WAR in Pieces: A Bottom-Up Approach to Player Evaluation in the NBA

Kenny Huang

Advisor: Ramon van Handel

Submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Science in Engineering
Department of Operations Research and Financial Engineering
Princeton University

May 2023

I hereby declare that I am the sole author of this thesis.

I authorize Princeton University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Kenny Huang

I further authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Kenny Huang

Abstract

Basketball is a dynamic sport that requires players to excel in a wide range of skills, including scoring, rebounding, passing, and defense. Traditionally, players have been evaluated based on simple metrics such as points, rebounds, and assists, which fail to capture the full complexity of their performance on the court. Moreover, these metrics do not account for the many contextual factors that can influence a player's performance, such as the quality of their opponents, the pace of play, and the role they play on their team. In recent years, a new generation of metrics have been created to address these limitations. These combine simple metrics with previously untracked aspects of the game to provide a more comprehensive and accurate assessment of a player's performance. Even so, all are flawed in some way, particularly in their failure to meaningfully capture player-player interactions.

In this paper, I create a new metric called Wins Above Replacement (WAR) that seeks to address those shortcomings and provide a more nuanced and reliable evaluation of player performance. The key idea is to create predictions at the lineup level and to rebuild the season from the ground up. I describe the full methodology used to calculate WAR, which is inspired in part by the baseball metric of the same name, and I present the results of its application to NBA players in the past six seasons. The results generally agree with preexisting evaluation metrics while also raising questions about new ways to think about evaluating players. Overall, this work represents a significant advance in the field of basketball player evaluation and has the potential to revolutionize the way that players are scouted, evaluated, and developed.

Acknowledgements

I'd like to take this opportunity to acknowledge all who made this thesis possible. First, I want to thank Professor Ramon van Handel. I appreciated you so much when you taught me probability many springs ago, and I am even more grateful now for your honest guidance and willingness to listen as an advisor. I would also like to acknowledge the ORFE department and the University for providing the means and opportunities to learn and grow both as a student and citizen these past four years.

To my best friend Howard who has had to suffer my stupid jokes since high school; to my underperforming thesis fairy and passenger princess Kathy; to my traveling salesman Jamie; to the rest of Stonks — Andrew, Alex, and Autumn — who have been at my side every step of the way; and to my sister Jenny who visited two days before the due date and reminded me of what's really important in life: y'all are my rocks, and I truly appreciate you so much.

Throughout my time at Princeton, I've encountered so many other peers that I now proudly call my friends. Neo and Josh, you have been with me since even before Day 0, and our shared experiences here have only strengthened that bond. Brian, your taste in sports teams is disappointing but nevertheless I love all our banter and jest, and I look forward to many more decades of Knicks and Cowboys mediocrity. To Jessica, Reha, and Callie, thank you for accepting me into your lovely abode; I'll always cherish my memories of birthday party decorations and baguettes of various sizes. And to all who I've crossed paths with these years at Princeton — Albert, Andrew A., Andrew T., Byrone, Ellie, Eric, John, OG, and everyone else — I will forever be grateful for all our shared laughs and golden memories.

Finally, I want to thank my parents, Jiaming and Haichen. There is nothing I can ever do to repay you for twenty-one years of unconditional love and support, but I will never stop striving to do so. Mom and Dad, I hope I have been a satisfactory son so far, and I will always love you.

To the people I love

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Basketball is a complex game. First developed by physical education professor James Naismith in 1891 as a way of keeping his students preoccupied during rainy days, the sport has evolved into a multi-billion dollar mainstay in the entertainment industry [12]. Its modest equipment and personnel requirements as well as attractiveness as an indoor sport during cold winter seasons fueled basketball's growing popularity at the college and high school levels in the early 20th century. By the 1920s, there were hundreds of professional local teams playing in armories and dance halls across the country, including the Harlem Globetrotters who still travel and perform today. In 1946, Walter Brown, owner of the Boston Garden arena ironically located in New York City, shrewdly realized that ice hockey arenas around the country could be used to host organized basketball games in front of crowds on an unprecedented scale. He established the Basketball Association of America, which merged with the National Basketball League in 1949 and the American Basketball Association in 1976 to create the National Basketball Association, or NBA as we know it now [15]. Today, the NBA is the top professional basketball league in the world in terms of popularity, salaries, talent, and level of competition.

Even in the early days, teams were always looking for ways to improve their performance on the court. Advances in sports science and medicine allowed players to become faster, stronger, and less susceptible to injury, and as financial incentives grew, more and more players trained for basketball as their primary occupation. Beginning in the 1970s, statisticians like Bill James introduced a quantitative, data-driven aspect to sports by applying statistical methods and empirical analysis to baseball, birthing what has become known as “sports analytics”. Since then, the sports analytics industry has grown dramatically; a study published by Grand View Research Inc states that the global sports analytics market size will expand at an annual rate of 31.2% and reach $4.6 billion by 2025 [8]. Within the NBA, every team now has some form of a department that focuses exclusively on quantitative analysis, as do most media outlets covering sports events.

One major result of this revolution is the proliferation of statistical evaluation of players into all corners of the sports world. Early on, teams naturally experimented with objective forms of evaluation to aid in decision-making, such as selecting which players to build around and which to let go. Aside from teams, however, evaluation metrics are increasingly accepted by others as well. For awards voters, they provide a supplementary tool when forming their opinions to support or rebut their conclusions from the eye test. For fans, they serve as fuel in a never-ending struggle to promote the greatness of their favorite players. Even for players, awards like MVP significantly impact their perceived value when negotiating salaries, and honors like All-NBA selections even have serious legal implications on financial matters, such as the maximum value they can earn with a new contract as allowed by the Collective Bargaining Agreement. For some, a single additional honor can make a difference of tens of millions of dollars per contract.

As such, statisticians have developed a multitude of different metrics in recent years to quantitatively compare players. Each has its advantages and disadvantages, which I will explore in the literature review section. My work starts with the statistic of Wins Above Replacement (WAR) in baseball. In simplest terms, WAR is a singular value that estimates how many wins a player contributes to his team compared to if he was replaced by a player at his same position at a skill level that could be added to the team at any time [10]. My goal is to apply this way of thinking to basketball and compute a value that numerically describes how each NBA player affects the most important part of the sport: winning.

Of course, some modifications must be made. Baseball is inherently relatively not reliant on player-player interactions, so analysts can more easily quantify individual players' contributions. However, as Professor Bill Gerard states, what he deems as "invasion team sports" like basketball are "much more complex and hence the separability of individual player contributions is considerably more difficult" [5]. One primary focus of my work is thus to account for these interactions as much as possible.

The remainder of this thesis can be summarized as follows: first, I will identify the difficulties and areas of improvement of evaluating players using previous literature and methods. Then, I will apply machine learning theory and experimentation to create a model that accurately predicts the number of games that a team will win using only player-related inputs, such as individual box score or advanced tracking statistics. Using this model, I will compute each player's WAR value by substituting a replacement-level player and taking the difference in predicted wins. Finally, I will analyze the computed WAR values in conjunction with preexisting evaluation metrics to ascertain the validity of this methodology.

Chapter 2

Literature Review

2.1 Peer Effects

One of the most central aspects and largest challenges when analyzing invasion sports is the interdependence between players. In baseball, offenses feature only one player at a time, and defensive plays require minimal coordination and rely mostly on individual execution. In basketball, all players on the court are involved at any time, whether directly or indirectly. On offense, for instance, only one player is in possession of the ball (the “ball-handler”), but the other four interact either directly with the ball-handler by running interference with the defense or indirectly by constantly adjusting their position given the locations and movements of the other players.

For instance, a common play in the modern NBA is the pick-and-roll. The ball-handler stands outside the three-point line, and the “screener”, usually a forward, approaches and stands next to the player defending the ball-handler. The ball-handler can then dribble around the screener, who by just standing still impedes the defender from staying close to the ball-handler. This is referred to as a “screen”. The ball-handler can now choose to dribble hard towards the basket, referred to as a “drive”; if the screener’s original defender is not able to contain him, a third defender may come and help as a “help defender”. This now leaves a third offensive player alone, whose responsibility now is to identify and move to a place on the court where the ball-handler can find and pass to him if needed. Overall, the pick-and-roll appears to be a surprisingly complex action, but the process described is considered a basic fundamental in the NBA and highlights the intricate interactions that can exist between offensive players. Among defensive players, the interactions are inherently even more complex. In recent years, switching schemes have gained popularity, where the screener’s defender is now responsible for guarding the ball-handler. As a result, teams increasingly emphasize communication between defenders and value agile, “switchable” big men who are able to defend against fast, smaller players.

Thus, when analyzing the effect of an individual player on the team’s performance, the influence of teammates must be included in some way. If not, omitted variable bias is introduced since the player’s teammates are relevant to his actions. Previous work in other sports have made corrections for this bias, such as when Thomas et. al [4] and Yurko et. al [17] sought to compute WAR equivalents in hockey and football, respectively. For instance, Thomas approaches this issue by modelling the scoring rate as a semi-Markov process and designates hazard functions that depend on the players on the ice. This Markov approach is not necessarily relevant for basketball, but the underlying reasoning behind its use is just as valid. Rein and Memmert add that the lack of contextual information of box scores limits them to have only “weak explanatory power for team outcomes” [9]. Thus, it is crucial in this work to incorporate other datasets in ways that account for interactions between teammates.

2.2 Existing Player Evaluation Metrics

To better understand what my work should prioritize, I first examined the ways in which analysts have tried evaluating players thus far. In sports, a large variety of actions (referred to as “statistics”) are tracked each game, such as points, rebounds, and assists. However, when one compares two players using only their statistics, there is no clear preference relation she can use to rank them. For instance, how do you compare a player averaging 20 points, 15 rebounds, and 10 assists with one averaging 30 points, 10 rebounds, and 5 assists? To attempt to answer this question, many analysts have created single catch-all metrics of a player’s contribution to their team. Currently, the most relevant single-value metrics to my work are Adjusted Plus-Minus, Box Plus-Minus, VORP, Win Shares, Wins Above Replacement Player, and RAPTOR.

2.2.1 Adjusted Plus-Minus

Adjusted Plus-Minus (APM), created by Wayne Winston and Jeff Sagarin, is among the most cited attempts to evaluate players [1]. Adjusted Plus-Minus is based directly on the differences between a team’s offensive and defensive ratings when a player is on the court compared to when he is off the court, hence the name “plus-minus”, and in simplest terms measures a player’s contribution in points above league average per 100 possessions. It does so by taking a high-level view and only caring about the scoring margin, so in effect it is able to indirectly account for anything that happens on the court that affects the scoring output of a team, including the elusive player-player interactions. In addition, Winston and Sagarin applied a linear regression model to account for contextual factors like opponent difficulty and teammate strength. As a result, it is considered among the most comprehensive metrics available and is particularly good at evaluating players over the course of multiple seasons.

However, there are still some limitations of APM. Primarily, APM struggles with small sample sizes, so for applications towards single seasons, it often fails the eye test and mistakes strong role players for high impact stars. Joe Sill created **Regularized Adjusted Plus-Minus (RAPM)** which uses the ridge regression instead of simple linear regression in order to create more predictive results even when using smaller sample sizes of data [3]. In terms of predicting long-term success, RAPM is considered the gold standard. On the other hand, ESPN's **Real Plus-Minus (RPM)** introduces a Bayesian statistical prior that leverages a prediction of box score values to better inform the model on what type a player is [7]. Finally, Basketball Index, a team of basketball consultants and analysts, combines the best of both with LEBRON, short for **"Luck-adjusted player Estimate using a Box prior Regularized ON-off"**. LEBRON utilizes RAPM data in addition to player tracking and box score metrics to gain the advantages of both RAPM and RPM [6].

One notable trait that all of of these plus-minus-based metrics share is the fact that they operate independently from how the game is played; they treat the game as a black box and focus only on the results that come out. This is not necessarily a bad thing, as it can eliminate any biases coming from preconceived notions about what the designer expects to see. However, it also reduces their intuitiveness and explaining power if one were to dig deeper and ask any question besides the scoring output. This is a limitation that I plan to address in my work.

2.2.2 Box Plus-Minus

A very similar-sounding but fundamentally different metric is Box Plus-Minus (BPM), developed by Daniel Myers at Basketball Reference. Instead of plus-minus values, BPM uses straightforward statistics like points, rebounds, and assists and regresses them to estimate RAPM while making adjustments relating to the team, position, and other factors. As such, it is also designed to reflect a player's contribution in points above league average per 100 possessions, so it is in a way a plus-minus statistic too.

The primary advantage of BPM is that it uses only basic statistics that have been tracked since the 1973 season, which allows one to evaluate players from before the more granular data needed for APM and other metrics were tracked. However, BPM still does not sufficiently address the previously discussed black-box limitation. It does account for some contextual information like player position, but otherwise the data is still treated as a sterile optimization problem.

2.2.3 Value Over Replacement Player

It would be remiss not to include Value Over Replacement Player in our discussion, as the name suggests that it is very similar to my methodology. In reality, VORP is a direct byproduct of BPM and describes the net rating that a player provides compared to a replacement-value player determined to have a BPM of -2.0. In other words, VORP is simply:

$$\text{VORP} = [\text{BPM} - (-2.0)] \times (\text{\% of possessions played}) \times (\text{team games}/82)$$

According to Basketball Reference, VORP can also be used to estimate Wins Above Replacement by multiplying by 2.7, the conversion rate of points-for-wins at league-average level in the Pythagorean formula of sports, which I will discuss later in the Methodology section [11].

2.2.4 Win Shares

Another popular evaluation metric is Win Shares (WS), which works from the top down: each team is given 3 Win Shares per win in a given season, and those Win Shares are distributed to the players on the team according to their offensive and defensive ratings relative to the rest of the team [2]. Simplifying a little, let's say an individual player performed δx_i amount better than league average. Then, that player's Win Shares value would roughly be $WS \approx \frac{\delta x_i}{\sum \delta x_i}(3w)$, if the team as a whole won w games this year. By only incorporating the high-level results of scoring margin, Win Shares can also be considered a plus-minus statistic. As designed, the sum of Win Shares for the players on a team is very strongly correlated with how many games that team wins, which makes sense objectively but also means that Win Shares suffers from the same black-box issues of the other plus-minus based metrics. In addition, it disproportionately inflates role players' contributions on good teams and deflates good players stuck on bad teams.

2.2.5 Wins Above Replacement Player

Wins Above Replacement Player is not a particularly well-known statistic, but its creator Kevin Pelton appears to have the most similar motivation to me of trying to replicate WAR in baseball [16]. However, his methodology replicates that of WAR much more closely than mine. In simplest terms, WAR in baseball is computed by assigning weights to each of the actions that affect the final score. Similarly, WARP first decomposes the game into offensive and defensive parts. Pelton computes the number of offensive points produced from scoring or assists and defensive points prevented from steals, blocks, and defensive rebounds, and he uses those to estimate the offensive and defensive ratings for a team consisting only of this player and four average teammates. From here, he applies the Pythagorean formula of sports to obtain an expected win percentage. He now defines a replacement level player by stipulating that a team of five replacement level players will win 10 games, a number he derives from empirical evidence from his previous work. By comparing the computed win percentage with the win percentage from a replacement level player and four average teammates, Pelton arrives at his WAR value.

The depth with which this method applies information about how the game itself works compares starkly to the methods examined thus far. My explanation is but a highly abridged version of all the math that Pelton does, but his work truly tries to incorporate and account for all the actions that can occur in a basketball game. However, one important aspect that is lacking is the context of teammates for each player. By computing the player's win percentage using four average teammates, he sets a baseline for the comparison with the replacement player but removes all notions of player-player interactions that the previous methods at least indirectly include. Moreover, Pelton meticulously describes his methodology, but there is no actual database of computed values. I feel that creating such a database is outside the scope of my work, so I am unfortunately not able to compare my results to his.

2.2.6 RAPTOR

Finally, we can examine the "Robust Algorithm (using) Player Tracking (and) On/Off Ratings" (RAPTOR), which is one of the most widely-cited metrics available. RAPTOR was developed by FiveThirtyEight, an entity that among other things performs rigorous statistical analysis in the realm of sports. Their RAPTOR model involves two main components: a "box" score component, which includes player tracking and play-by-play data, and an "on-off" component, which follows a similar line of logic as RAPM of comparing how well a team plays with and without a given player [20]. However, an improvement upon the plus-minus metrics mentioned earlier is that the "box" component incorporates an impressive variety of facets of the game, including passing and rebounding measures as well as spacing, fast breaks, and fouls drawn. This echoes a goal of my work, to use more tangible and visible non-plus-minus statistics in the inference process. Similarly to VORP, RAPTOR can also be used to estimate WAR by assuming a replacement level of -2.75. The formula takes into account minutes played and even how much each player impacts the pace of his team:

$$\begin{aligned}\text{WAR} &= (RAPTOR + 2.75) \times \text{MinutesPlayed} \\ &\quad \times ((\text{LeaguePace} + \text{IndividualPaceImpact})/\text{LeaguePace}) \times \text{WARmultiplier}\end{aligned}$$

That being said, RAPTOR still has its limitations. The creators themselves acutely state in their description of RAPTOR works:

> “Namely, these statistics assume that player performance is largely linear and additive, that is, that you can roughly add up the ratings from individual players to project team performance. In other words, RAPTOR does not account for coaching, systems or synergies between teammates.”

Another contentious property is that RAPTOR is a patented methodology whose exact details are known only to FiveThirtyEight. It is impossible to verify its validity, and thus their conclusions must be taken with a grain of salt.

2.2.7 Summary

In summary, the player evaluation metrics mentioned above are the best available, but each has its own limitations: plus-minus-based statistics like BPM and APM as well as Win Shares are black boxes that don’t take into account any characteristic aspects of basketball; WARP does so but is unable to account for any peer effects; and RAPTOR uses some game knowledge but only indirectly accounts for player-player interactions and is difficult to verify. As such, one way to address this gap in the literature is to create a methodology that applies game knowledge in an intuitive and clear way that also incorporates player-player interactions.

Chapter 3

Methodology

As stated in the introduction, I wish to create a model that predicts the number of games that a team will win using only player-related inputs, such as box score or advanced tracking statistics. However, the literature review reveals that this model must be more than just plugging in players into a linear regression due to the entanglement of player-player interactions. How I approach this problem is by shifting my focus from the players on the roster every season to those on the court every minute. Simply put, my goal is now to create a model that predicts how well a group of five players plays together on a per-minute basis. Using data on how long each 5-man lineup played during a given season, I can reverse engineer the season from scratch to predict how many games that team won. At first glance, however, it appears difficult to quantify exactly how much a lineup playing 50 minutes together will affect a season lasting 4000 minutes in total.

For this, I will introduce the Pythagorean formula, one of the fundamental formulas in sports analytics incorporated in multiple preexisting methodologies. The formula states that that the winning percentage of a given team is related to the offensive and defensive outputs in the following way:

$$\text{win\%} = \frac{(\text{offensive input})^a}{(\text{offensive input})^a + (\text{defensive input})^a}$$

where a is an intrinsic property of each sport. In basketball, a is empirically found to be around 16, and the inputs used are the offensive and defensive ratings, rate statistics describing the expected number of points scored by the team and their opponents per 100 possessions, respectively. For example, if on average I score 120 points per 100 possessions and my opponent 110 points, my expected win percentage would be

$$\text{win\%} = \frac{(120)^{16}}{(120)^{16} + (110)^{16}} = 0.808$$

or about 81%. Empirically, I found that this applies well with an error of about 0.04, so I can use this tried-and-true method to break down wins into pieces that are easier to work with. Specifically, instead of finding how a given lineup playing 50 minutes together will affect the number of wins in a season, I try to quantify how that lineup playing 50 minutes together will affect their average offensive and defensive ratings for the season. If the lineup is strong offensively, then it will improve the overall offensive rating of the team and in theory correspond to a larger win percentage prediction from the Pythagorean formula; a similar logic applies for defense. This illuminates my path forward: create models to predict the offensive and defensive ratings of any lineup of five players and use the predictions with how often those lineups played to recreate the season. From this, I get the overall offensive and defensive ratings for a given team in a given season, which I can plug into the Pythagorean formula to yield the final win percentage prediction.

3.1 Limitations

One major limitation that I can foresee is that this methodology is very era-dependent. The game of basketball and how it is played are not stationary, so the methodology will invariably lose predictive power as time passes and the game evolves. As a clear example, the three-point revolution in recent years led to an increased emphasis on three-point field goals and floor spacing [19], drastically reducing the value of dominant centers prevalent during the early 21st century like Shaquille O'Neal. "Stretch" forwards who are skilled shooters and can "stretch" the defense, such as Kristaps Porzingis, gained popularity in their wake. For this reason, I will be limiting the data used to the six most recent full seasons, those between 2016-17 and 2021-22.

Another assumption that is made in this methodology is that the players themselves are static and do not adjust their inherent characteristics or approach when with different teammates. This ties into an issue of a lack of high-level context. I have no access to contextual information about when these lineups play, such as whether they are against the other team's best players, in "garbage time" when the game is out of reach and there are no stakes at all, or in "clutch" situations when the game is close with less than five minutes to play and every point matters exponentially more. This lack of contextual information forces me to treat each minute equally, which is a very strong assumption that can impair the methodology's performance.

Finally, a point worth mentioning is that incorporating the Pythagorean formula introduces an additional source of noise. My reliance on the Pythagorean formula to compute wins means that I cannot avoid the aforementioned error of 0.04. For consistency, all errors mentioned will be the actual errors, but the standard to reach is now 0.04, not 0.

Chapter 4

Data

For decades, all notational analysis was based on box scores, which tracks a number of different "statistics" for each game and presents each player's and team's performance in a tabular form. The earliest form of box scores with the formation of the BAA in 1946 only counted the bare minimum: Field Goals, Field Goal Attempts, Free Throws, Free Throw Attempts, Assists, Personal Fouls, and Points. After the NBA-ABA merger in 1976, this was expanded to include Steals, Blocks, Turnovers, and Offensive and Defensive Rebounds, and in 1979 the 3-point shot was introduced and rounds out the statistics tracked in the box score as we know it today [14]. Explanations of the relevant statistics will be included later for context. However, even though box scores can reveal much of a player's involvement, it is unable to fully capture the context in the game. For instance, there are many in-game actions that are not recorded in box scores but contribute to a winning team, such as setting screens and boxing out after field goals. Many players excel in these areas, but an outsider who only has access to the box score would not be aware of their contributions.

In recent years, an emphasis on data-driven analysis, augmented by innovations in video tracking and database management, has led to an explosion in the scope of data collection beyond traditional box scores. Today, thousands of independent companies actively track, analyze, and store previously unseen aspects of the game, providing analysts an unprecedentedly fine-grain view of exactly what is happening on the court. The NBA itself even partners with data providers like SAP to release publicly-available data that serves as an important basis for my work. The data used are drawn from the following datasets.

4.1 Box Scores

Despite their limitations, box scores are a valuable place to start as they contain information about the most visible aspects of the game. In particular, I used the season summary box score from Basketball Reference, a publicly-accessible and reliable database. These summaries contain per-game averages for each NBA player for a given season and track a variety of statistics, including but not limited to:

- GP - Games Played
- MIN - Minutes Played
- 2PA - 2-Point Field Goal Attempts - shots taken from inside the three-point arc not including those missed due to personal fouls (see PF)
- 2P% - 2-Point Field Goal Percentage - made two-point field goal attempts divided by total two-point field goal attempts
- 3PA - 3-Point Field Goal Attempts - shots taken from outside the three-point arc not including those missed due to personal fouls (see PF)
- 3P% - 3-Point Field Goal Percentage - made three-point field goal attempts divided by total three-point field goal attempts
- FTA - Free Throw Attempts - unobstructed shots taken as compensation for fouls by the opposing team (see PF)

- FT% - Free Throw Percentage - made free throw attempts divided by total free throw shots
- ORB - Offensive Rebounds - when a missed field goal or free throw attempt stays in possession of the offense; credited to the player who gains control of the ball
- DRB - Defensive Rebounds - when a missed field goal or free throw attempt becomes in the possession of the defense; credited to the player who gains control of the ball
- AST - Assists - when a player's pass leads directly to a made field goal attempt; credited to the passer
- STL - Steals - when a defender gains possession of the ball before any field goal attempt, usually by catching a pass from the offense or forcing the offense to lose control of the ball in some way
- BLK - Blocks - when a defender deflects a field goal attempt
- TOV - Turnovers - when an offense gives the ball back to the defense without a field goal attempt; examples include stepping out of bounds, steals, and taking too long to shoot
- PF - Personal Fouls - any illegal play, usually by the defense
- PTS - Points - 1 per made free throw attempt, 2 or 3 per made field goal attempt

These statistics are included in my analysis because they are tangible manifestations of how different players tend to act on the court. For instance, tall players, colloquially referred to as "big men", are usually less agile and less accurate at shooting and thus spend much of their time on offense close to the basket. This naturally increases their rate of shooting 2-point field goals relative to 3-point field goals. On defense, they guard other big men who are also close to the basket, which increases the likelihood of getting defensive rebounds and blocks. The implication is that by examining these statistics and their values relative to each other, I can glean insights about the way they behave on the court.

In addition, to compare players with drastically different playing times, I normalized the counting statistics, including points, rebounds, and assists, to be on a per-36 minute basis. In other words, for each player I multiplied their stats by 36 divided by the average number of minutes he played per game to estimate what his production would have been had he played 36 minutes, which is roughly how much an average starter plays.

I'd also like to quickly note that I chose to omit the Made quantity and only use the Attempts and Make Percentage for each shot type. Since Make Percentage hovers within a relatively thin range, Made Shots and Attempted Shots will track each other strongly and the multi-collinearity will not add any information to the model.

4.2 Drive Data

There are some other datasets that I want to incorporate into the player analysis, all of whom are released by the NBA and SAP. To start, one of the best indicators of a player's role in the offense is his drive frequency. Running hard to the basket while dribbling the ball and penetrating the defense is a common catalyst to create movement within the defense. If other defenders come in to help, the ball handler can pass it out to open shooters, and if not, he can make something happen on his own now that he is likely past his defender. Thus, the number of times a player drives (DRIVES) is a strong indication of what his responsibilities are within the offense.

4.3 Passing Data

In the same vein, passing data is also a reliable indicator. Aside from passes coming from drives, players can also generate offense through passing and off-ball movement outside the three-point line, since the ball moves faster than any player can and can punish defenders who are out of position. Once a defense is compromised this way, continuous and thoughtful passing can lead to a scrambling situation which often leads to easy layups or open shots for the offense. In general, teams that frequently pass often are more efficient. Thus, I want to include in my player analysis the number of passes made by each player (PASSES) as well as the points created from assists (PTS_CREATED), which appropriately adjusts the assists statistic for three-point field goals as well.

4.4 Shooting Data

Finally, many players' defining feature is the way that they get their points. Generally, field goal attempts are classified into three types: short-range, pullups, and catch-and-shoot. Short-range attempts, defined as those from ten feet or closer from the basket, are often layups or dunks and usually result from a specific set of technical moves called "post moves". As such, these shots are distinct from the others and thus treated as their own category. On the other hand, attempts from farther than ten feet are almost exclusively jump shots. Here, the success rate of jump shots varies heavily with the context of each shot. For example, I have briefly mentioned the term "open" in previous sections, whose generally accepted definition is when a shooter has no defenders within a 6-foot radius. These attempts have a much higher success rate than contested shots, where the threat of a deflection and obstruction of the view of the basket can significantly alter the shot (for more information, search for the "Kornet Kontest"). In addition, there are two main ways that a player can enter their shooting motion: pullups and catch-and-shoot. In the former, a player is in control of the ball and can "pull up" and shoot whenever he wants; the latter is exactly what it sounds like. However, in the case of the latter, the fact that the player is not in control of the ball until right before the shot occurs introduces a source of disorder that significantly reduces the quality of the shot. In practice, this means that catch-and-shoot attempts are almost never contested, since a shooter in that situation usually elects to pass it to a teammate instead.

In the context of our analysis, the relative rates at which players shoot catch-and-shoot shots compared to other types is highly reflective of how often they have the ball, so I include the percentage of each player's field goal attempts that are classified as catch-and-shoot as FREQ and the actual number of those attempts as CATCH.

4.5 Lineup Data

The most valuable data from the NBA and SAP is a dataset describing different combinations of five players, or lineups. For each season and five-man lineup, they provide the number of minutes that those five players were all on the court together that year as well as a myriad of metrics that describe what happened during that time. Most important among those metrics are the offensive and defensive ratings of each lineup, which gives me the labelled dataset that I need to predict the ratings of each lineup. This dataset, as well as the others from the NBA and SAP, was manually copied from the NBA website, which was a painstaking process as there was no way to easily export the data like there was for Basketball Reference. The relevant columns are:

- LINEUPS - the 5-man lineup (ex. M. Gortat - J. Wall - M. Morris - B. Beal - O. Porter Jr.)
- TEAM, SEASON

- MIN - number of minutes played together
- OFFRTG - offensive rating
- DEFRTG - defensive rating

4.6 Standings Data

In order to confirm how well the predicted offensive and defensive ratings correlate to actual win counts, I needed the true values of how many games each team won. This was extracted from Basketball Reference as well:

- TEAM, SEASON
- WINS - number of wins
- WIN_PERC - winning percentage

Normally, WINS would simply be WIN_PERC × 82, but the COVID-shortened 2020-21 season, which only had 72 games, necessitates the use of WIN_PERC instead of WINS.

4.7 Ratings Data

In addition, I extracted from Basketball Reference the true values of each team's offensive and defensive ratings in each season. This was used to compute the outputs of a replacement-level lineup in order to fill in the minutes not covered by the lineups in the lineups dataset.

- TEAM, SEASON
- OFFRTG - offensive rating
- DEFRTG - defensive rating

Chapter 5

Model Pipeline

5.1 Archetype Matching

At this point, I now had a dataset of inputs with labels but faced some issues regarding dimensionality. Box scores alone had 13 values per player, meaning that the concatenated 5-player lineup dataset would have 65 features. In addition, lineups with less minutes played are extremely susceptible to noise, so filtered for lineups with at least 100 minutes played, the dataset consisted of only 600 samples.

Fortunately, dimensionality reduction solves this problem while also addressing the goal of incorporating game knowledge. In the NBA today, players can be roughly grouped into one of a few categories:

1. Primary ball-handler - players who dribble the ball up the court each possession and are in charge of initiating the offense

2. Shot creator - players with high rates of isolation plays who create their own shots within the offense

3. 3-and-D (also called "wings") - tasked with defending the best players on the opposing teams; offense role often limited to catch-and-shoot 3-point attempts

and off-ball movement

4. Big men - taller players who linger close to the basket; often also rim protectors who are the last line of defense if the offense gets close to the basket

5. Stretch big - a variant of big men who are more skilled at shooting and have a higher volume of mid-range or 3-point field goal attempts

More specific breakdowns exist, but these are examples of the most common archetypes [13]. Ideally, we are able to reduce each player into how well they embody each type. To begin, I applied PCA on the per-36 minute rates of statistics for players playing at least 15 minutes per game for at least 30 games, who presumably are serious contributors to their teams and played in enough games throughout the season to reduce noise. Prior to PCA, the per-36 values were normalized to mean 0 and standard deviation 1 to account for differences in scale between statistics, such as point and steals.

After some preliminary testing of the columns to include in the PCA, I found that the output is not able to sufficiently capture both a player's tendencies and his proficiency at those tendencies. For instance, the principal components can reconstruct the relative rates at which a player shoots from different parts of the court but isn't able to take into account how well they shoot, and vice versa. To fix this, I adjust the inputs to the PCA to be action-dependent and append additional efficiency-based metrics onto the PCA projections. By action-dependent, I am referring to statistics that count actions that the player chooses to do, such as field goal attempts or passes made, as opposed to those dependent on the results, such as field goals made or assists. This isolates the tendencies of the player, and his proficiency is accounted for with three additional columns: 2-point field goal percentage (2P%), 3-point field goal percentage (3P%), and points created through assists (PTS_CREATED). These were chosen to represent shooting and passing abilities to augment the PCA projections.

Table 5.1: Principal Components.

	PC1	PC2	PC3	PC4	PC5
2PA	0.416*	−0.128	−0.152	−0.215	0.121
3PA	−0.255*	−0.327*	−0.312	0.367*	−0.019
FTA	0.386*	−0.176	−0.239	−0.033	0.196
CATCH	−0.311*	−0.063	−0.446*	0.404*	−0.087
PTS	0.282	−0.304	−0.419*	0.038	0.140
ORB	0.290	0.378*	−0.054	−0.060	−0.010
DRB	0.316	0.260*	−0.185	0.223	−0.208
PASSES	0.226	−0.173	0.354*	0.275*	−0.633
DRIVES	0.131	−0.454*	0.207*	−0.084	0.022
STL	0.052	−0.101	0.455*	0.455	0.649
BLK	0.221	0.342*	−0.154	0.199	0.157
TOV	0.344	−0.283	0.070	0.237	−0.169
PF	0.123	0.309	0.076	0.463	0.010

2PA: 2-point field goal attempts. 3PA: 3-point field goal attempts. FTA: free throw attempts. CATCH: catch-and-shoot field goal attempts. PTS: points. ORB: offensive rebounds. DRB: defensive rebounds. STL: steals. BLK: blocks. TOV: turnovers. PF: personal fouls.

Once the PCA is finalized, the values of its principal components as found in Table 5.1 and the most significant elements in each can sketch a rough idea of what signal is extracted. Here are the noteworthy takeaways, specifically the relevant stats and the top 15 players in order of value, with the numbers in parentheses indicating how many times they appear in the top 15:

- PC1: 2PA, −3PA, FTA, −CATCH
  - Implied type: Shot creator - creates their own shot and gets fouled often, less reliant on catch-and-shoot type shots
  - Top 15: Giannis Antetokounmpo (4), Russell Westbrook, Joel Embiid (5), Andre Drummond (2), Nikola Jokic, Anthony Davis, Jonas Valunciunas
- PC2: ORB, DRB, BLK, −3PA, −DRIVES

  - Implied type: Big men - rebounds and blocks well due to proximity to basket, significantly less 3-point field goal attempts and drives to the basket as a result
  - Implied type: Opposite of Shot creator (3-pointer variety) - similar to other Shot creators but relatively high volume of three-point attempts
  - Top 15: Mitchell Robinson (3), Nerlens Noel, Hassan Whiteside (2), Onyeka Okongwu, Robert Williams III, Moses Brown, Clint Capela (2), Taj Gibson, Joakim Noah, Rudy Gobert, JaVale McGee
  - Bottom 15: Trae Young (3), James Harden (4), Russell Westbrook, Isaiah Thomas, Luka Doncic (3), Stephen Curry, Ja Morant, Derrick Rose

- PC3: PASSES, DRIVES, STL, −CATCH, −PTS
  - Implied type: Primary ball-handler / Slasher - high passing and driving rates, lower shooting and scoring rates; often are smaller players who are pesky on defense and get steals
  - Top 15: TJ McConnell (3), Kris Dunn (3), Elfrid Payton, Rajon Rondo (2), Draymond Green, Ben Simmons (2), Tyus Jones, Dejounte Murray, Tyler Ulis

- PC4: 3PA, CATCH, PASSES
  - Implied type: Catch-and-shoot specialist - high volume of 3-point attempts and catch-and-shoot type shots, often part of scramble drills and thus relative high rate of passes made
  - Top 15: Kristaps Porzingis (4), Kevin Love (2), Brook Lopez, Joel Embiid (4), Nikola Vucevic, Klay Thompson, Dirk Nowitzki, Malik Beasley

Through inspection, the first four PCAs extract information closely resembling the archetypes that we had laid out before. Interestingly, PC2 kills two birds with one stone: at one end are shot creators who specialize in 3-point shooting, like Harden and Trae, and on the other end are big men in the mold of the traditional center whose responsibilities boil down to rebounding and blocking shots. When looking at the principal components themselves, there are some interesting quirks; for example, one would not expect catch-and-shoot specialists to be so outstanding in steals and personal fouls, but these irregularities can reasonably be attributed to the requirement that the principal components maintain orthogonality.

For the most part, the players shown pass the eye test. Giannis and Embiid are two of the most gifted individual scorers in the game, Mitchell Robinson and Hassan Whiteside perfectly reflect the desired skill set of rebounding and blocking, and TJ McConnell and Kris Dunn come to mind as great examples of slashers whose passing and driving are their best abilities. Understandably, the later principal components become less applicable since most of the information has already been extracted through the first two principal components, but even so the five components do an acceptable job at separating players and together explain 85% of variance in the dataset. Given these results, I have confidence to project all players onto $\mathbb{R}^5$, and to complete the dataset I then append the three proficiency columns to create an 8-dimensional representation of each player.

One last adjustment was made to account of differences in minutes played. Since the PCA was performed on rate values, the projections were extremely susceptible to noise for players with very limited playing time. I applied a simple weighted average calculation to regress everyone towards the mean.

$$x_i^* = \frac{MP_i}{MP_i + \text{min\_mins}}(x_i^{proj}) + \frac{\text{min\_mins}}{MP_i + \text{min\_mins}}(\bar{x}),$$

where x_i^{proj} is the PCA representation of a player, MP_i is the number of minutes played, and $\bar{x} = (-0.09, 0.25, 0.09, 0.07, -0.01)$, the average of all PCA representations. min_mins is a pre-determined threshold of 450, which is roughly the number of minutes per season of what I later define to be replacement-level players. The proficiency columns were not adjusted in this way.

5.2 Permutational Invariance

Once I had the reduced forms of each player as well as the dataset of 5-man lineups and their offensive and defensive ratings, I could generate the models. However, one last consideration that I must tackle is the permutational invariance of the lineups. When feeding my five sets of reduced PCA coordinates in $\mathbb{R}^8$ into a model, how much should I care about the order? How do I account for the fact that the five players play different roles and should be treated as such, but strictly speaking there are no defined rules of how they must play and so any permutation of them fed into the model should result in the same output? In the end, I tried six different ways to dealing with this permutational invariance.

5.2.1 Method 1: As Is

When looking through the dataset in its original form, the ordering of the five appears to have no rhyme or reason. However, for the first method I assume that there is some underlying signal and use the order as found in the original dataset.

5.2.2 Method 2: Data Augmentation

Another idea is to use data augmentation. This requires the assumption that the differences in the roles of the five players are not substantial, implying pure permutational invariance. For each datapoint, I create 120 sister datapoints by permuting the five players' values and label each new point with the offensive and defensive ratings of the original. Now, the model has at its disposal a more robust 72000 datapoints.

5.2.3 Method 3: Positional Ordering

This idea is to take advantage of the positional assignments of each player to gain some insight about how the group of five players might play together. Specifically, I am referring to the traditional assignments of point guard, shooting guard, small forward, power forward, and center, whose delineations were widely used and quite rigid in previous decades. In theory, arranging the players according to what position they generally identify as should make things more clear for the model, although the loosening of positional restrictions in recent years likely reduces the information that this provides.

5.2.4 Method 4: PCA Ordering

The next thing to try is similar to the positional method, except to use the computed PCA coordinates. In particular, I arrange them in the following way, roughly in order of perceived importance:

- P1: the player maximizing PC1 - primary shot creator
- P2: the remaining player maximizing PC3 - primary ball handler / slasher
- P3: the remaining player maximizing PC2 - big man
- P4: the remaining player minimizing PC4 - catch-and-shoot specialist
- P5: the remaining player

5.2.5 Method 5: Selective PCA Ordering

The previous methods assume either pure invariance or full determinism. For the next method, I create a combination of the PCA method and data augmentation where I order only the primary shot creator and primary ball handler and take the 6 permutations of the remaining players as augmented data points. This hopefully provides some more context for the model while also improving robustness by increasing sample size to about 3600.

5.2.6 Method 6: Super Selective PCA Ordering

My final method is a simple extension of the previous method by only anchoring the primary shot creator. This loses some information at the cost of quadrupling the sample size to 14400.

5.3 Rating Predictions

The primary model in this work is a neural network, whose primary draw is its ability to capture nonlinear relationships without the need to explicitly create interaction terms. Specifically, the model used in the remainder of this work is a conventional fully connected deep network with hidden layer sizes of 16, 12, 8, and 4, as shown in Figure 5.1. Some architecture testing and hyperparameter tuning was done, but the final model architecture works well enough and the marginal gains of future testing were too small. The number of hidden layers and their sizes were chosen to be sufficiently deep to capture nonlinear relationships without making the parameter space too big, considering that only about 600 lineups have played 100 or more minutes together.

As a reminder, the goal of this modeling step is to predict the performance of a given set of five players. Once the model is trained, I can use it to infer the performances of each lineup or permutations of each lineup if applicable. For instance, when analyzing the data augmentation and selective PCA ordering methods, I set the final output for each lineup as the median prediction of the 120, 24, and 6 sister datapoints, respectively. Originally, the goal of the data augmentation was that the model would be able to learn how to automatically adjust any permutation to

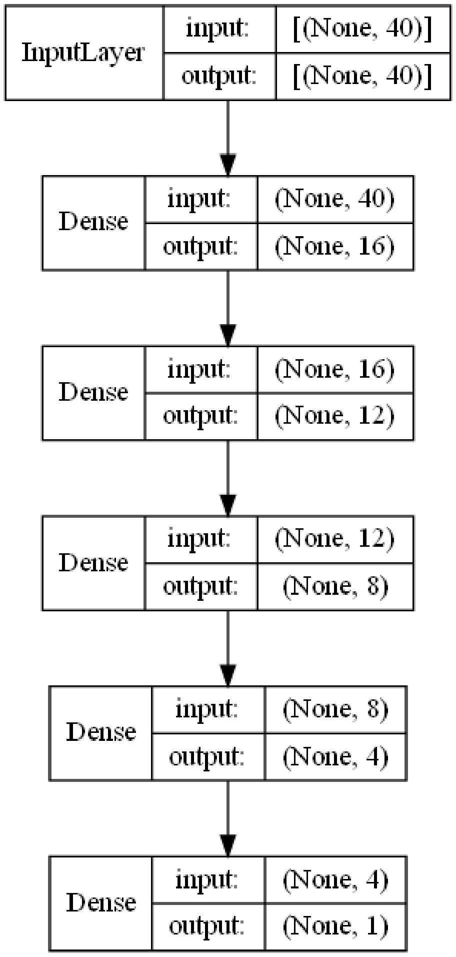


Figure 5.1: Architecture of offensive and defensive models

extract the relevant information, but I surmise that the sample size is too limited for full learning to occur. In my experiments, the models output a sufficiently large range of values for the permutations of each lineups that the median approach is necessary, emulating more of an ensemble effect.

Once inferred, these lineup predictions are combined with how much each was actually played together to reconstruct the season and compute an overall expected performance. The realized performances from the 2016-2021 seasons then serve as a benchmark for the evaluation of these models.

Chapter 6

Results

6.1 Testing Permutation Invariance Methods

To simplify the testing, I first check the validity of each of the permutation invariance methods. I am acutely aware of the relatively small sample size: out of the 12000 lineups in the past six years, only 600 played enough minutes to qualify, so I know that the dataset is susceptible to overfitting as is. As mentioned before, the scope of the dataset is a (mostly) unavoidable problem caused by the non-stationarity of the game, which I will discuss later. In order to check for overfitting, I can train the models and consolidate lineup predictions when applicable for each method and then compare how well each method does on the training and testing data.

When comparing the models' performances in Table 6.1, it is clear that the methods that didn't involve any data augmentation heavily overfit on the training set. This makes sense as my proposed architecture has a total of 1005 parameters, which is problematic and indicates that some kind of data augmentation is needed. However, even the selective PCA ordering method with 3600 datapoints appears prone to overfitting, and the performance suffers about as much as for the other methods. Of the methods outlined in the previous section, the only one that performs respectably

Table 6.1: RMSE of Permutational Invariance Methods

		Offensive Model		Defensive Model		
	Count	Training	Testing	Training	Testing	Overall
As-Is	579	6.138	7.446	5.862	9.285	0.1054
Augmentation	69480	6.709	7.063	7.083	7.774	0.1021
Positional Ordering	579	6.615	7.340	7.872	8.585	0.1177
PCA	579	6.180	8.029	6.673	8.984	0.1107
Selective PCA	3474	6.200	7.082	7.158	8.509	0.1122
Super Selective PCA	13896	6.881	6.748	8.306	7.593	0.1160

on the testing data is the data augmentation method, which employs almost 14000 data points. Experimentally, RMSE only increases from 6.7 to 7.1 on offense and 7.1 to 7.8 on defense, and the the R^2 of the overall model was the best of all the methods at 0.484. The impractibility of most of the methods is admittedly disappointing, but given the nature of the problem and the limitations of the dataset there is not much else I can do. Moving forward, all testing will use the data augmentation method.

I'd like to point out here that while testing the methods involving data augmentation, I nearly made a major mistake in creating the train-test split. Generally, randomizing the data and then selecting the first 70% of the randomized order is a straightforward and computationally inexpensive method to create the two sets. However, in this case, because the samples in the enlarged dataset are no longer independent, this approach is not viable. By taking a random selection of augmented points, I essentially choose 70% of lineup permutations for each lineup to be in the training set, meaning that all lineups are still substantially represented in both training and testing. Recall that I only take the median of the permutation predictions as the final prediction for each lineup; since the permutations are all drawn from the same distribution, their median predictions are almost identical, which results in the test set significantly overperforming. After fixing this mistake, I can better trust the results of the training and testing sets. Fortunately, the models still perform well on the test set as shown in Table 6.1, indicating that they generalize well.

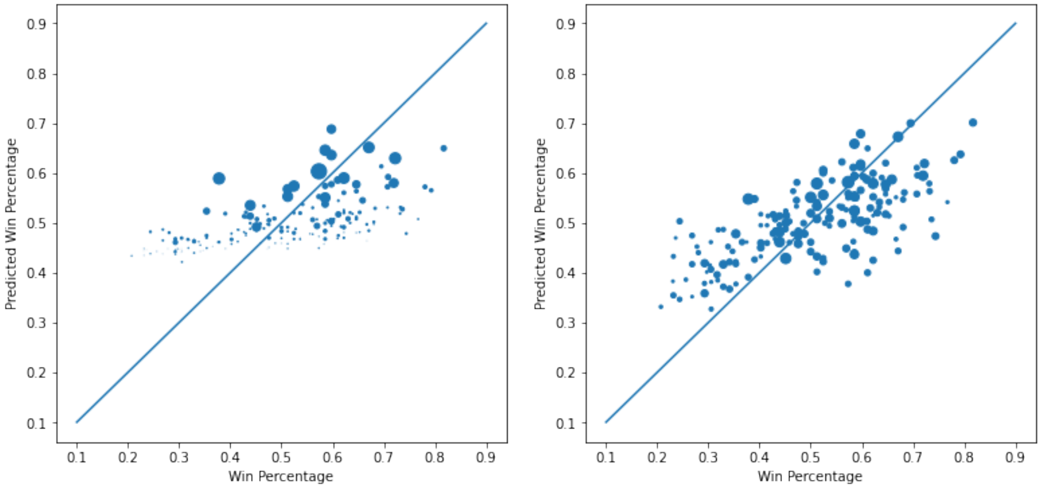
6.2 Evaluating the Model

A closer look at the models reveals some concerns. First, using only the training and testing sets as inputs (i.e. lineups that play at least 100 minutes), the model does not perform well in predicting overall team performance, with a Pearson correlation coefficient of 0.62 and an error of 0.1180 with respect to the true values (Figure 6.1). This is not too surprising, as there are such a large number of possible lineups that the well-behaved lineups of each team only make up 10% or less of the total possible minutes played.

To remedy this, I assume that the model is able to pick up some perceptible signal from the well-behaved lineups that generalizes to the volatile lineups as well. In the same way that turtles lay hundreds of turtle eggs at a time in the hopes that at least a few will survive, I feed thousands of unreliable inputs and hope that in the long run the overperformers and underperformers cancel out and give a consistent estimate of the team's performance. In addition, to smooth out the team-level predictions, I use the real ratings for each team to estimate a "replacement-level" lineup to fill in whatever minutes are not covered by the SAP lineups dataset, which excludes lineups that play less than a threshold of 12 minutes. Teams decimated by injuries or with more roster turmoil can have more than half of their total minutes played by these outliers, so this method can greatly reduce variance in the team-level predictions. Applying these changes, the updated lineup predictions and the overall outputs are as follows in Figure 6.1.

As expected, the updated model improves significantly, with a stronger correlation coefficient of 0.72 and a lower error of about 0.1021, or 8 games per season; recall that the use of the Pythagorean theorem automatically bakes in a relative error of 0.04. The cumulative error-frequency plot appears solid as well, with an AUC of 0.699 (Figure 6.2). Overall, two-thirds of predictions lie within the relative error of 8 games, and 80% within 10 games.

Figure 6.1: Final team predictions using only high-volume lineups vs full dataset



Marker sizes indicate the number of minutes included in the prediction

However, looking at the offensive and defensive models individually in Figure 6.4 gives reason for pause. The offensive model performs even better than expected, with an even stronger correlation of 0.76 and more than half of variance accounted for. However, the defensive model lags behind, with a moderate correlation of 0.48 and only 15% of variance covered. When considering the inputs of the model, this is not at all surprising since the statistics used are almost all offensive-based. I will explore these choices further in the discussion section, but as a result of this disparity, in the next section I will introduce Offensive and Defensive WAR, variants pertaining to only offensive and defensive contributions.

One final note that I want to point out is that the model is unfortunately not calibrated. It is clear from Figure 6.3 that the model overestimates underperforming teams and underestimates overperforming teams. For now, I will leave this as is. Fortunately, the inverse looks fine, so the predictions are consistent within themselves.

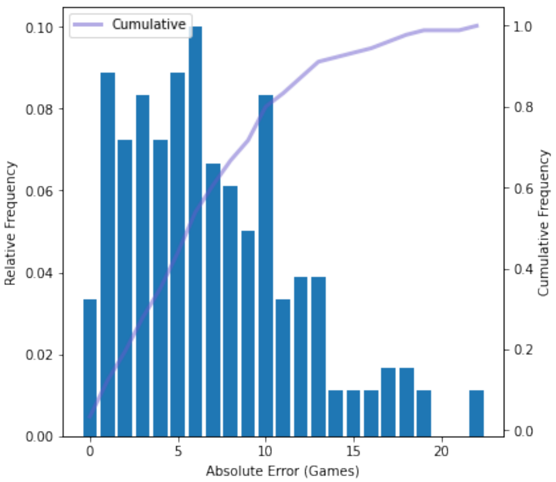


Figure 6.2: Error-frequency plot with cumulative form of final team predictions using full dataset

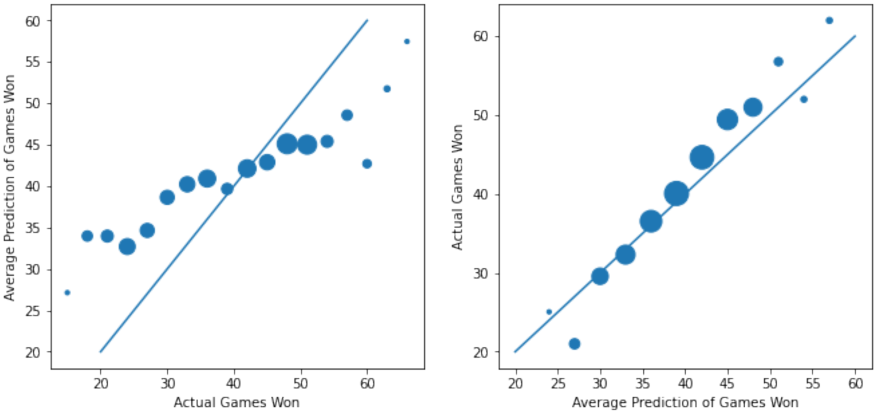


Figure 6.3: Reliability curves of final team predictions using full dataset

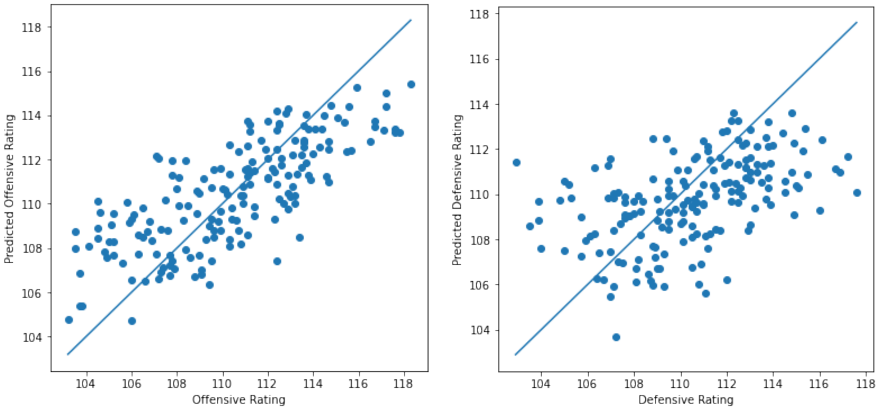


Figure 6.4: Final team predictions using full dataset of offensive vs defensive model

6.3 Visualizing the Model

One of the initial goals was to create a more transparent methodology than previous metrics like APM or Win Shares. This work already addresses this to a satisfactory degree with the creation of player archetypes and Pythagorean formula. However, the final piece that is still a black box is the neural network itself. The inputs are understandable and the outputs are reasonable, but how does the model get from one to the other?

In deep machine learning architectures, one interpretation is that each node in a hidden layer looks for and extracts a specific nugget of information, and the nodes in the next hidden layer piece them together into their own nuggets of information. This is especially prevalent in image processing, where the layers describe larger and more general features the deeper they are. In the context of my work, the hope is that the layers work similarly and steadily build pieces until the model can decide how those pieces fit together.

In Figure 6.5 are the weights of the first hidden layer in the offensive model. Since the input is 40-dimensional and there are 16 nodes in the first hidden layer, this figure represents a 16 × 40 matrix containing the 640 parameters.

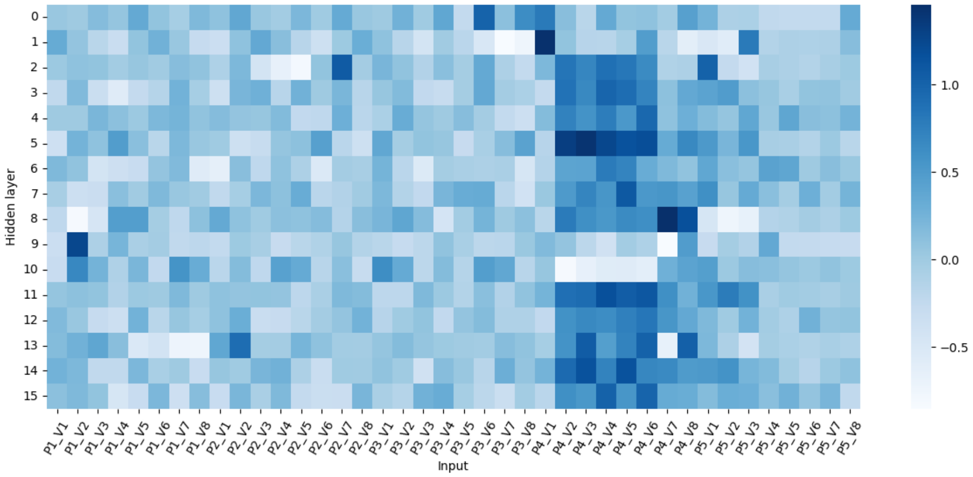


Figure 6.5: Matrix of weights in offensive model

Clearly, the model strongly favors Player 4. Given the permutational invariance assumption, this is not cause for concern, as the model will arbitrarily choose one of the players as an anchor point. Within Player 4, his attributes are valued by different nodes in the first hidden layer. For instance, the second node strongly weights the first value in his representation, which corresponds to his tendency to play like a shot creator. This node also heavily weights the third value in Player 5's representation, which describes his ability as a ball-handler. Thus, I can infer that this node likely focuses on the interaction between ball-handler and shot creator. Similarly, different parts of the representation of Player 4, and by extension all players, are targeted by different nodes in the first hidden layer, and each node captures some form of what is happening on the court.

It is admittedly difficult to fully piece together the behavior of all the hidden nodes with just the naked eye, and more work in designing the architectures and creating visualizations would likely further improve the interpretability of this piece of the puzzle. But it is undeniable that the underlying structure is there, that some parts of it make sense, and that as a whole it works. With four hidden layers, the model is able to capture complex relations between the forty inputs that arguably shouldn't be fully graspable by human eyes, and I have already experimentally shown that the culminating lineup predictions are indeed reflective of reality. There exists an inherent tradeoff between explainability and explaining power, and I believe that the neural network as is lies not too far from the optimum.

6.4 Computing WAR

Once the model is verified, I can complete the final step of the puzzle: computing WAR. For a given player, season, and team, the methodology is as follows:

1. I first compute the average replacement-level player. Each of the 30 NBA teams

can roster 15 players, so I take the average of the $30 \times 15 \times 6 + 1 = 2701$th to the $30 \times 16 \times 6 = 2880$th ranked players by minutes played as the replacement-level player in the past 6 seasons, which is $(-0.26, 0.68, 0.33, 0.13, 0.04, 0.51, 0.29, 2.92)$. This checks out, as there are generally many young centers with raw potential relative to pure scorers near the end of the bench.

2. I extract from the augmented dataset all lineups for that team and create a copy where I replace all instances of the given player with the replacement player.

3. I use the model to predict the offensive and defensive ratings for each lineup for both lineup datasets and use the minutes played of each lineup to compute the expected ratings of the entire team that season under both scenarios.

4. I apply the following calculations to compute WAR, Offensive WAR (WAR_OFF), and Defensive WAR (WAR_DEF). Let o_{off} be original offensive rating, o_{def} be original defensive rating, r_{off} be replacement offensive rating, r_{def} be replacement defensive rating, and $W(o, d)$ be the expected win percentage of a team with ratings o, d as described by the Pythagorean formula. Then,

$$\begin{aligned} \text{WAR} &= 82 \times \big(W(o_{off}, o_{def}) - W(r_{off}, r_{def})\big) \\ \text{WAR\_OFF} &= 82 \times \big(W(o_{off}, r_{def}) - W(r_{off}, r_{def})\big) \\ \text{WAR\_DEF} &= 82 \times \big(W(r_{off}, o_{def}) - W(r_{off}, r_{def})\big) \end{aligned}$$

In other words, WAR computes the difference of the player compared to the replacement, and WAR_OFF and WAR_DEF describe the difference of the player compared to the replacement, assuming that the player has the defensive or offensive abilities of a replacement player, respectively.

6.5 WAR Results

Table 6.2 displays the computed leaders in WAR for each season, as well as their RAPTOR-based WAR, VORP-based WAR, and Win Share values and their percentiles. At first glance, this table makes no sense. I for one do not consider Rudy Gobert to be a top 5 player, not to mention the single most valuable player four times in the past six seasons. However, this surprising result is actually incredibly encouraging. For context, Rudy Gobert is a controversial and unique player in that he is extremely gifted on the defensive side of the ball but below average on the offensive side and maligned as a waste of money by proponents of the new age of free-flowing offense and stretch bigs. Given the overwhelming emphasis placed on offense compared to defense, an objective measure that equally weights both sides would naturally appear to favor defensive-minded players who have been generally underrated by traditional offense-centric philosophies, when in fact it is giving those players their deserved laurels.

Table 6.2: Values and percentiles of WAR metrics of top 5 players by year in WAR

	PLAYER	**WAR**		**RAPTOR**		**WS**		**VORP**	
		Value	Perc	Value	Perc	Value	Perc	Value	Perc
2016	DeAndre Jordan	9.0	100	7.8	84	11.8	96	9.5	85
	Ricky Rubio	8.1	-	5.8	71	6.1	67	4.9	61
	Hassan Whiteside	7.4	-	4.7	58	9.5	90	5.9	69
	Tristan Thompson	6.9	-	3.4	39	7.3	78	1.9	27
	Myles Turner	6.6	-	5.3	65	8.0	82	5.4	66
2017	Paul George	7.5	100	8.8	87	8.9	88	9.7	86
	Steven Adams	7.2	-	6.8	79	9.7	91	5.4	66
	Kyle Lowry	6.8	-	10.7	93	10.2	93	12.2	91
	Nicolas Batum	6.5	-	1.8	20	3.5	30	3.2	45
	Draymond Green	6.3	-	6.9	80	6.1	67	7.6	79
2018	Rudy Gobert	10.9	100	11.4	95	14.4	99	13.0	93
	Paul George	9.5	-	17.4	100	11.9	96	17.8	98
	Ben Simmons	8.7	-	3.0	35	8.2	84	10.3	88
	Steven Adams	7.2	-	6.4	76	9.1	89	5.7	67
	Anthony Davis	7.1	-	9.6	90	9.5	90	14.3	95
2019	Rudy Gobert	8.4	100	10.6	93	10.7	94	8.9	83
	Hassan Whiteside	7.5	-	6.1	74	8.5	85	7.0	76
	Damian Lillard	6.6	-	11.9	96	11.6	96	15.9	97
	Duncan Robinson	5.5	-	6.0	73	5.3	57	3.5	47
	Kawhi Leonard	5.4	-	11.7	95	8.7	86	13.8	94
2020	Rudy Gobert	11.4	100	12.4	97	11.3	95	10.3	88
	Draymond Green	7.2	-	7.5	84	4.6	47	5.1	63
	Clint Capela	6.9	-	8.8	87	8.2	84	5.9	69
	Ben Simmons	5.4	-	4.7	58	6.0	65	6.2	71
	Russell Westbrook	5.4	-	1.4	16	3.7	32	9.2	84
2021	Rudy Gobert	8.6	100	10.8	94	11.7	96	9.7	86
	Clint Capela	8.1	-	5.7	70	8.3	85	5.7	67
	Mitchell Robinson	8.0	-	5.3	65	8.5	85	5.4	66
	Steven Adams	7.5	-	5.1	63	6.8	73	5.4	66
	Jakob Poeltl	5.8	-	6.8	79	6.9	75	4.6	58

The columns refer to WAR, RAPTOR WAR, Win Shares, and VORP WAR, respectively.

When comparing each pair of these categories as seen in Figure 6.6, the metrics indeed all align to a certain degree. Unsurprisingly, the battle-tested RAPTOR, WS, and VORP values align very closely, but encouragingly my fledgling WAR metric is weakly correlated with each. As WAR is refined further, this correlation should improve, but this is not the main priority. I am not aiming to recreate the other metrics but instead introducing another angle to extract new information, so it is only necessary here to ensure that the results are at least reasonably aligned with prior observations.

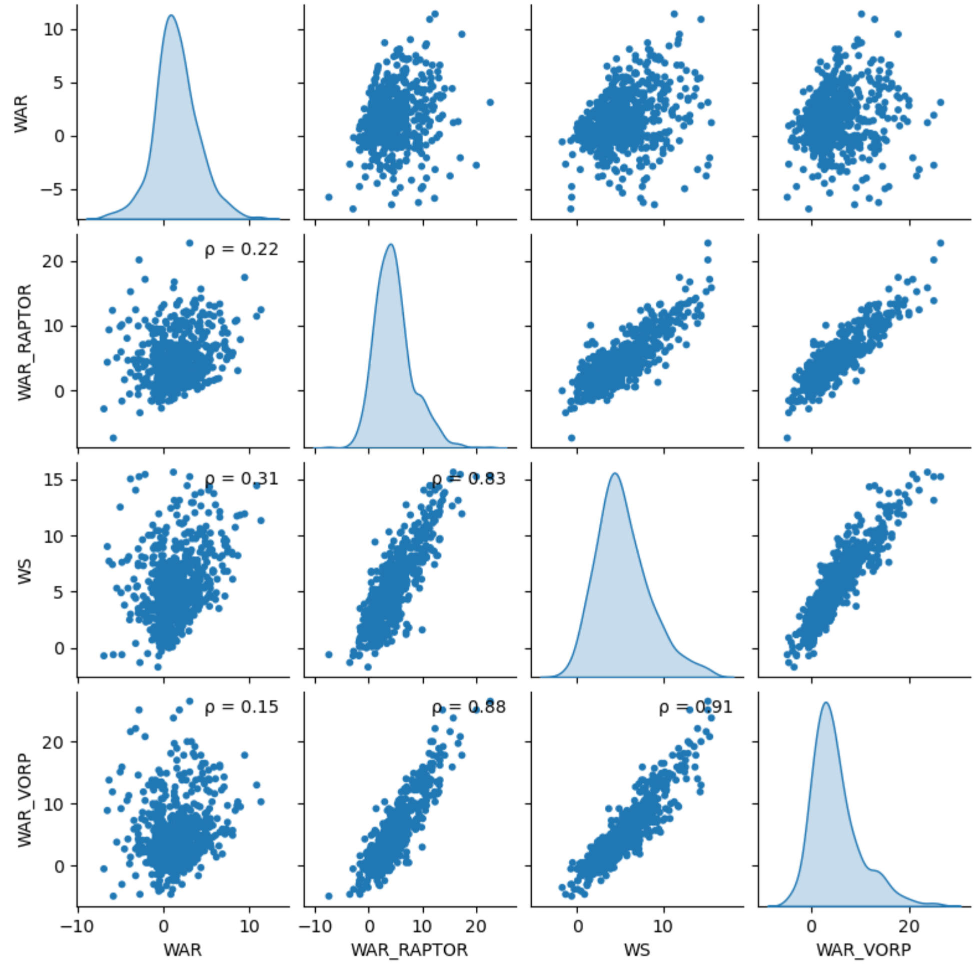


Figure 6.6: Pairwise comparisons of WAR metrics

6.6 Offensive WAR Results

Given how well the offensive rating model performs, it is enlightening to compare offensive metrics as well. Table 6.3 displays the leaders in Offensive WAR for each season, as well as the offensive forms of the other metrics.

A quick glance at Table 6.3 confirms that the players listed are indeed generally regarded as strong offensive players. To me, the most surprising result is the string of centers at the top of the list from 2019-2021 in Hassan Whiteside, Rudy Gobert, and Jakob Poeltl. This is notable because they are not generally regarded as elite offensive players, especially Gobert as previously discussed. However, upon a deeper dive, the results are not as surprising as they are at first glance. In their respective years, the leaders had individual offensive ratings of 125, 131, and 123 as the starting centers on their teams, significantly higher than the league average of 113. Even though they may not be considered elite individual offense weapons, they still operate and contribute in a way such that their teams perform exceptionally efficiently while they are on the court, and it is incredibly encouraging that the model is able to catch that. Of course, it is worth acknowledging that many of the highest rated players are centers, which raises concerns that the model may have biases towards traits or actions that favor this type of player.

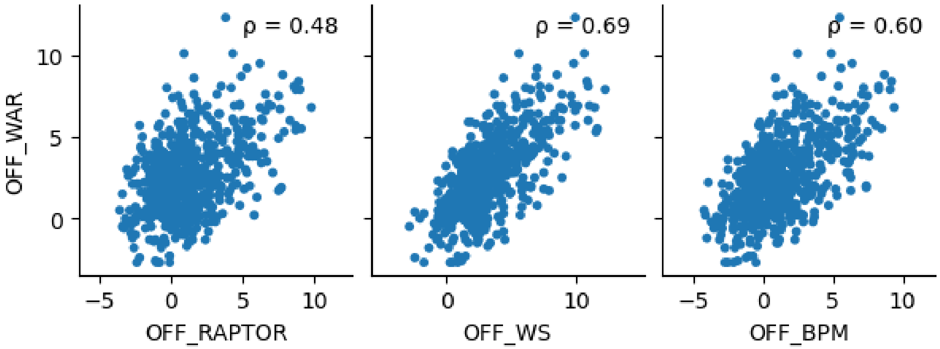


Figure 6.7: Offensive WAR vs other offensive metrics

Table 6.3: Values and percentiles of offensive metrics of top 5 players by year in offensive WAR

	PLAYER	**WAR**		**RAPTOR**		**WS**		**BPM**	
		Value	Perc	Value	Perc	Value	Perc	Value	Perc
2016	K.A. Towns	12.3	100	3.8	87	9.9	98	5.5	93
	LeBron James	9.5	-	6.2	96	9.8	98	6.4	95
	Russell Westbrook	8.8	-	7.8	98	8.5	98	8.7	100
	Bradley Beal	8.7	-	4.9	92	6.9	92	4.2	87
	DeAndre Jordan	7.6	-	1.6	68	7.2	93	2.5	75
2017	K.A. Towns	10.1	100	4.3	89	10.6	99	4.9	90
	LeBron James	6.9	-	7.0	97	11.0	99	7.3	98
	Damian Lillard	6.6	-	6.6	97	9.9	98	7.0	97
	Bradley Beal	6.3	-	2.5	78	4.4	77	2.9	79
	Klay Thompson	6.3	-	1.7	69	3.1	59	1.7	65
2018	Paul George	9.2	100	5.3	93	7.0	92	5.6	93
	K.A. Towns	8.1	-	3.5	86	7.2	93	5.4	92
	Nikola Vucevic	7.2	-	1.8	71	5.4	84	5.0	90
	LaMarcus Aldridge	7.1	-	0.7	50	6.4	90	2.4	74
	Damian Lillard	7.0	-	6.8	97	9.7	98	6.6	96
2019	Hassan Whiteside	10.1	100	0.9	54	5.6	85	2.5	75
	Damian Lillard	8.1	-	8.6	99	10.9	99	8.3	99
	Duncan Robinson	7.6	-	3.2	83	3.7	67	1.5	63
	Devin Booker	7.0	-	4.8	91	6.0	87	3.4	83
	Rudy Gobert	6.4	-	-0.4	28	6.5	90	1.7	65
2020	Rudy Gobert	8.0	100	-0.3	30	6.1	88	2.1	70
	Stephen Curry	7.9	-	8.7	99	6.5	90	8.3	99
	Nikola Jokic	7.9	-	9.0	100	12.2	100	9.1	100
	Damian Lillard	7.5	-	7.5	98	9.6	98	7.5	98
	Mikal Bridges	6.8	-	1.9	72	5.4	84	2.4	74
2021	Jakob Poeltl	8.6	100	1.6	68	4.4	77	0.9	52
	Nikola Jokic	8.4	-	8.9	100	10.8	99	9.2	100
	Joel Embiid	8.0	-	4.2	88	7.9	95	7.2	98
	Mitchell Robinson	7.5	-	0.5	47	5.3	83	1.0	54
	Rudy Gobert	7.4	-	0.1	37	7.3	94	2.9	79

The columns refer to Offensive WAR, Offensive RAPTOR, Offensive Win Shares, and Offensive Box Plus-Minus, respectively.

In terms of comparison with other metrics, the similarities are even more clear here. The correlation coefficients increase to as high as 0.7, indicating that the offensive mechanism is able to capture much of the information found by the other metrics. In particular, this implies a drastic discrepancy between my offensive and defensive WAR metrics. This is true, as my defensive WAR values are at best weakly correlated with a coefficient of 0.29 with RAPTOR and at worst a negligible coefficient of 0.01 with Win Shares. However, given the differences between the methodologies — offensive and defensive ratings by lineup for mine, a direct weighted sum method by player for RAPTOR, plus-minus results for WS and BPM — it is comforting that the results from both the offensive and defensive metrics align to such a degree.

With that being said, there are some differences between WAR and the other metrics that I'd like to briefly explore here. For instance, Offensive RAPTOR highly rates Chris Paul, placing him at fifth and second in 2016 and 2017, while he is only at the 38th and 40th percentiles by my metric. As for other outliers, Luka Doncic and Trae Young place in the top 5 by RAPTOR most years but are only in the 60-70th percentiles by WAR. The fact that these three players all rank in top 10 in assists during the past six seasons implies that perhaps my model heavily weights primary scoring threats and is not as impressed by those who create offense through passing and ball movement. There is no objective system to determine whether this preference is correct, but a new metric that passes the eye test but is not parallel to previous metrics has undeniable value. When used in conjunction, WAR and the other metrics can provide a more nuanced evaluation than if each is used alone.

6.7 Addendum: A Different Architecture

When I discuss my methodology with others, a very reasonable question often arises: why create two separate models for the offensive and defensive ratings instead of directly predicting win percentage? This is a great point, and when this first came to my attention, I enthusiastically tested its veracity by creating a combined model with the same functionality but with only one output, as seen in Figure 6.8.

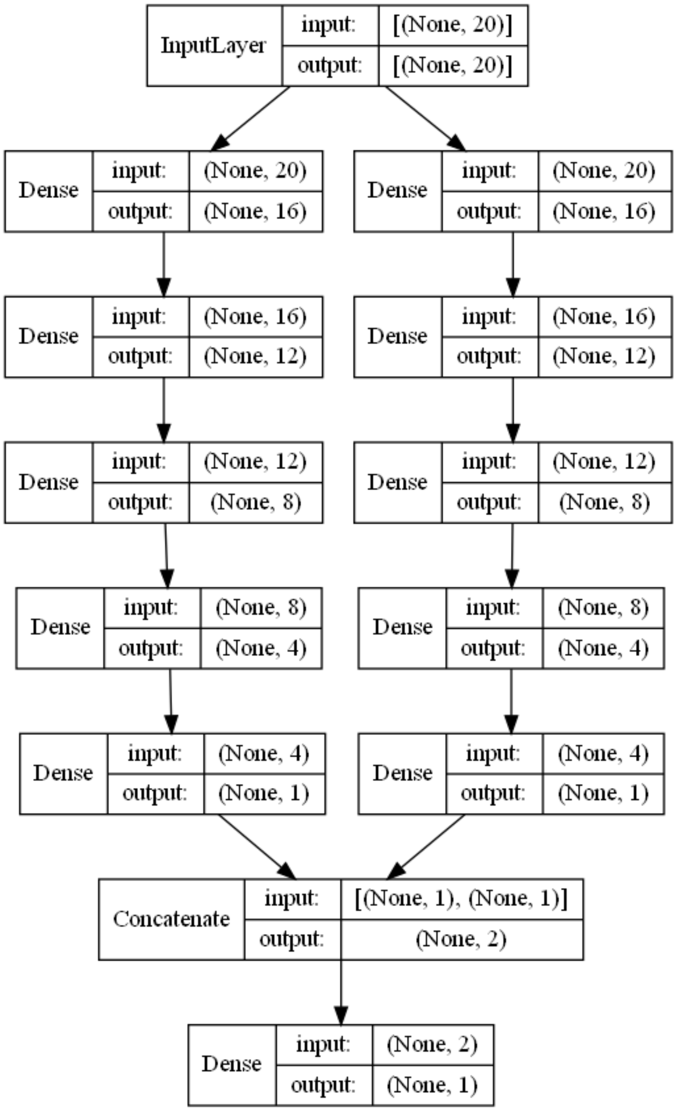

InputLayer | input: [(None, 20)] | output: [(None, 20)]
Dense | input: (None, 20) | output: (None, 16)
Dense | input: (None, 20) | output: (None, 16)
Dense | input: (None, 16) | output: (None, 12)
Dense | input: (None, 16) | output: (None, 12)
Dense | input: (None, 12) | output: (None, 8)
Dense | input: (None, 12) | output: (None, 8)
Dense | input: (None, 8) | output: (None, 4)
Dense | input: (None, 8) | output: (None, 4)
Dense | input: (None, 4) | output: (None, 1)
Dense | input: (None, 4) | output: (None, 1)
Concatenate | input: [(None, 1), (None, 1)] | output: (None, 2)
Dense | input: (None, 2) | output: (None, 1)

Figure 6.8: Architecture of combined model

Note: the Concatenate layer performs the Pythagorean formula computation.

Unfortunately, empirically the model severely underperformed. Overall, the win percentage predictions had a Pearson correlation coefficient of only 0.63 and the error was 0.1130, compared to 0.72 and 0.1021 for the two-model approach. A deeper look reveals why: after training for only 10 epochs, the model already does not generalize well to unseen data as shown in Figure 6.9. At the lineup level, the model performs well on the training data: correlation is 0.67 and the model accounts of 42% of the variance in the data with an error of 0.203. However, on the testing data, correlation plummets to 0.39 and the model now only accounts for 9% of the variance for an error of 0.264.

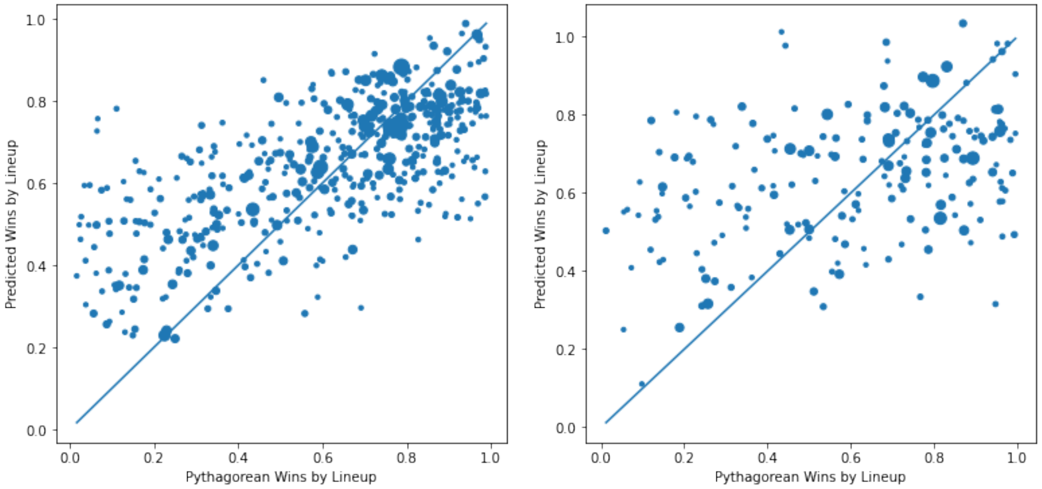


Figure 6.9: Training vs testing results of combined model on lineup predictions

Reasonably, I can assume that this is due to the omission of the intermediate steps of the rating predictions. By breaking down win percentage into two subparts that are proven to recreate win percentage well, I provide additional, signal-rich information to the model in the middle of its process and thus simplify its task and improve generalizability. This result reiterates one of my initial goals of reducing the black-box abstractions of previous evaluation metrics and shows the value of applying game knowledge in a meaningful way.

Chapter 7

Discussion

7.1 Future Work

Of course, the work is not done yet. Throughout my work, I identified multiple areas of improvement that I unfortunately did not have the bandwidth or expertise needed to complete.

7.1.1 Defensive Inputs

One of the largest issues in sports analytics is the underrepresentation of defense in player evaluation. Indeed, the complex interactions on defense coupled with a dearth of defensive statistics results in poor defensive metrics. FiveThirtyEight puts this well in their explanation of RAPTOR:

> “Defense is more of an uphill battle. Some of the statistics RAPTOR uses to rate defensive performance are really more like proxies for other unmeasured statistics. We expect that the state of publicly available defensive metrics will improve in future years, and RAPTOR will improve along with them.”

It is clear that this trend continues in my work. The player archetypes drawn from the PCA are heavily driven by offensive philosophies, and some size and speed generalizations provide some vague insights into how players might fit together defensively. Moreover, I did not include any defensive efficiency metrics; this oversight is due to both my being unsure of which metrics would be explanatory as well as the general lack of defensive-oriented data. Even when I attempted to take inspiration from RAPTOR's calculation, many of the statistics they used, such as offensive fouls drawn and positional opponent data, are not publicly available. In future work, I would likely use opponent field goal percentage as a starting point.

7.1.2 Net Rating Approach

In this work, I took the approach of using offensive and defensive ratings to most accurately compute how many games a team would win. An alternative that came to mind near the completion of my work is the possibility of using the net rating, the difference between the two, alone to estimate win percentage. With a reduced number of degrees of freedom, the precision may improve enough to overcome the net rating estimate.

In a way, this is melding the plus-minus and counting statistics approaches of RAPM and RAPTOR, the gold standards in player evaluation. By using the same approach of building up and using player statistics to predict net rating instead, I could separate players' contributions in a distinct way from the top-bottom approach of RAPM. This is a promising area of experimentation and could possibly be a new direction of player evaluation research.

7.1.3 Model Calibration

As stated in the results section, my model is too conservative in its predictions. There are a few methods to correct for this, including isotonic regression and spline calibration [18], but considering that I was more interested in comparing players than the model accuracy itself, I did not see this as a major priority. This would be an important area to correct in future work.

7.1.4 More Data

A repeated obstruction that I encountered was the limited sample size. As stated earlier, I purposefully limited the sample size to only the past six seasons, those after the height of the three-point revolution, for fear of muddying the data space with a different distribution of teams and play styles. However, I may have been too careful, and the inclusion of additional years' data likely benefits this analysis in terms model robustness to a degree that outweighs the harms I was trying so hard to prevent.

7.1.5 Transformers

One additional idea to take care of the permutational invariance that I was very excited to test was transformers, a machine learning architecture often used in natural language processing applications. Transformers take in inputs like words iteratively and apply the concept of self-attention, which estimates the relative importance of each input to all other inputs and introduces symmetry that can lead to permutational invariance. In the context of my work, the words would be players, and the self-attention should capture their interactions. Typically for NLP applications, positional embeddings are introduced to provide more information about the players relative to the others; using these would essentially recreate the positional and PCA ordering methods, and omitting them preserves the permutational invariance of the data augmentation method. Overall, transformers would both improve computing power and maintain the flexibility of being able to implement the various methods outlined in this work.

However, there are some issues that prevented their inclusion. Primarily, transformers require significant sample sizes, which is a salient issue in this context. To combat this, I could use pretrained models such as Hugging Face, but the exclusion of the permutational invariance would introduce significant biases and reduce the usefulness of the pretrained model. In total, there were too many unknown factors to justify using transformers over refining other areas of the analysis. However, given more time and experience with transformers, this is an area with enormous potential.

7.1.6 Media Reports

Originally, one source of input that I planned to incorporate was the use of media reports. The idea is to apply language processing techniques to analyze articles written by the media about each team to extract additional information about players. This added nuance would ideally adjust the PCA models to adhere more closely to expectations and thus reduce variance in the PCA projections. Expected sources included ESPN, the Athletic, and each team's own journalism team, but this approach ended up getting tabled and could be interesting to check out in future work.

7.1.7 Supervised Learning

A final possibility that I previously explored is using supervised learning methods instead of the PCA analysis done here. This would entail manually labelling certain players as the most ideal form of some pre-chosen archetypes and then applying clustering algorithms like KNN to get a different reduced representation of each player. Given this approach's clear propensity for preconceived biases, I did not explore it, but its potential is worth at least mentioning.

7.2 Conclusion

Overall, I believe that this work is a major step forward for player evaluation. The crux of this work is realizing that player-player interactions manifest when those players are on the court together. Thus, analyzing how well lineups of five players play together most directly captures what happens on the court. PCA is used to simplify this analysis into examining how archetypes fit together, and the season as a whole can be subsequently rebuilt using lineup data and the Pythagorean formula.

By attacking the problem from a bottom-up approach, I lay the foundation for the creation of newer metrics in this novel way. Technological advances and the collection of increasingly granular datasets allow for more nuanced analysis and even new approaches like mine. As these areas continue to improve, player evaluation will only get better. Of course, it is impossible to completely capture a player's contributions using a single number; teammates, scheme fit, and a myriad of other factors concurrently affect how much a player contributes to his team. However, these metrics are not in vain, as they each answer their own question. When used in conjunction, they can tell a compelling story and reveal truths otherwise unseen.

Bibliography

[1] Deep dive on regularized adjusted plus-minus i: Introductory example. https://squared2020.com/2017/09/18/deep-dive-on-regularized-adjusted-plus-minus-i-introductory-example/.

[2] Nba win shares. https://www.basketball-reference.com/about/ws.html. Basketball Reference.

[3] Regularized adjusted plus minus (rapm). https://www.nbastuffer.com/analytics101/regularized-adjusted-plus-minus-rapm/.

[4] S. T. J. S. M. A. C. Thomas, Samuel L. Ventura. Competing process hazard function models for player ratings in ice hockey. *Annals of Applied Statistics*, 7(3):1497–1524, 2013.

[5] B. Gerard. Is the moneyball approach transferable to complex invasion team sports? *International Journal of Sport Finance*, 2(4):214–230, 2007.

[6] B. Index. Lebron introduction. https://www.bball-index.com/lebron-introduction/.

[7] T. D. Jocks. What is real plus minus? https://thedatajocks.com/what-is-real-plus-minus/.

[8] A. I. Magazine. A primer on sports analytics: A new dimension of sports. https://analyticsindiamag.com/a-primer-on-sports-analytics-a-new-dimension-of-sports/.

[9] R. R. . D. Memmert. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *Springer Plus*, (5):1410, 2014.

[10] MLB. Wins above replacement (war). https://www.mlb.com/glossary/advanced-stats/wins-above-replacement.

[11] D. Myers. About box plus/minus (bpm). https://www.basketball-reference.com/about/bpm2.html. Basketball Reference.

[12] J. Naismith. *Basketball: Its Origin and Development*. University of Nebraska Press, 1996.

[13] B. Nation. Offensive roles. https://www.bball-index.com/offensive-archetypes/.

[14] S. Nation. The story in a box score. https://www.blazersedge.com/2017/1/19/13755264/nba-box-score-history.

[15] NBA. History. https://careers.nba.com/history/.

[16] K. Pelton. The warp rating system explained. http://sonicscentral.com/warp.html.

[17] M. H. Ronald Yurko, Samuel Ventura. nflwar: a reproducible method for offensive player evaluation in football. *Journal of Quantitative Analysis in Sports*, 15(3):163–183, 2019.

[18] R. Sangani. A comprehensive guide on model calibration: What, when, and how. https://towardsdatascience.com/a-comprehensive-guide-on-model-calibration-part-1-of-4-73466eb5e09a.

[19] S. Shea. The 3-point revolution.
https://shottracker.com/articles/the-3-point-revolution. Shot Tracker.

[20] N. Silver. How our raptor metric works.
https://fivethirtyeight.com/features/how-our-raptor-metric-works/.
FiveThirtyEight.