	0.350	-1.975
Bismack Biyombo	8	-5.601	-1.636	0.448	-1.188
Noah Vonleh	2	-5.622	-1.112	-0.765	-1.876
Doug McDermott	1	-5.740	-0.755	-0.437	-1.192
Dante Cunningham	3	-5.762	0.030	-1.515	-1.485
Dwight Powell	7	-5.865	-1.998	1.274	-0.724
Kyle Singler	1	-6.061	-0.755	-0.774	-1.529
Brandon Rush	1	-6.289	-0.565	-1.485	-2.051
Rashad Vaughn	1	-6.295	-0.837	-0.609	-1.446
Nik Stauskas	1	-6.356	-1.703	-1.825	-3.528
Nerlens Noel	8	-6.643	-3.678	0.549	-3.129
Alonzo Gee	1	-7.151	-1.671	-0.455	-2.126
Richard Jefferson	1	-7.319	-1.859	-0.819	-2.678
Wayne Ellington	1	-7.449	-2.404	0.002	-2.402

Table 4.5: Top fifteen and bottom fifteen players measured by expected point differential per 100 possessions playing with nine replacement-level players.

partially on the prior 2014-15 season. Moreover, the motivation of the player profiles model is to evaluate how the play styles of the members of a lineup interact with each other, so having a low rating in table 4.5 simply indicates that the player would not contribute well to a team of replacement-level players against a team of replacement-level players. However, it is still possible that the same player would be an effective contributor to different lineups or against different lineups.

That said, the results of the model are very interesting. First, the names at the top of the list should be familiar to even the casual NBA fan; this is a good sanity check, as Russell Westbrook, Steph Curry, and LeBron James were widely considered the best three players in the league in 2016, and Curry won MVP for his performance. However, he did not earn the top spot according to the model; instead, Westbrook edged out Curry by about 0.7 points per 100 possessions. An intuitive explanation for this is that, when playing with only replacement players, it is important to be able to score and rebound for oneself without relying on mediocre teammates. When inspecting their player profiles, one finds that Russell Westbrook's top attribute is assist rate, or the percentage of teammate field goals on which he assists while he's on the floor, and his lowest attributes are the percent of his two-point field goals that are assisted and the percent of his three-point field goals that are assisted. In other words, Westbrook is elite at both helping others score and at generating offense on his own; these are the two most important offensive skills when playing with replacement-level

players, whereas Steph Curry, while also elite in both categories, relies slightly more on being assisted on his three-point shots. This can likely be attributed to the Warriors' superior lineup and their style of play, which emphasizes ball movement.

Overall, the top of the list is littered with players from the fourth and sixth clusters. As mentioned previously, these are the clusters that emphasize facilitating on offense; in fact, these are the only clusters that had means with above-average assist rates. This suggests that, when playing with replacement-level players, having an elite point guard is critical, presumably because they are able to elevate the level of their teammates. Indeed, inspecting these players' player profiles shows that many of them have some of the lowest rates of being assisted on two-point field goals in the league; this typically means that they are able to score on their own and that they deliver a large portion of their team's assists. The only players in the top fifteen who are not in one of the two clusters are Kevin Durant, who is perhaps the best pure scorer in the NBA, Draymond Green, whose combination of offensive and defensive prowess is virtually unmatched, and Klay Thompson, who is one of the best pure shooters to ever play in the NBA.

On the other hand, seven of the bottom eight players belong in the "three-point specialist" category. As mentioned above, these players tend to require assists to set up their shots. When playing with replacement-level players who tend to be bad at facilitating, the primary strength of a three-point specialist is significantly weakened. As a result, these players falter

horribly when playing with only replacement-level players.

4.5 Lineup Ratings

Just as players are evaluated by comparing them to replacement level, lineups can be evaluated using the model by computing a lineup's expected point differential per possession against a lineup of five replacement players. The 100 lineups with the most possessions played in the 2015-16 season were evaluated, and the results are given in table 4.6. Note again that 2016 replacement players were used as the opposing team.

These results are very reasonable and unsurprising. First, it is unsurprising that the Warriors' so-called "Lineup of Death" featuring Steph Curry, Draymond Green, Klay Thompson, Harrison Barnes, and Andre Iguodala was rated the best lineup versus replacement players; this lineup gained notoriety during the 2015-16 season as it became the Warriors' third most used lineup as it played 173 minutes for a staggering +44.4 net points per 100 possessions, a couple points higher than the player profile model's estimate against replacement-level players[2,20]. After this lineup follow several other Warriors lineups, with a smattering of Thunder, Clippers, and Cavaliers lineups rounding out the top 15. What is more interesting about these ratings, however, is the correlation between successful lineups and lineup composition in terms of cluster types. The majority of the highly rated lineups contain at least one three-point specialist from cluster 1, and in many cases two such special-

Player 1	Player 2	Player 3	Player 4	Player 5	Cluster Types	Total RAPM	Rating
H. Barnes	S. Curry	D. Green	A. Iguodala	K. Thompson	(3, 4, 7, 1, 1)	17.580	42.345
S. Curry	F. Ezeli	D. Green	A. Iguodala	K. Thompson	(4, 8, 7, 1, 1)	17.922	40.757
H. Barnes	S. Curry	F. Ezeli	D. Green	K. Thompson	(3, 4, 8, 7, 1)	15.597	39.483
A. Bogut	S. Curry	D. Green	A. Iguodala	K. Thompson	(8, 4, 7, 1, 1)	17.154	39.105
H. Barnes	A. Bogut	S. Curry	D. Green	K. Thompson	(3, 8, 4, 7, 1)	14.829	36.398
K. Irving	L. James	K. Love	J. Smith	T. Thompson	(6, 4, 5, 1, 8)	10.743	36.383
S. Adams	K. Durant	S. Ibaka	D. Waiters	R. Westbrook	(8, 3, 3, 6, 4)	10.966	36.260
B. Griffin	D. Jordan	L. Mbah	C. Paul	J. Redick	(3, 8, 7, 4, 1)	11.361	35.205
M. Dellavedova	L. James	K. Love	J. Smith	T. Thompson	(1, 4, 5, 1, 8)	12.355	34.864
A. Bogut	S. Curry	D. Green	B. Rush	K. Thompson	(8, 4, 7, 1, 1)	12.240	34.570
K. Irving	L. James	K. Love	T. Mozgov	J. Smith	(6, 4, 5, 2, 1)	8.069	34.159
S. Adams	K. Durant	S. Ibaka	A. Roberson	R. Westbrook	(8, 3, 3, 7, 4)	12.504	33.594
B. Griffin	D. Jordan	C. Paul	J. Redick	L. Stephenson	(3, 8, 4, 1, 3)	8.417	33.593
J. Crawford	W. Johnson	D. Jordan	C. Paul	J. Redick	(6, 1, 8, 4, 1)	8.761	33.587
D. Jordan	L. Mbah	C. Paul	P. Pierce	J. Redick	(8, 7, 4, 1, 1)	9.481	33.092
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
G. Antetokounmpo	O. Mayo	K. Middleton	J. Parker	M. Plumlee	(3, 1, 4, 3, 2)	-0.931	17.845
E. Bledsoe	T. Chandler	B. Knight	M. Morris	P. Tucker	(4, 8, 6, 7, 7)	-3.743	17.399
W. Ellington	J. Johnson	B. Lopez	D. Sloan	T. Young	(1, 6, 2, 6, 5)	-3.988	16.835
N. Bjelica	G. Dieng	Z. LaVine	K. Martin	S. Muhammad	(7, 5, 6, 6, 3)	-5.712	16.317
G. Dieng	T. Prince	R. Rubio	K. Towns	A. Wiggins	(5, 1, 4, 5, 6)	-0.746	15.583
A. Baynes	S. Blake	S. Johnson	M. Morris	A. Tolliver	(2, 6, 3, 1, 1)	0.790	15.542
K. Garnett	T. Prince	R. Rubio	K. Towns	A. Wiggins	(7, 1, 4, 5, 6)	1.791	15.439
K. Bryant	J. Clarkson	R. Hibbert	J. Randle	L. Williams	(6, 6, 2, 5, 6)	-11.552	14.621
A. Bradley	J. Jerebko	K. Olynyk	M. Smart	E. Turner	(1, 7, 7, 4, 4)	1.369	14.376
D. Booker	T. Chandler	B. Knight	A. Len	P. Tucker	(6, 8, 6, 8, 7)	-5.648	13.933
C. Aldrich	J. Crawford	W. Johnson	P. Prigioni	A. Rivers	(8, 6, 1, 4, 6)	-3.865	13.794
B. Bogdanovic	W. Ellington	B. Lopez	D. Sloan	T. Young	(1, 1, 2, 6, 5)	-6.243	13.633
K. Bryant	J. Clarkson	R. Hibbert	L. Nance	L. Williams	(6, 6, 2, 7, 6)	-9.438	13.532
D. Booker	T. Chandler	A. Len	R. Price	P. Tucker	(6, 8, 8, 4, 7)	-3.679	12.691
K. Bryant	J. Clarkson	R. Hibbert	J. Randle	D. Russell	(6, 6, 2, 5, 6)	-10.898	12.536

Table 4.6: Top fifteen and bottom fifteen lineups by expected point differential against replacement players, out of the 100 lineups that played the most possessions in 2015-16.

ists. The only lineups that do not contain a three-point specialist are the two Thunder lineups; however, Kevin Durant is an exception in cluster 3 in that he is an excellent three-point shooter, despite the cluster mean. This demonstrates the importance of having players who are threats from beyond the 3-point arc.

It is also interesting to note that the total RAPM of a lineup is not completely correlated with the lineup's points above replacement rating. For example, the second Thunder lineup has a far lower total RAPM than the Cavaliers lineup featuring Mozgov, but the Mozgov lineup is ranked higher than the Thunder lineup because, according to the model, it is able to take better advantage of complementary skill sets.

On the other side of the table are several Lakers lineups featuring a 37-year-old Kobe Bryant; these lineups are rated very poorly and do not have any three-point specialists on them. Also noteworthy is the number of Timberwolves lineups in the bottom 15 lineups; in particular, the lineup of Rubio, Wiggins, Prince, Garnett, and Towns has a positive total RAPM, almost 10 more points per 100 possessions than the get moreworst Lakers lineup by RAPM, and yet they are projected to perform quite poorly against replacement players, relative to the other teams.

4.6 Matchups Between Starting Lineups

Another useful application of the model is that we can test how well we would expect each team's starting lineup to perform offensively and defensively against every other team's starting lineup from the 2015-16 season. In order to determine a team's starting lineup, I selected whichever lineup started the most games for a team during the season. The expected points per possession for each possible matchup is given in figure 4.2, where the rows indicate the team on offense and the columns indicate the team on defense; the values in the table represent the how many points per 100 possessions the offensive lineup would be expected to score against the defensive lineup. Note that values on the diagonal represent how well a given team's starting lineup would be expected to fare against itself.

Further, for any two teams, we can compute the difference between the offensive advantages each team was expected to have against the other based on figure 4.2. This gives an overall expected point differential per possession for each matchup; these overall expected point differentials per possession, prorated to 100 possessions, are given in figure 4.3. Note in this figure that 104.5 points per 100 possessions is roughly average in the NBA, so the colorbar is centered at 104.5, with green representing above average and red representing below average.

These figures by and large make sense; we can see that teams like Golden State, Oklahoma City, and Cleveland are very strong, whereas teams like Philadelphia, New York, and

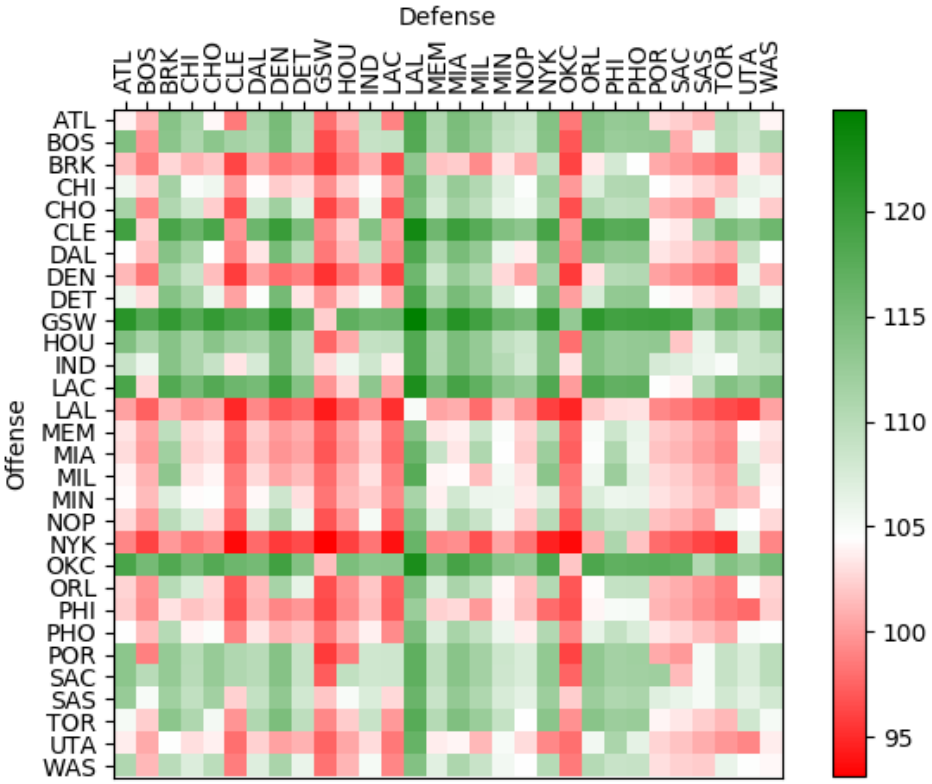


Figure 4.2: How different teams' starting lineups would be expected to perform against one another on a given side of the ball.

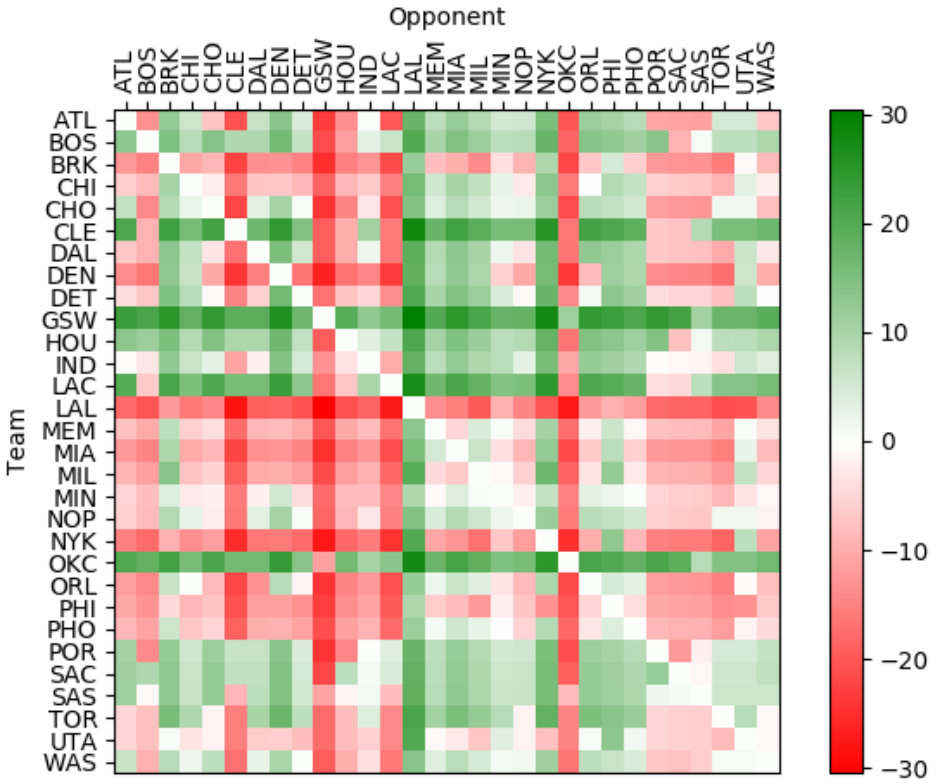


Figure 4.3: How different teams' starting lineups would be expected to perform against one another, computing the point differential between the offensive and defensive points.

the Los Angeles Lakers are weaker. It is interesting to note that because of the way that different play styles interact, these values are not transitive; in other words, if team A has a positive expected point differential against team B and team B has a positive expected differential against team C, it does not necessarily mean that team A would play well against team C. For example, Boston has a very poor expected point differential against Oklahoma City, whereas Indiana is a better matchup with Oklahoma City. However, the model favors Boston's starting lineup against Indiana's. Intuitively, it makes sense that a situation like this could arise in the NBA due to different matchups and play styles within the lineups, so it is good that the model agrees that situations like this can arise.

5

Conclusion

As shown in chapter 4, the model trained on player profile data provided significantly better out-of-sample prediction accuracy compared to the baseline RAPM model. This provides evidence for the hypothesis that chemistry and complementary play styles are critical in order to be successful in the NBA. Indeed, while RAPM certainly has its use

cases, such as when attempting to break down team outcomes and distribute responsibility to the individual level, the success of the model herein suggests that there is more to building a strong lineup than simply maximizing the squad's combined RAPM ratings.

The primary implication of this work is that it indicates that NBA coaches and executives should dedicate more resources to understanding the way in which players' play styles and abilities interact with one another from a statistical perspective. Often, the focus in the NBA is solely on acquiring star talent; instead, teams should make it a first-class goal to build a roster that allows the coach to play players with complementary play styles. As seen in section 4.5, some lineups are able to overachieve their RAPM expectations due to the way that their play styles interact; this fact could be advantageous for a team who wishes to stay competitive while maintaining enough available cap space to sign high-impact players in a coming offseason for example. The principle of getting as much out of one's players as possible is especially crucial when a coach has to turn to his bench during a game. With twelve players available to play, there are 792 possible lineups a coach can choose from during a game; choosing the optimal lineup based on his players' play styles and on the lineup selected by the opposing team's coach without taking a principled, analytical approach would thus be nearly impossible. The results presented here motivate an increase in the degree to which basketball analytics are used to make in-game decisions. We have seen similar movements in other sports; for example, Microsoft has partnered with the National

Football League, where players and coaches use Microsft Surface tablets to employ analytics software in real-time[17]. A similar movement in professional basketball could allow coaches to utilize models such as the one presented here in order to evaluate substitution decisions.

There is also room for extensions of the research presented here. One possible extension would be to add additional features to the regression model. For example, one drawback of the model is that it assumes that the point outcome is only a function of the players on the court; additional features could include terms measuring fatigue of a player, such as the number of possessions that he has been on the court. The model could also incorporate player-tracking data, which the NBA began collecting at the beginning of the 2013-14 NBA season; this data could be incorporated into the player profiles and into the regression model directly. In particular, the player-tracking data could increase the signal in the player profiles; for example, one could quantify how a player moves without the ball or how often a player is involved in different types of plays, such as pick-and-rolls.

Another extension of the research would be to analyze complementary play styles within rosters. While the research here focuses on evaluating lineups, NBA executives need to make decisions on a season level about which players should be acquired and which should not; these decisions come in many forms, as players can be acquired through trades, free agency, or the NBA draft. In constructing a roster, the goal is to acquire talented players while giving the coach more flexibility in his lineup and substitution decisions; therefore,

the model presented in this thesis could be extended to determine the cohesion of a roster's play styles. Similarly, this model could be used to analyze whether a trade would have a positive impact on a team by determining how the new player or players would complement the players that remain on the roster after the trade.

While the conclusion that interactions between players' skill sets make a significant impact on the outcome of possessions and games in the NBA is not remarkable in itself, it is an important fact to consider when thinking about or watching basketball. Indeed, the results of this research confirm systematically what many NBA fans, analysts, and teams have assumed intuitively since the game was invented: there is more to a lineup's value than the simple sum of its parts.

A

Code

Code related to scraping and cleaning the data can be found in the Python package I have authored for easily scraping sports data; the library is open source and its code can be found at `https://github.com/mdgoldberg/sportsref`.

Code specific to this thesis has also been maintained on GitHub and can be found at `https://github.com/mdgoldberg/nba_lineup_evaluation`.

References

[1] Arcidiacono, P., Kinsler, J., & Price, J. (2015). Productivity spillovers in team production: Evidence from professional basketball. http://public.econ.duke.edu/~psarcidi/multiable.pdf.

[2] Basketball Reference (2017). Basketball reference. http://www.basketball-reference.com.

[3] Beech, R. (2003). Nba roland ratings. http://www.82games.com/rolandratings.htm.

[4] Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

[5] Breiman, L. (2001). Random forests. *Machine Learning*.

[6] Fearnhead, P. & Taylor, B. (2011). On estimating the ability of nba players. *Journal of Quantitative Analysis in Sports*.

[7] Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*.

[8] Goldberg, M. (2017). Sportsref. `https://github.com/mdgoldberg/sportsref`.

[9] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*.

[10] Ilardi, S. & Barzilai, A. (2008). Adjusted plus-minus ratings: New and improved for 2007-2008. `http://www.82games.com/ilardi2.htm`.

[11] Kitanidis, P. & Bras, R. (1979). Collinearity and stability in the estimation of rainfall-runoff model parameters. *Journal of Hydrology*.

[12] Kuehn, J. (2016). Accounting for complementary skill sets when evaluating nba players' values to a specific team. *Sloan Sports Analytics Conference*.

[13] Maymin, A., Maymin, P., & Shen, E. (2013). Nba chemistry: Positive and negative synergies in basketball. *International Journal of Computer Science in Sport*.

[14] Rosenbaum, D. (2004). Measuring how nba players help their teams win. `http://www.82games.com/comm30.htm`.

[15] Roweis, S. T. & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*.

[16] Sill, J. (2010). Improved nba adjusted +/- using regularization and out-of-sample testing. *Sloan Sports Analytics Conference*.

[17] Soper, T. (2016). Microsoft happy with its nfl partnership thus far as surface tablet becomes mainstay on sidelines. http://www.geekwire.com/2016/microsoft-happy-nfl-partnership-thus-far-surface-tablet-becomes-mainstay-sidel

[18] Sporting Charts (2015). Basketball plus-minus. https://www.sportingcharts.com/dictionary/nba/basketball-plus-minus.aspx.

[19] Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*.

[20] Tjarks, J. (2016). The nba's new lineups of death. https://theringer.com/the-nbas-new-lineups-of-death-7e52303954a1.