Creating Dream Teams under the Salary Cap: NBA Player Selection Optimization

Pipe Galera

18 of August, 2019

Abstract

How can teams use computer modeling to increase their winning rate? This project-paper proposes optimizing NBA player selection based on stats and salary of the players. I use a linear regression model to estimate the weight of offensive and defensive stats on winning an NBA game. Then, I use these weights to set up an optimization problem to select the best NBA line-up to achieve two main goals: i) Maximum wins under the 2018 salary cap ii) Minimum team salary that match the win threshold of the last league champion.

The project applies:

- Scrapping static and dynamic webpages with Selenium and BeautifulSoup.
- Preprocessing data with Pandas and Numpy.
- Descriptive analysis visualization with Seaborn.
- Modeling with scikit-learn.
- Optimization with PuLP.

Code available at: https://github.com/pipegalera/Optimizing-NBA-Player-Selection

Contents

1 Introduction

The National Basketball Association -NBA- is the biggest basketball league in the world, with more than 15 million spectators every year watching the finals. It is also the most profitable sports business for the teams, with an average team net value of $1.9 billion [2]. Given the impact of the NBA, more and more teams are using data analytics to try to gain some advantage on and off the court. An early example of the develop of this data revolution is the praised statistician Dean Oliver, who developed several advanced metrics to analyze the performance of teams and players, now used widely [10].

This relative new line of analytics referred to as *Sabermetrics* or just *Sport Analytics* started to rise with the eruption of big data, and therefore the ability to match and train the models with real data. As basketball is a competitive game, the main focus of this field is usually match prediction and forecasting of the results.

The current project focuses on the same objective, using linear regression and optimization to predict the team win rate. The regression outperforms other models, explaining 90% of the variance in team wining rates using only four features ($R^2 = 0.90$). This model is used to tune the later linear optimization, that selects the players among the 2018-2019 Regular Season to create a fictional team, constrained by a limited budget (League salary cap). The resulting team have a theoretical win rate close to 97%. Also, the optimization model can construct a team that achieve the same game-record of the 2018-2019 Regular Season winner (Milwaukee Bucks with a 76% win rate) with 12 times lower team payroll ($10.6m against $126m).

The rest of the project is divided as follows: Section 2 gives the context of the model, explaining the basic structure of the NBA as well as the definition of the different metrics used. Section 3 describes the model strategy and the optimization problem. Section 4 remark the results. Section 5 explains the strengths and weaknesses of the paper with respect of the most recent paper with the same modelling goal. Finally, Section 6 concludes the project paper.

2 The NBA: Concepts and Statistics

2.1 Brief overview of the game

The NBA is constituted of by 30 teams divided equally in 2 Conferences, the East Conference and the West Conference. Every year, the teams play each other 82 times in what is referred to as the *Regular Season*. Half of them in their own stadium and half of them away. After the reglementary 82 games, the best 8 teams of each conference enter a postseason championship: the *Playoffs*.

This championship consist of several eliminatory rounds. 2 teams face each other and the best of 7 games passes the round. The final winners of each Conference meet in the best-of-seven NBA Finals [9]. This paper focuses on the Regular Season, as the data is richer and the Playoffs constitute another championship in itself, with different incentives and strategies.

To ensure parity and competitiveness between teams, the NBA established a *"Salary Cap"* in the early 1980s. This wage cap is a monetary threshold decided every year that dictates the maximum payroll a team can spend on players's salaries. Every team that spends more than the salary cap on player's salary have to pay heavy taxes ranged from 50% to 400% for every dollar spent over the salary cap [6].

Because of this constraint, the team's Presidents and General Managers have to be meticulous hiring players. They have to engineer a team that can perform efficiently and is balanced between offense and defense. The team design selection is fundamental for the success of the squad. Given the importance of the player selection process, this project paper purpose a data-solution for selecting and hiring players based on their efficiency.

2.2 First, Second and Third Unit Players

A basketball game consist of two teams of five players each, and the goal is to outscore the opponent team. Each of the five play different positions. On a typical team the five different positions will be:: Center (C), Power Forward (PF), Small Forward (SF), Point Guard (PG), and Shoot Guard (SG). Every team generally has 15 players in the roster, that rotate their time on court during the game. The first 5 players that accumulate more court-time are considered the first unit and the 5 players afterwards are considered the second unit. On average, the First Unit spent 60% of the playing time on the court last season. This is double the time the Second Unit spent on the court, which was 30%. (see Table 5 in the Appendix).

The least used players conform a Third Unit, that only play under special circumstances. Third Unit players may for instance play if the game is already over because of the points unbalance, because the team is trying to develop very young players, or just before the Playoff competition to the other players rest (10% playing time).

Due to this reason, the remaining players that do not conform the First and Second Unit have little or no impact in the performance of the team. The NBA, realizing this fact, allows teams to keep most of these contracts out of the salary cap. Some examples are the *Mid-level Exception* contract, *Two-Way* contracts or contracts with the affiliated team the Development League. Therefore, in this project I assume that every team consist of 10 players. While it simplifies the linear modeling, it is also not far from the real situation of the teams for the referred limited performance, minutes in court and no impact in salary cap constraint.

2.3 NBA Statistics and Feature selection

One the most used compound statistic to predict player performance is the Player Efficiency Rating. *PER*, which sums up the player's positive accomplishments, subtracts the negative accomplishments and returns a per-minute rating of a player's performance [4]. While this is one of the most successful attempts to describe in one single digit the overall performance of a player, the creator of the metric admits that it does not depict the defensive qualities of a player [5]. As a main weakness in this aspect, PER only accounts for blocks and steals as defensive accomplishments.

That fact undermines the defensive factor of wining. Taking data for the last 8 seasons show only a 56,25% predicting the ranking of the teams according to their PER [7]. For the 2018-2019 season, overall, the PER does not show an accurate prediction of the team win rate with a Pearson coefficient close to 0 (Table 4 in the Appendix).

However, when we cluster for the different Conferences and only take into account the most used players, PER starts to gain more correlation with winning rates. The correlation of the East Conference team ranking with the PER of the 5 most used players (FU), and the 10 most used players (FU+SU) by team is shown in Table 1 below:

Table 1: Expected Rank by Summed FU and FU+SU PER and Real Rank by W/L%

Team	W/L%	Real Rank	Expected Rank FU PER	Expected Rank FU+SU PER
Milwaukee Bucks	0.73	1.0	1.0	2.0
Toronto Raptors	0.71	2.0	3.0	6.0
Philadelphia 76ers	0.62	3.0	2.0	3.0
Boston Celtics	0.6	4.0	4.0	1.0
Indiana Pacers	0.58	5.0	8.0	8.0
Orlando Magic	0.51	7.0	5.0	7.0
Brooklyn Nets	0.51	7.0	7.0	12.0
Detroit Pistons	0.5	8.0	10.0	14.0
Charlotte Hornets	0.48	10.0	11.0	4.0
Miami Heat	0.48	10.0	13.0	9.0
Washington Wizards	0.39	11.0	6.0	5.0
Atlanta Hawks	0.35	12.0	14.0	11.0
Chicago Bulls	0.27	13.0	12.0	15.0
Cleveland Cavaliers	0.23	14.0	9.0	10.0
New York Knicks	0.21	15.0	15.0	13.0
Pearson Coefficient wr. W/L%		1	0.85	0.65

Despite the PER not being an accurate metric overall, it shows that the performance in the game of the First Unit (FU 5 PER) is more correlated with winning than the overall performance. The same observation is true for the West Conference. Intuitively, the efficiency of the First Unit is more correlated with winning precisely because this unit plays more and stays more on the court. Consequently, the model design of the project weighs the performance of the player differently depending on whether they belong to the First or Second Unit.

Another successful attempt is the Offensive Rating (*ORtg*) and Defensive Ratings (*DRtg*) metrics, developed by Dean Oliver [10]. At a team level, these metrics explains the points produced by 100 possessions (ORtg) and the points allowed by 100 possessions (DRtg):

$$ORtg_{Team} = 100 \times \frac{PointsProduced}{Possessions} \tag{1}$$

$$DRtg_{Team} = 100 \times \frac{PointsAllowed}{Possessions} \tag{2}$$

where:

$$Possessions = 0.5 \times (TeamPossession + OpponentPossesion) \tag{3}$$

$$\begin{aligned} TeamPossesion = {} & FGA_{Team} + 0.4 \times FTA_{Team} + TOV_{Team} \\ & - 1.07 \times \frac{ORB_{Team}}{ORB_{Team} + DRB_{Opponent}} \times (FGA_{Team} - FG_{Team}) \end{aligned} \tag{4}$$

$$\begin{aligned} OpponentPossesion = {} & FGA_{Opponent} + 0.4 \times FTA_{Opponent} + TOV_{Opponent} \\ & - 1.07 \times \frac{ORB_{Opponent}}{ORB_{Opponent} + DRB_{Team}} \times (FGA_{Opponent} - FG_{Opponent}) \end{aligned} \tag{5}$$

The formula of Offensive and Defensive Rating are shown in equations (1) and (2). Even though the idea of points per 100 possessions looks simple, calculating the possessions themselves is not straightforward. Total possession is calculated as the number of plays a team have. A team can finish a play attempting to score (Field Goal Attempts or *FGA*), going to the free-throw line after receiving a fault (Free Throw Attempt or *FTA*), losing the ball (Turnover or TOV) or losing a rebound after an attempt in one of the sides of the court (Offensive and Defensive Rebounds or *ORB/DRB*). Equations (4) and (5) weigh the different events mentioned for both the team and the opponent team respectively. Finally, total possessions are summarized in (3), which is the total of possessions used for calculating *ORtg* and *DRtg*.

At a player level the metric is more complex given how possession is measured at an individual level [1, 10], but the sense is the same: it measures points by possessions. In both team and player level, the higher the Offensive Rating the better the team or player's offense. The lower the Defensive Rating, the better the team or player's defense.

3 Modeling Strategy and Optimization

3.1 Team Model

A winning team must offset its offense and defense performance. The following linear model resemble a balance relationship between the efficiency of the team in both sides of the court:

$$Win\% = \beta_0 + \beta_1 \times ORtg_{team} + \beta_2 \times DRtg_{team} \tag{6}$$

To test this model (6), I have extracted the data from the difference between Team Offensive and Defensive Ratings (Net Rating) and Win Rate of every team from 2012 to 2019. In total, there are 210 data points. Figure 1 shows a strong linear relation(R^2= 0.92) :

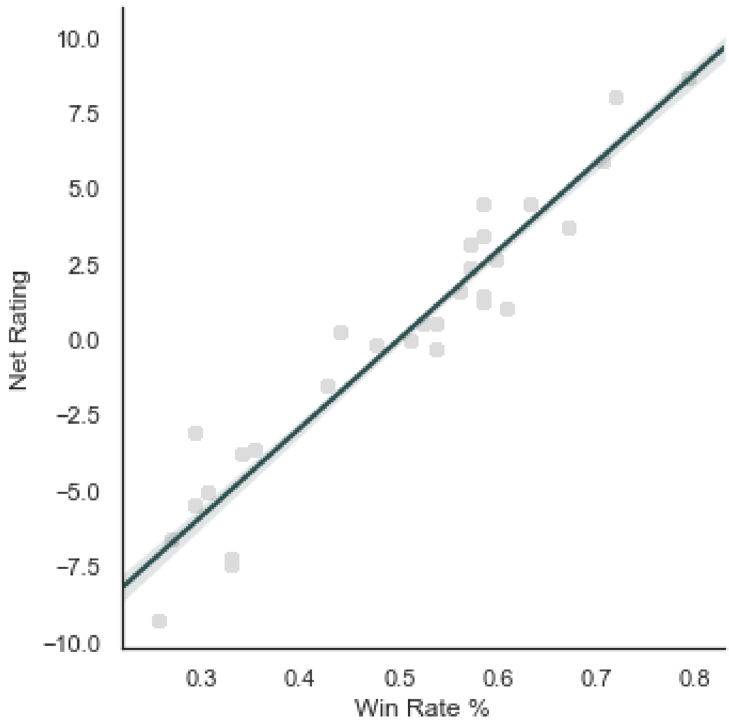

Figure 1: Team Offensive/Defensive Rating and Win Rate

Since a team wins by outscoring the opponent and Net Rating measures the outscoring every 100 possessions, the close relation is not surprising.

3.2 Players Model

Given that we have a model that "predicts" or explain winning rates, the first challenge is matching individual players' efficiency with the team's statistics. The strategy is picking players metrics that parallels the model (6), at individual level. We are interested in accurate player metrics that are correlated with win rates, since later we will use those to select the best players.

For this project paper I have taken Player Offensive Rating and Player Defensive Rating as the measuring units of player's efficiency. While *Team* ORtg/DRtg and *Player* ORtg/DRtg share the same name and goal (calculating efficiency) they are calculated differently, as briefly discussed in Section 2.2. The latest, attempts to capture the individual contribution of single players by their singular possession time.

For being able to reformulate the model with individual player's data, I needed to make the following assumptions: (i) Every team consist of 10 players, with a First Unit that plays more than the second; (ii) The First Unit of the teams have a greater influence on the game outcome; (iii) Teams can hire players without contract impediments all (players are free agents). Using the matching proposed, the team win rate described in (6) can be converted to the following model:

$$\begin{aligned} Win\% = \beta_0 + \beta_1 \times \sum FirstUnitORtg_{player} + \\ \beta_2 \times \sum SecondUnitORtg_{player} + \\ \beta_3 \times \sum FirstUnitDRtg_{player} + \\ \beta_2 \times \sum SecondUnitDRtg_{player} \end{aligned} \tag{7}$$

Other models with different specification and features were considered, but they have less predictive power (Models I.8 to I.11. Appendix).

3.3 Optimization

The aim of the paper is creating a team of 10 players that according to their performance and salary can reach the maximum possible win rate. The contribution of every player to the win rate is based on their time on court (First and Second Unit). This status is constant in the model, a Second Unit selected player will still contribute as a Second Unit player, and vice versa.

The key approach is selecting the coefficients, or feature weighs, of Model (7) and use them to design the optimization problem that selects the players according to the needs of the team.

3.3.1 Maximizing Win Rate under Salary Cap

This strategy aims to reach the maximum team winning percentage under financial constraints. The objective function captures Model (7), and the constraints reflect the salary cap, number of players, roles and positions.

Maximize

$$\beta_0 + \beta_1 \times \sum_{i=1th}^{5th} (S_i \times ORtg_i) + \beta_2 \times \sum_{i=6th}^{10th} (S_i \times ORtg_i) + \beta_3 \times \sum_{i=1th}^{5th} (S_{ii} \times DRtg_i) + \beta_4 \times \sum_{i=6th}^{10th} (S_i \times DRtg_i) \tag{8}$$

Subject to:

$$\sum_{i=1th}^{10th} (S_i \times Salary_i) \leq SalaryCap \tag{9}$$

$$\sum_{i=1th}^{5th} (S_i) \leq 5 \tag{10}$$

$$\sum_{i=1th}^{10th} (S_i \times Center) \leq 2 \tag{11}$$

$$\sum_{i=1th}^{10th} (S_i \times PointGuard) \leq 2 \tag{12}$$

$$\sum_{i=1th}^{10th} (S_i \times ShootGuard) \leq 2 \tag{13}$$

$$\sum_{i=1th}^{10th} (S_i \times PowerForward) \leq 2 \tag{14}$$

$$\sum_{i=1th}^{10th} (S_i \times SmallForward) \leq 2 \tag{15}$$

Player are sorted in First and Second Unit according to their minutes played for their teams, being the first to the fifth player i the First Unit and the sixth to the tenth the Second Unit. S_i describe the Selection Status of the player i, $ORtg_i$ the offensive rating of player i and $DRtg$ the defensive rating. $Salary_i$ reflect the player i salary and the $SalaryCap$ is a financial constant threshold. Finally, constraints (10) to (15) limit the positions and number of First Unit players the model can choose from.

Here, we set the highest Salary Cap possible to not incur in team taxation, but the model can be used to maximize performance under different financial constraints. For example, a team may want to just have enough victories to classify for Playoffs in order to save the team's money and prepare for the next season.

3.3.2 Minimizing Salary Cap

In this strategy, the aim is to get the minimum total team's payroll able to achieve a certain win rate.

Minimize

$$\sum_{i=1th}^{10th} (S_i \times Salary_i) \tag{16}$$

Subject to

$$\begin{gathered}\beta_0 + \beta_1 \times \sum_{i=1th}^{5th} (S_i \times ORtg_i) + \beta_2 \times \sum_{i=6th}^{10th} (S_i \times ORtg_i) + \\ \beta_3 \times \sum_{i=1th}^{5th} (S_i \times DRtg_i) + \beta_4 \times \sum_{i=6th}^{10th} (S_i \times DRtg_i) \geq Win\% \end{gathered} \tag{17}$$

$$\sum_{i=1th}^{5th} (S_i) \leq 5 \tag{18}$$

$$\sum_{i=1th}^{10th} (S_i \times Center) \leq 2 \tag{19}$$

$$\sum_{i=1th}^{10th} (S_i \times PointGuard) \leq 2 \tag{20}$$

$$\sum_{i=1th}^{10th} (S_i \times ShootGuard) \leq 2 \tag{21}$$

$$\sum_{i=1th}^{10th} (S_i \times PowerForward) \leq 2 \tag{22}$$

$$\sum_{i=1th}^{10th} (S_i \times SmallForward) \leq 2 \tag{23}$$

Again, we can modify the threshold $Win\%$ if the win rate goal is different. For example, a team may want just have enough victories to classify for Play-Offs in order to save team's money and prepare to the next season.

4 Analysis and Results

The data-set used contains all the current NBA players, their respective metrics and statistics, and their salaries from 2018 until 2024. The Team Model was trained with team data from the season 2012 until the season 2019. All the data was scrapped from www.basketball-reference.com using *BeautifulSoup* library, and double checked with data scrapped from NBA.com with *Selenium* library. Basketball-reference have data feeds from the league, and the variation in the basic statistics (for example, minutes played per game) was almost none. The data was scrapped because the official data from the proper NBA is not available for neither academic or personal use [9].

The library *sklearn-learn* was used for the different logistic and linear regression, and *PulP's* library for the optimization problem. All the programming code is provided in a *GitHub* public repository [3].

4.1 Players Model

I have found that the most correlated model with winning rates was the linear regression specified in Model (7), with a coefficient of determination of 0.90 between the win rate predicted and the real one. The coefficients are the following, used for the optimizing the player selection:

$$0.2953 + 0.00372 \times \sum FirstUnitORtg_{player} + 0.002056 \times \sum SecondUnitORtg_{player} \\ -0.00065 \times \sum FirstUnitDRtg_{player} - 0.00064 \times \sum SecondUnitDRtg_{player} \tag{24}$$

Other variations of the linear model were tried with less predictive power (I.27 - I.28. Appendix). The logistic regression(I.25 and I.26. Appendix) also predict with the same accuracy at decimal level, with different coefficients that give more importance to the offense side in comparison with the linear regression: $\beta_0 = 0.5831, \beta_1 = 0.00086, \beta_2 = 0.00048, \beta_3 = -0.00065, \beta_4 = -0.00066$. For all the models the training/test ratio was 70/30. I selected the most accurate model for the optimization problem, the Linear Model (7).

The coefficients are in conformity with the descriptive evidence: Offensive efficiency is more correlated with win rates than defensive one, and the offensive coefficient of the First Unit almost double the Second One as they play the double minutes (see Table 5. Appendix).

4.2 Optimization

4.2.1 Maximizing Win Rate under Salary Cap

Following the strategy, we make use of the coefficients to weigh the player selection optimization problem. For the maximization win rate problem, we set the Salary Cap constraint to the League actual Salary Cap of the last season: \$101.900.000. The table below shows the 10 players selected:

Table 2: Player Selection under Salary Cap to maximize win rate.

Player	Team	Possition	Role	ORtg	DRtg	Salary
Rudy Gobert	Utah Jazz	C	Fist Unit	133.0	100	\$23,241,573
Giannis Antetokounmpo	Milwaukee Bucks	PF	Fist Unit	121.0	99	\$24,157,304
Andre Iguodala	Golden State Warriors	SF	Fist Unit	124.0	109	\$16,000,000
Monte Morris	Denver Nuggets	PG	Fist Unit	124.0	111	\$1,349,383
Malcolm Brogdon	Milwaukee Bucks	SG	Fist Unit	121.0	108	\$1,544,951
Mitchell Robinson	New York Knicks	C	Second Unit	140.0	106	\$1,485,440
Dante Cunningham	San Antonio Spurs	PF	Second Unit	121.0	111	\$2,500,000

Table 2 continued from previous page

Player	Team	Possition	Role	ORtg	DRtg	Salary
Tony Snell	Milwaukee Bucks	SF	Second Unit	119.0	109	$10,607,143
George Hill	Milwaukee Bucks	PG	Second Unit	113.0	107	$19,000,000
Pat Connaughton	Milwaukee Bucks	SG	Second Unit	122.0	107	$1,641,000

The team win rate predicted is 96,66%, with a salary close to the maximum payroll allowed: $101.526.794. The model choose some expected players, like the most valuable player Giannis Antetokounmpo and the NBA Defensive Player of the Year for the last 2 seasons, Rudy Gobert.

The problem can be turned into a minimization problem to find the players that would get the worst win rate. This selection could be useful to detect not worth players, or evaluate their performance/salary ratio. The selection is provided in the appendix (Table 6). The list includes players that have left the NBA this year (Kosta Koufos or Wayne Selden), retired players (Dirk Nowitzki) or players that their team want to transfer due to bad performance (Antonio Blakeney).

4.2.2 Minimizing Salary Cap

For the minimization Salary Cap problem the win rate selected is the one of the last Regular Season Champion, the Milwaukee Bucks. This team won 75,98% of their matches. Therefore, we decided to create the cheapest team possible that can reach the 2018-2019 Milwaukee Bucks win rate. The following table present the players selection:

Table 3: Player Selection under Salary Cap to minimize win rate.

Player	Team	Position	Role	ORtg	DRtg	Salary
Thomas Bryant	Washington Wizards	C	First Unit	130.0	112	$1,378,242
Ryan Arcidiacono	Chicago Bulls	PG	First Unit	121.0	115	$1,349,383
Monte Morris	Denver Nuggets	PG	First Unit	124.0	111	$1,349,383
Malcolm Brogdon	Milwaukee Bucks	SG	First Unit	121.	108	$1,544,951
Maxi Kleber	Dallas Mavericks	PF	First Unit	115.0	109	$1,378,242
Mitchell Robinson	New York Knicks	C	Second Unit	140.0	106	$1,485,440
Bruno Caboclo	Memphis Grizzlies	SF	Second Unit	109.0	109	$670,211
Danuel House	Houston Rockets	SF	Second Unit	123.0	113	$94,003
Gary Clark	Houston Rockets	PF	Second Unit	108.0	110	$596,872
Wesley Matthews	Indiana Pacers	SG	Second Unit	106.0	111	$737,715

This selection of players would win 0.7598% according to the linear model, exactly the same win ratio of the last Regular Season leader. However, the payroll sum of the players is $10,584,442, 12 times less than the budget of the team with the same theoretical performance.

One way to see if the model is selecting the right players is to check what happened to these players at the end of the real season. If these players are performing above their same salary peers , we would expect that teams offer them better contracts. Indeed, 8 of these 10 players improved their salary substantially with new contracts and the 2 resting players could not because they are still under a limited rookie contract that prevents it.

Malcom Brogdon, a player that is selected in both optimization models, went from a $1,5 million contract per year to sign a multiannual-contract of more than $20 millions per season at the end of the season. Another example is Daniel House, whose salary has been multiplied 37 times. Overall, this composition of players signed today would be 5 times more expensive (Table 7. Appendix).

5 Standing on the shoulders of giants

This model idea is original by Ramya Nagarajan and Lin Li [8]. This project rest on their effort, but propose the following changes looking to improve the accuracy of the model. The modifications can make the model depict better the relation between team performance and player performance more accurately.

5.1 Measurement of First and Second Unit

In their paper [8], the First and Second Unit are defined as the "usual" starters and the "usual" bench of the team, respectively. Instead, I have used the 5 players with more total minutes played as the First Unit and the Second Unit as the next 5 with more minutes. This separation is made for calculating differently the weight or impact of the players on global teams performance. While the difference seems trivial, it avoid several problems:

i) Consistency. Starters and Bench players are not a constant role and change during the season due to performance. 25 of the 30 teams started and finished with different line-ups. Setting Starters and Bench involves inevitably choosing between players according to a criteria that should be well defined.

ii) Influence on court. Bench players influence the results more to than the starters, given that they can play more. This is the case for example for Lou Williams, an awarded 6th Man of The Year for 2 years in a row. He plays more minutes and have better statistics per possession than some starters of his own team. Therefore, he influences the game outcome more but he is better starting in the bench. The same is true for Montrezl Harrell, Eric Gordon or Andre Iguodala among other examples.

iii) Team Rank bias by injuries. Starters can get injured and this could limit his influence on court.

To clarify this point, we may analyze the case of Victor Oladipo. Victor Oladipo is a very high performance player (*All-Star Player*) playing for the Indiana Pacers. If we take him measured as Ramya Nagarajan and Lin Li [8], he is an starter and therefore his influence weighted more explaining the rank of his team. The problem is that he was injured during most of the season and therefore his statistics are not an indicative of how well Indiana Pacers performed.

If I measure team win rate in the same way, the model would be attributing Indiana Pacers an up-bias win rate because of Victor Oladipo, a starter that could not play as such most of the season. Measured by minutes played, we can see that Victor Oladipo is the 9th player ranked by minutes and his influence diminish and weighed accordingly as Second Unit.

5.2 Matching of the Parameters

These authors use PER as the individual offensive metric and the team average defensive rating for the individual defensive metric. I found that *Player* ORtg and *Player* DRtg are more correlated with *Team* ORtg and *Team* DRtg than the mentioned, and therefore better match. Using PER as the offensive measure instead of Player DRtg reduced accuracy of the model significantly (R^2= 0.62 against R^2= 0,90).

Also, in the paper they remark that making the team average the defensive rating was the only solution given the lack of defensive statistics the official NBA provide. The alternative match used in this project, have the drawback that they cannot be extracted from the official NBA site, that was probably the main reason of the authors of not using them.

5.3 Overall interpretation

While the specification by Ramya Nagarajan and Lin Li [8] provide high precision predicting win rates (R^2=0.845 reported), it also produce a Player Model with difficult interpretation:

$$\begin{aligned} Win\% = 2.040343 + 0.011658 \times \sum tmStarterPER + 0.00599 \times \sum tmBenchPER \\ -0.002605 \times \sum tmStarterDefRtg - 0.002851 \times \sum tmBenchDefRtg \end{aligned} \tag{25}$$

The resulting equation, extracted from the original paper, show an intercept more than twice the maximum win rate possible (1). Despite that, the coefficients indicate the same pattern as in the model used for these paper: Offensive efficiency is more correlated with win rates than defensive one, and the offensive coefficient of the Starters almost double the Bench players.

6 Conclusion and limitations

This project studied players' efficiency measures in order to predict the win rate of NBA teams in the Regular Season. A linear regression composed of the player offensive and defensive ratings per 100 possessions of the main 10 players can explain 90% of the variance between the different winning paths of every team.

The regression is used to tune the later linear optimization that selects the players among the 2018-2019 regular season to mainly achieve: i) Maximizing Win Rate under Salary Cap and ii) Minimizing Salary Cap under a win threshold.

For the first goal, the linear optimization generates a selection of 10 players that would construct a theoretical team with win rate close to 97%. For the second goal, the optimization model can construct a team that imitates the record of the 2018-2019 Regular Season winner, Milwaukee Bucks, with a 12 times reduction in the the team's total salary.

Within this theoretical framework, a team performs as well as the sum of the 10 players together. However, it is a limited tool for guiding team decisions as it does not take into account the synergies between the players, the effect if the coach's vision and strategies or long-run team formation perspective. Nevertheless, it is still useful for spotting undervalued players. It is surprising that almost all the players selected under the "Minimizing Salary Cap" optimization strategy have seen improvements in their contracts at the end of the season in real life.

It could be also a useful tool to complement experts decisions. A manager may for instance use it to search for players in the league that would increase their winning projection as much as possible, after she has already decided on which players to keep based on her own judgment (setting their statistics as constants in in the linear model). Overall, more complex and sophisticated prediction models would be necessary to drive team-creation design, but this modest setup was successful in to creating highly efficient teams.

7 Appendix

Table 4: Expected Rank by PER and Real Rank by W/L%

Team	PER	W/L%	Expected Rank	Real Rank
Boston Celtics	274.1	0.598	1.0	10.0
Atlanta Hawks	265.8	0.354	2.0	26.0
Washington Wizards	263.6	0.39	3.0	25.0
Philadelphia 76ers	263.3	0.622	4.0	7.0
Houston Rockets	258.1	0.646	5.0	6.0
Memphis Grizzlies	254.1	0.402	6.0	24.0
Minnesota Timberwolves	254.0	0.439	7.0	21.0
Phoenix Suns	240.6	0.232	8.0	29.0
Brooklyn Nets	237.9	0.512	9.0	15.0
Portland Trail Blazers	228.0	0.646	10.0	6.0
Toronto Raptors	227.7	0.709	11.0	2.0
New Orleans Pelicans	224.7	0.402	12.0	24.0
Golden State Warriors	224.5	0.695	13.0	3.0
Utah Jazz	221.1	0.609	14.0	8.0
Chicago Bulls	220.3	0.268	15.0	27.0
New York Knicks	216.9	0.207	16.0	30.0
Sacramento Kings	213.0	0.476	17.0	19.0
Milwaukee Bucks	211.8	0.732	18.0	1.0
Los Angeles Lakers	203.9	0.451	19.0	20.0
Indiana Pacers	203.3	0.585	20.0	13.0
San Antonio Spurs	202.4	0.585	21.0	13.0
Charlotte Hornets	201.2	0.476	22.0	19.0
Cleveland Cavaliers	200.8	0.232	23.0	29.0
Detroit Pistons	197.1	0.5	24.0	16.0
Denver Nuggets	196.7	0.659	25.0	4.0
Los Angeles Clippers	193.9	0.585	26.0	13.0
Dallas Mavericks	192.0	0.402	27.0	24.0
Miami Heat	188.4	0.476	28.0	19.0
Orlando Magic	184.0	0.512	29.0	15.0
Oklahoma City Thunder	177.0	0.598	30.0	10.0

Table 5: % Minutes Played On Court by the First, Second and Third Unit.

Team	% First Unit	% Second Unit	% Third Unit
Cleveland Cavaliers	0.65	0.25	0.1
Toronto Raptors	0.64	0.31	0.05
Washington Wizards	0.63	0.25	0.12
Boston Celtics	0.55	0.37	0.07
Chicago Bulls	0.53	0.31	0.16
Miami Heat	0.61	0.37	0.02
Indiana Pacers	0.58	0.33	0.09
Brooklyn Nets	0.52	0.32	0.16
Charlotte Hornets	0.57	0.28	0.15
Orlando Magic	0.64	0.3	0.06
New York Knicks	0.53	0.32	0.16
Milwaukee Bucks	0.59	0.31	0.1
Atlanta Hawks	0.52	0.36	0.12
Detroit Pistons	0.61	0.28	0.1
Philadelphia 76ers	0.7	0.23	0.07
Dallas Mavericks	0.63	0.26	0.11
Denver Nuggets	0.55	0.38	0.08
Golden State Warriors	0.59	0.3	0.12
Houston Rockets	0.66	0.28	0.07
Los Angeles Clippers	0.69	0.2	0.1
Los Angeles Lakers	0.55	0.36	0.09
Memphis Grizzlies	0.55	0.25	0.2
Minnesota Timberwolves	0.53	0.33	0.14
New Orleans Pelicans	0.53	0.29	0.18
Oklahoma City Thunder	0.66	0.26	0.08
Phoenix Suns	0.55	0.3	0.16
Portland Trail Blazers	0.59	0.34	0.08
Sacramento Kings	0.65	0.29	0.06

Table 5 continued from previous page

Team	% First Unit	% Second Unit	% Third Unit
San Antonio Spurs	0.58	0.37	0.05
Utah Jazz	0.61	0.3	0.09
Mean among Teams	0.593	0.303	0.104

Alternative models:

- Logistic Regression with the same features specifications ($R^2 = 0.90$):

$$\begin{aligned} y = \beta_0 + \beta_1 \times \sum FirstUnitORtg_{player} + \\ \beta_2 \times \sum SecondUnitORtg_{player} + \\ \beta_3 \times \sum FirstUnitDRtg_{player} + \\ \beta_2 \times \sum SecondUnitDRtg_{player} \end{aligned} \tag{I.26}$$

Where:

$$Win\% = \frac{1}{1 + e^{-y}} \tag{I.27}$$

- Linear Regression with different features specifications ($R^2 = 0.78$):

$$\begin{aligned} Win\% = \beta_0 + \beta_1 \times \sum FistAndSecondUnitORtg_{player} + \\ \beta_2 \times \sum FistAndSecondUnitDRtg_{player} \end{aligned} \tag{I.28}$$

- Linear Regression with different features ($R^2 = 0.62$):

$$\begin{aligned} Win\% = \beta_0 + \beta_1 \times \sum FirstUnitPER_{player} + \\ \beta_2 \times \sum SecondUnitPER_{player} + \\ \beta_3 \times \sum FirstUnitDRtg_{player} + \\ \beta_2 \times \sum SecondUnitDRtg_{player} \end{aligned} \tag{I.29}$$

Table 6: Player Selection under Salary Cap to minimize win rate.

Player	Team	Position	ORtg	DRtg	Salary
Élie Okobo	Phoenix Suns	PG	94.0	117	\$1,238,464
Wayne Selden	Chicago Bulls	SG	94.0	115	\$1,544,951
Chandler Hutchison	Chicago Bulls	SF	101.0	114	\$1,991,520
Antonio Blakeney	Chicago Bulls	SG	94.0	116	\$1,349,383

Table 6 continued from previous page

Player	Team	Position	ORtg	DRtg	Salary
Bobby Portis	Washington Wizards	PF	104.0	112	$2,494,346
Kosta Koufos	Sacramento Kings	C	101.0	108	$8,739,500
Mario Hezonja	New York Knicks	SF	94.0	112	$6,500,000
Frank Ntilikina	New York Knicks	PG	88.0	115	$4,155,720
Dirk Nowitzki	Dallas Mavericks	PF	96.0	111	$5,000,000
Ante Žižić	Cleveland Cavaliers	C	116.0	117	$1,952,760

Table 7: Next Contract of Player Selection under win threshold to achieve minimum salary.

Player	Salary Season 2018-2019	Salary Season 2019-2020
Malcolm Brogdon	$1,544,951	$22,600,000
Ryan Arcidiacono	$1,349,383	$3,000,000
Thomas Bryant	$1,378,242	$8,000,000
Maxi Kleber	$1,378,242	$8,000,000
Monte Morris	$1,349,383	$1,588,231
Bruno Caboclo	$670,211	$1,882,867
Wesley Matthews	$737,715	$2,564,753
Mitchell Robinson	$1,485,440	$1,559,712
Danuel House	$94,003	$3,540,000
Gary Clark	$596,872	$1,416,852

References

[1] Basketball-reference.com. Calculating individual offensive and defensive ratings. https://www.basketball-reference.com/about/ratings.html.

[2] Forbes.com. Nba team values 2019. https://www.forbes.com/sites/kurtbadenhausen/2019/02/06/nba-team-values-2019-knicks-on-top-at-4-billion/, Feb. 2019.

[3] Galera, F. Optimizing nba player selection. https://github.com/pipegalera/Optimizing-NBA-Player-Selection.

[4] Hollinger, J. *Pro Basketball Forecast, 2005-06.* Potomac Books, 2005.

[5] Hollinger, J. What is per? https://www.espn.com/nba/columns/story?columnist=hollinger_john&id=2850240, Aug. 2011.

[6] Hoopshype.com. How does the nba's luxury tax work? https://hoopshype.com/2018/10/11/nba-luxury-tax/.

[7] Hsu, P.-H., Galsanbadam, S., Yang, J.-S., and Yang, C.-Y. Evaluating machine learning varieties for nba players' winning contribution. In *2018 International Conference on System Science and Engineering (ICSSE)* (2018), IEEE, pp. 1–6.

[8] Nagarajan, R., and Li, L. Optimizing nba player selection strategies based on salary and statistics analysis. *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)* (2017), 1076–1083.

[9] NBA.com. Nba frequent asked question. https://www.nba.com/news/faq.

[10] Oliver, D. *Basketball on paper: rules and tools for performance analysis.* Potomac Books, Inc., 2004.