Moneyball for the NBA: An Ensemble Approach

Charles Zimmerman

05/10/2021

I. Overview/Executive Summary

The term “Moneyball” refers to the use of statistical analysis in professional sports to assemble a team of players based on their overall value, i.e., their contributions to winning relative to the money spent to acquire and retain the players. Moneyball strives to identify undervalued players, who are predicted to help the team win even though the cost to acquire them is relatively low. The strategy was famously used by the Oakland Athletics baseball team in the early 2000s, which posted a win-loss record of 103-59 in the year 2002—the second best record in baseball with the third lowest payroll. Michael Lewis’ book Moneyball and a movie starring Brad Pitt based on this book helped make this team and the architect of the stategy, Billy Beane, famous.[1]

This project will attempt to identify a possible Moneyball approach for the National Basketball Association (NBA), the American professional basketball league. It will execute and compare an ensemble of five different statistical models for predicting success in the NBA at the team level, and then extrapolate those models to individual players. In order to compare teams to players, player statistics will be transformed to project how the player would perform as if he were the entire team—i.e. if he played every minute and if every player on the team had the same statistical characteristics.

If the models relied solely on equating team statistics to player statistics, they would likely fail, because there are cases where values within statistical categories could have different implications for a team vs. a player. Two examples:

- A team that makes a high percentage of its shots (high field goal percentage) is probably a good team. But players might have a high field goal percentage because they shoot infrequently and only when they are very close to the basket. Such players do not add as much value to a team compared to others with more well-rounded skills.
- Shot attempts per game could also be a better indicator of team effectiveness than player effectiveness. If a team takes many shots per game, it usually means they are getting many opportunities to score, limiting turnovers, and rebounding well. For

[1] Lewis, Michael (2003). *Moneyball : the art of winning an unfair game.* New York : W.W. Norton

> players, many shots per game, is sometimes a negative indicator, if the player has a low shooting percentage or prefers to shoot rather than pass the ball to an open teammate.

For this reason, the models to select "moneyball" players, though developed based on team statistics, were validated by confirming they accurately identified good players. Specifically, if a model did not rank four players known to be elite superstar players highly, it was discarded.

II. Methods/Analysis

The entire project was developed in the R programming language, using version 4.0.3 of R and version 1.3.1903 of R Studio as the IDE.

A. Data Sourcing

Three data sources were used: one for team statistics, one for player statistics and one for player salaries. The team statistics data, for seasons 2000-2001 through 2018-2019, was originally sourced from https://www.kaggle.com/mharvnek/nba-team-stats-00-to-18, Player stats for the season 2018-2019 were originally sourced from https://www.kaggle.com/mharvnek/nba-team-stats-00-to-18, and Player salaries were for 2018-2019 were sourced from http://www.espn.com/nba/salaries/_/year/2019/.

Minimal manual cleanup was done, such as standardizing statistical category names in the team and player stats file and the player names in the salary and player stats file. The three .csv files were then uploaded to the author's github site, from where the code for the model development accesses the data.

The 2018-2019 season was chosen because that was the last "normal" NBA season as of this writing. The 2019-2020 season was shortened and partially played only in one location due to COVID-19 restrictions, and the 2020-2021 season, in addition to also being modified due to COVID, was not complete at the time this study was conducted.

Below displays the statistics used for the analysis. All except the percentages represent per-game averages, and except where noted, all were evaluated as predictors for both the team and player data sets.

Abbreviaton	Description
WPct	Winning Percentage (predicted outcome, team stats only)
PTS	Points
FGM	Field Goals Made
FGA	Field Goals Attempted
FGPct	Field Goal Percentage
ThreePM	Three Point Shots Made
ThreePA	Three Point Shots Attempted

Abbreviaton	Description
ThreeP_Pct	Three Point Shot Percentage
FTM	Free Throws Made
FTA	Free Throws Attempted
FT_Percent	Free Throw Percent
ORB	Offensive Rebounds
DRB	Defensive Rebounds
TRB	Total Rebounds
AST	Assists
TOV	Turnovers
STL	Steals
BLK	Blocks
PF	Personal Fouls

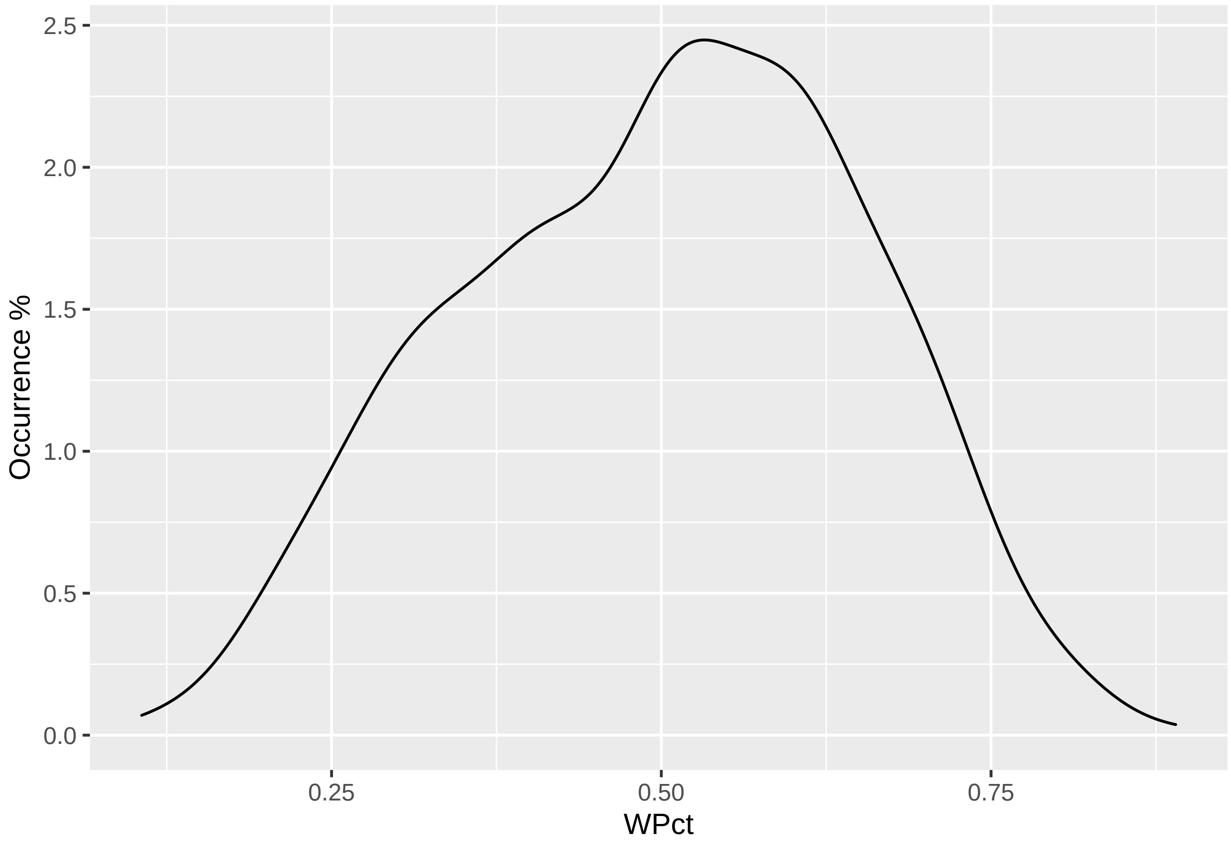
B.Data Preparation

Additional data cleanup and preparation was handled via code. Tasks included:

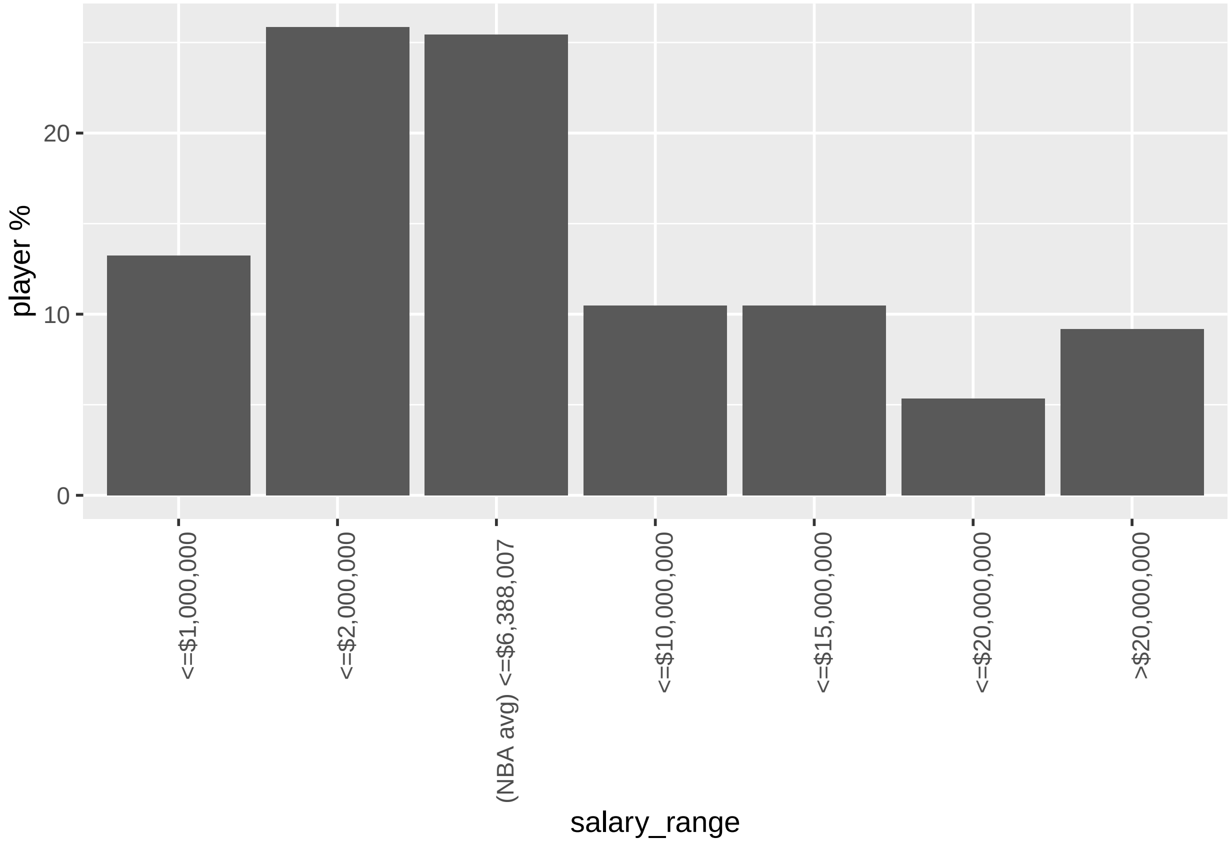
- Separating Player Name and Player Id (an alphanumeric unique identifier) in the player data from one column into two.
- Dropping un-needed columns that provided no meaningful information from the player data.
- Filtering out player records that were not present in both the salary and the player stats data. Players who received a salary but did not play in the 2018-2019 season might be present in the salaries but not stats data. Some players who played fewer than 20 games (they would have signed temporary, short-term contracts) were missing from the salary data.
- Removing duplicate records from the player stats data. Players who played on more than one team during the season were represented at least 3 times—one for each team they played for and a row summing their totals for the season. In these cases, the totals were retained and the data for single teams were discarded.
- Filtering out players who did not play at least 15 minutes per game, those who did not play at least 25 games out of the 81 in the season, and those who did not earn at least $838,464, the minimum salary for the 2018-2019 season. Players who earned less than that would have been playing on a partial season contract.
- Adding the "player-as-team" transformation. First the overall average minutes per NBA game was calculated—a regulation NBA game is 48 minutes but the overall average is slightly higher due to tie games going to overtime. Player totals for each statistical category were multiplied by that overall average divided by the player minutes per game, so as to project the statistic out for the entire game. Then that result was multiplied by five (the number of team players on the court at one time), so as to model a whole team as that one player. The below code illustrates, MP=Player minutes per game:

```
nba_player_stats_salaries_adj<-nba_player_stats_salaries %>%
    filter(MP>=15 & Salary >nba_min_salary_2018_2019 & G>25) %>%
  mutate(PTS=PTS*(avg_mins_game/MP)*5,FGM=FGM*(avg_mins_game/MP)*5,
         FGA=FGA*(avg_mins_game/MP)*5,ThreePM=ThreePM*(avg_mins_game/MP)*5,
         ThreePA=ThreePA*(avg_mins_game/MP)*5, FTM=FTM*(avg_mins_game/MP)*5,
         FTA=FTA*(avg_mins_game/MP)*5, ORB=ORB*(avg_mins_game/MP)*5,
         DRB=DRB*(avg_mins_game/MP)*5,TRB=TRB*(avg_mins_game/MP)*5,
         AST=AST*(avg_mins_game/MP)*5,TOV=TOV*(avg_mins_game/MP)*5,
         STL=STL*(avg_mins_game/MP)*5,BLK=BLK*(avg_mins_game/MP)*5,
         PF=PF*(avg_mins_game/MP)*5)
```

The final result was one dataset for teams and another one for players, both with statistics. The below shows the distribution of team winning percentages in the team data, for the 566 records:



And below shows the salary distribution within ranges, for the 303 players included in the analysis.



– Finally, all of the statistics were transformed to standard (z) scores with a mean of 0 and a standard deviation of 1, so that statistics with different ranges of values could be more accurately used for their purpose as inputs to the model.

C. Models

The below describes the prediction methods used for the analysis. All of these methods, use cross-validation, or partitioning the data into subsets and using each subset as a validation sample, in the algorithms to generate their resuts, meaning that in no cases was a custom-developed cross validation step required.

1. Stepwise Regression: Stepwise regression involves adding and removing predictors iteratively in order to identify the subset of predictors that produces the model with the lowest error.[2] The R caret's package implementation of the algorithm was used, configured to include five cross-validation samples.

2. Simple linear multiple regression: The second model, simple linear regression, was formulated based on a review of the correlations between predictors and team winning percentage, as well as between the predictors themselves in both the team and player data set. The goal was to choose variables that made sense while maximizing predictor correlations

[2] Kassambara, Alboukade (2018) *Stepwise Regression Essentials in R.* http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/

with team winning percentage while minimizing predictor correlations with each other. The below shows the correlations for the variables chosen with team winning percentage, for team data, and predictor correlations with each other for both player and team data:

```
## TEAM

##                WPct     AST   FGPct ThreeP_Pct     DRB
## WPct       1.00000 0.34793 0.57928    0.48188 0.33947
## AST        0.34793 1.00000 0.50514    0.29525 0.45702
## FGPct      0.57928 0.50514 1.00000    0.54816 0.31026
## ThreeP_Pct 0.48188 0.29525 0.54816    1.00000 0.20866
## DRB        0.33947 0.45702 0.31026    0.20866 1.00000

## PLAYER

##                  AST     FGPct ThreeP_Pct       DRB
## AST        1.000000  0.018666  0.2009477 0.3787346
## FGPct      0.018666  1.000000 -0.2214835 0.3967037
## ThreeP_Pct 0.200948 -0.221483  1.0000000 0.0015706
## DRB        0.378735  0.396704  0.0015706 1.0000000
```

3. Partial Least Squares Regression: Partial Least Squares Regression (PLSR) uses Principal Components Analysis to group predictor variables into linear combinations based on their correlations with the outcome variable and each other, and uses these combinations (or components, or latent variables) as predictors.[3] The R pls package was used for this prediction method.[4] Two predictions were generated, one using the first component (the one accounting for the most variability in the outcome) only, and one using the first two components.

4. Ridge Regression: Ridge regression is a form of regularized regression that penalizes or shrinks the coefficient weights for predictors if they are too large, by a penalty equivalent to square of the magnitude of the coefficients.[5] The R glmnet package's implementation was used.[6]

5. Lasso Regression: Lasso regression is also form of regularized regression that penalizes or shrinks predictor coefficients. It is similar to Ridge regression, but its penalty term is based on the the absolute sum of the coefficients.[7] The R glmnet package was used.

[3] Kassambara, Alboukade Principal (2018) *Component and Partial Least Squares Regression Essentials.* http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/152-principal-component-and-partial-least-squares-regression-essentials/

[4] Helge Mevik, Bjorn;Wehrens, Ron; Liland, Kristian Hovde; Hiemstra, Paul (2020). *Package 'pls': Partial Least Squares and Principal Component Regression.* https://cran.r-project.org/web/packages/pls/pls.pdf

[5] Bhattacharyya, Saptashwa (2018): *Ridge and Lasso Regression: L1 and L2 Regularization.* https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b

[6] Friedman, Jerome, et. al.: *Package 'glmnet:' Lasso and Elastic-Net Regularized Generalized Linear Models.* https://cran.r-project.org/web/packages/glmnet/glmnet.pdf

[7] Bhattacharyya, 2018.

All five of the methods were evaluated in exactly the same way:

- The models were fitted and used to generate predictions of the teams' scaled winning percentage.
- The residual mean square error (RMSE) for these predictions was calculated.
- A second set of predictions was run against the player data set, using the "player as team" statistics as inputs and the predicted scaled winning percentage as the output. In this case, residuals could not be calculated because "player as team winning percentage" is a made-up statistic which could not be compared to an existing one.
- Instead, to roughly estimate the model's accuracy for these predictions, the predicted results were appended as a column to the player data set and then ordered by those results. The ranking was added as a column, and results were reviewed for four elite players: Giannis Antetokounmpo, Lebron James, Stephen Curry, and James Harden. All won the league Most Valuable Player (MVP) award at least once in the 2010s. At least one of the teams Curry and James played for appeared in all 9 of the championship series' preceding and including the 2018-2019 season (James' team played in 8, Curry's team in 5). Harden and Antetokounmpo are superstar players; the latter was the league MVP in 2018-2019 and the former in 2017-2018. In sum, any model that that does not rank these players near the top would be inappropriate for evaluating a player's contribution to winning.

The ranking for these players was reported individually and summed to create statistic that could be used to compare the models, with the best possible score being 10 (1+2+3+4), the lower the better.

The code below illustrates what was done for each model, it is for the second model evaluated, but the same process flow was followed in each case:

```
#specify 5 cross validation folds
train.control <- trainControl(method = "cv", number = 5)
#train the model
theory_fit <- train(WPct ~ FGPct+ThreeP_Pct+DRB+AST ,
                    data = nba_team_stats_for_modeling,
                 method="lm",
                    trControl = train.control)
#predict winning pct for team
Y_hat_theory_fit <- predict(theory_fit,nba_team_stats_for_modeling)
#check rmse
theory_rmse<-RMSE(nba_team_stats_for_modeling$WPct, Y_hat_theory_fit)
#predict player stats
Y_hat_player_theory_fit <-predict(theory_fit,nba_player_stats_salaries_adj)
#generate ranking statistic used to validate model
theory_player_ranking<-nba_player_stats_salaries_adj %>%
  cbind(WinPctAsTeam=Y_hat_player_theory_fit)%>%
  arrange(desc(WinPctAsTeam)) %>% mutate(rank=row_number()) %>%
  select(Player, WinPctAsTeam,rank)%>%
  filter(Player %in% c("James Harden", "Stephen Curry",
  "Giannis Antetokounmpo","LeBron James")) %>%
  summarize(sum(rank))
```

Finally, the best model was chosen based on the RMSE and the rank statistic, though the latter was given priority, with the RMSE used as a tiebreaker. The moneyball team was then selected, one point guard, one shooting guard, one small forward, one power forward and one center. Players earning over the average salary in the 2018-2019 season of $6,388,00710[8] were excluded from consideration. The following snippet of code illustrates:

```
  #mutate(worth=WinPctAsTeam/as.numeric(Salary))  %>% arrange(desc(worth)) %>%
  #filter(Pos  %in% c("PG")& (Salary<  avg_salary_2018_2019)) %>% slice(1) %>%
#pull(Player)
```

The worth of all players is derived based on the ratio of his predicted winning percentage as a team to his salary and the player with the highest value at each position was selected.

III. Results

The below shows the results for the two statistics computed for each model, residual mean square erroring in predicting team win percentage, and sum of the rankings of selected elite players, based on their predicted "player-as-team" win percentage.

Model	RMSE	Player Ranking Score
Stepwise Regression	0.59510	886
Simple Multiple Regression	0.77433	276
PLSR (2 components)	0.61994	270
PLSR (1 component)	0.79353	16
Ridge Regression	0.99531	16
Lasso Regression	0.76028	579

Partial Least Squares Regression, one component, and Ridge regression performed equally well in terms of ranking elite players as such. Both ranked the elite players as follows:

Player	rank
Giannis Antetokounmpo	2
James Harden	3

[8] Heavy.com (2019): *NBA Players' Salary: How Much Money Is the Average?* https://heavy.com/sports/2019/05/nba-player-average-salary-how-much-highest-paid/

Player	rank
Stephen Curry	4
LeBron James	7

Since the PLSR error in predicting team success was lower than the error for Ridge Regression, PLSR with one component was used to generate the MoneyBall team.[9] The players chosen by position are:

Position	Player	2018-2019 Salary
Point Guard	Derrick Rose	$1,512,601
Shooting Guard	Malcolm Brogdon	$1,544,951
Small Forward	Kelly Oubre Jr.	$3,208,630
Power Forward	Pascal Siakam	$1,544,951
Center	Thomas Bryant	$1,378,242

IV. Discussion

A. Player Selection

The sum of the salary for these players in 2018-2019 was $9,189,375 vs. the average salary at this time for five players of $31,940,035, so if these players are as good as the model suggests, they appear to represent great bargains and be true MoneyBall players.

Below is a brief review of the players' career since 2018-2019 with thoughts about how well the prediction model fared in light how they have been playing and their salaries since that time.[10]

Derrick Rose

Rose has been chosen for the NBA all-star team multiple times and was the NBA's MVP for the 2010-2011 season. His salary was low in 2018-2019 because he was struggling to comeback from multiple injuries. 2018-2019 was a turnaround year for him; his minutes played per game and performance in all other statistical categories shot up that year. He continued to play well in 2019-2020 and 2020-2021 and is now a key player on the New York Knicks, a team whose performance in 2020-2021 is better than it has been for at least 20 years. His yearly salary is currently 7.3 million, lower than the average of 8.32 million for 2020-2021. He seems very much the type of player the Moneyball method is designed to identify.

[9] A follow up review confirmed that the same team would have been picked using the ridge regression result.

[10] Salary and statistical information obtained from https://www.basketball-reference.com/. 2020-2021 statistics are as of May 7, 2021.

Malcom Brogdon

Brogdon played for the Milwaukee Bucks in 2018-2019. After that season, he was signed by the Indiana Pacers as a free agent. In two seasons with the Pacers, his minutes and his statistical performance have noticeably gone up. For instance, he averaged 28.6 minutes and 15.6 points per game in 2018-2019; in 2020-2021, those numbers are up to 34.5 minutes and 21.2 points. His current salary of $20,700,000, perhaps suggests he is too expensive to be considered a Moneyball player. However, Moneyball players are also those whom statistical techniques suggest could contribute more to their teams if they are allowed more playing time, and based on his performance since 2018-2019, he fits the bill in this regard.

Kelly Oubre Jr.

Kelly Oubre Jr. played for the Phoenix Suns in 2018-2019 and the following season; currently he plays for the Golden State Warriors. He is relatively highly regarded, but his 2020-2021 statistics do not show improvement relative to 2018-2019. He averaged 16.9 points and 29.5 minutes per game in 2018-2019. Those numbers improved in 2019-2020 when he still played for the Suns, but in 2020-2021, he averages 15.4 points per game in 30.7 minutes per game as of this writing. This drop-off could be attributable to playing with Stephen Curry, who is the go-to scorer on the Warriors, but his statistics in other categories also don't show improvement. He currently earns $14,375,000. He is a a young player with lots of potential, but the model's selection of him as a Moneyball player is not supported by the statistics.

Pascal Siakam

Pascal Siakam emerged as an all-star and elite player in 2018-2019 and was a key member of the Toronto Raptors team that won the NBA championship that year. He currently earns $29,000,000 per year. His selection is a good sign that the model recognizes good players, but also that it probably requires tweaking to exclude players who improve to his degree during the season their performance is evaluated.

If the model were revised to exclude Siakam, it would have selected John Collins as the power forward. Collins has played on the Atlanta Hawks for his whole career, starting in 2017-2018 until now. He is averaging 7.5 rebounds and 17.6 points per game this season as of this writing, in 29.6 minutes per game. Though these numbers are down slightly from 2018-2019, his current salary of $4,137,302, up from $2,299,080 in 2018-2019, is much lower than Siakam's. Though he is not as good a player as Siakam, he may be a better value for the money, and it seems appropriate to classify him as a Moneyball player.

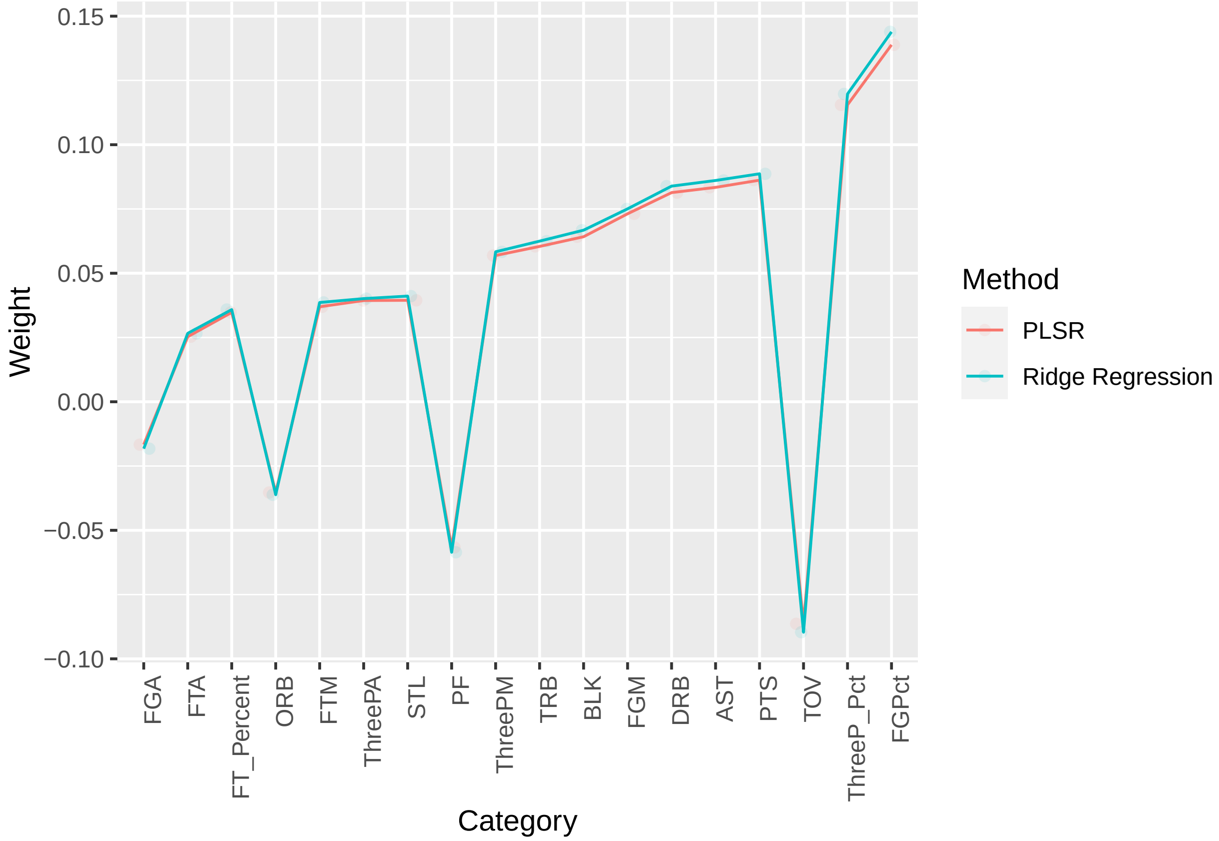
Thomas Bryant

Thomas Bryant, center for the Washington Wizards missed time in the 2019-2020 season due to a COVID diagnosis, and only played 10 games in 2020-2021 before suffering a season ending injury. But the model is not designed to predict illness or injury and based on his statistics and salary, it appears that it made a superb choice in this case. His minutes per game in 2019-2020 and 2020-2021 were 24.9 and 27.1, up from 20.8 in 2018-2019. His scoring average for the three seasons went from 10.5 to 13.2 to 14.3. Perhaps most impressively, his overall shooting percentage combined was 59.3% in 2019-2020 and 2020-2021, and his three point shooting percent during this time was 41.1%. Bryant is 6-10 and the model, rightly so, particularly favors big centers who can score from anywhere on the court including three-point range, such as Anthony Davis and Giannis Antetokounmpo. His 2020-2021 salary is $8,333,333, right at the league average

B. Theoretical Implications

As noted in the introduction, when it comes to evaluating basketball statistics in terms of what they mean for players vs. what they mean for teams as a whole, there is some overlap and some difference. The results in this study bear this notion out. Specifically, two separate forms of regression analysis independently supported the notion that models partially accounting for variability in team winning percentage might do very well in predicting player value.

These methods, Partial Least Square Regression and Ridge Regression, both arrived at similar player rankings and similar relative variable weightings for predicting the score used to define player value, the made-up "player as team winning percentage." The weightings, which reflect the relative importance of predictor variables, are displayed below (note: variable importance is reflected by the absolute value of the weighting; weights for ridge regression were multiplied by 100 for comparison purposes).



In other words, in addition to helping to identify players who add value to teams, this technique might be useful to some degree in identifying the relative importance of specific statistical categories, or combinations of them, in terms of how well they reflect a player's contribution to team success.

C. Conclusion

Sabermetrics, or the empirical analysis of sports performance, is no longer a new field, and it has been used in basketball to weight and combine different statistical categories to assess the value of players. For instance, the player efficiency rating statistic combines and weights performance across categories to measure a player's per-minute performance.[11] This study suggests that a statistical model identifying factors contributing to team success, extrapolated to players, might complement existing techniques used in assembling players. Such a model might assist teams both in identifying players who would emerge as great contributors if they were allowed more playing time as well as those with a high skills to salary ratio.

The results, though promising, suggest that the model developed here could be improved further; for instance, perhaps it can be implemented with parameters so users can configure it to focus on players who are undervalued in terms of their minutes played as opposed to their salary, or vice versa; or to only consider veteran players, as younger players are more likely to expect higher salaries as their career progresses. But in light of the players selected for the Moneyball team, the results suggest that sabermetric approaches, whether the specific one evaluated here or similar ones, can assist teams in their monetary investment and player selection decisions,

[11] Greenberg, Neil *What is Player Efficiency Rating?* The Washington Post,04/13/2017. accessed from https://www.washingtonpost.com/what-is-player-efficiency-rating/37939879-1c08-4cfa-aff3-51c2a2ae060e_note.html