Northwestern MSDS-460 Final Project

Major League Baseball Roster and Lineup Optimization

Group 5

Andrew Stevens, Alexander Wang, Joshua Roberts, Kris Heller

MSDS 460: Decision Analytics

November 7, 2021

Abstract

Major League Baseball strategy has been revolutionized by the application of statistics and simulation. Among the potential pursuits is workforce optimization applied to a roster to maximize the success of the team over a given season. The 2016 Chicago Cubs season serves as our target set, taking into consideration the career performance statistics of each player to select a lineup. The objective function is based on "Wins Above Replacement" with constraints developed to simulate realistic conditions, such as maximum games played, fielding entire teams for each game, and pitcher rotations. The resulting program produces a lineup for each game of the season.

Keywords: Sabermetrics, WAR, roster optimization, salary cap, workforce analytics

Introduction

Optimization is a method to find solutions that maximize or minimize some parameters. This method is used all around us in our daily lives, such as in GPS systems, financial companies, scheduling, etc. One area where optimization has been of increasing interest over the years is in sports (Duarte., Ribeiro, Urrutia, & Haeusler, 2007). A subclass of increasing interest is SABRmetrics (Lewis, 2003). Sabermetrics is the mathematical analysis of baseball statistics that measure in-game activity.

Our group is particularly interested in how to combine sabermetrics and linear programming to optimize a MLB baseball roster over the course of a season. To do this, we focused on the 2016 Chicago Cubs. Our problem statement for the 2016 Chicago Cubs was: What is the best possible roster, and the best lineup from game to game, when taking into

consideration constraints of innings pitched and games played (due to injury, fatigue, rest, etc.), with the goal of maximizing total team Wins Above Replacement (WAR)?

The statistic WAR is the measure of how much better or worse a player is compared to the average replacement player. WAR is generally regarded as one of the most comprehensive statistics and its value is said to reflect the number of additional wins the individual player contributes to their team over a replacement level player.

For the MLB and its teams, the league is not only a game, but a business too. There is a luxury tax that is charged to teams that spend over a set limit, and this tax rate climbs with each successive season that a team stays over the total team salary threshold. Because of this, there is a limit that many teams are willing to spend on their player salaries. Some small market teams take it a step even further by seemingly trying to spend as little as possible on total player salaries for their team.

In recognition that total team salary is an important factor in roster construction for a team, our research and analysis looked at multiple objectives. As stated before, our main objective was to field the best team with the goal of maximizing total team WAR, while setting a threshold for total team salary to not exceed.

Literature review

With analytics and sabermetrics increasing in popularity in professional baseball, there have been numerous studies attempting to better optimize baseball operations. There have been numerous attempts at modelling an optimized roster including the various constraints that an MLB organization may run into. Greene and Hirsch (2011) perhaps use the most similar method, but do not explore the number of games each player will play in. They apply workforce analytics

and calculate the efficiency of each player based on their WAR, expected playing time, and salary. One major oversight to this study is the exponential cost and length of star players' contracts. Greene's study, however, does a great job showing the relative efficiency of each player in the roster. This study aims to address these issues by using the available salary and current contracts as constraints as well as using WAR as a major deterministic variable in the problem.

Other such methods do not use WAR statistics and incorporate models to simulate or represent player performance. With WAR being the most important variable in determining performance for this study, using some other model or metric can alter the results drastically. Schorsch and Valera (2015) perform a study that uses a Markov chain to calculate the optimal lineup order and calculates the maximum amount of runs scored per game. This has limitations as the optimization can only be run on the starters and does not incorporate pitching or full roster optimization. Schorsch's model as stated in the journal states that the run expectancy calculation is the weakest link and this project opts to use WAR as a better replacement.

Another model examined uses stochastic modeling for run expectancy similarly to Schorsch, but goes far more in depth while modeling more than just roster optimization by using a nested Dirichlet distribution (Null 2009). This study goes as far as to model each individual plate appearance using this distribution which is out of the scope of this project. To compromise, WAR was decided as the objective variable with salary and player fatigue used as constraints. This creates a solvable optimization problem that captures the problem at hand without adding any extra levels of complexity. The method chosen may not be as accurate as some other studies or models, but provides a good measuring stick as WAR has grown in popularity as a statistic recently.

Methodology

Data Collection and Preparation

We obtained 100 years of baseball data from https://www.baseball-reference.com/ including all manner of statistics from total games played, hits and earned run average to times caught stealing and intentional walks thrown by pitchers. Given that we are interested specifically in the 2016 Chicago Cubs season, much of this data is of no use to use and will be dropped from consideration. Observations that we are interested in from this dataset will be up to five years of historical data from players that appear on the 2015 Cubs roster. For these players, their average statistics since 2011 will be calculated (regardless of whether they played for different teams during this period) with the goal of obtaining an accurate view of the performance and stamina that can be expected from them. Appended to each player's individual season data is also their WAR rating for that year. We have used the WAR calculations performed by FanGraphs.com for this exercise as the information used to calculate this is more granular than what we have available to us from Baseball-Reference.com.

Exploratory Data Analysis

In exploratory data analysis for solving this linear programming problem, the most important aspect is ensuring that all data is available for each player; NA, missing, or incorrect zero values could affect the solution as players with missing data could be heavily favored in minimization situations and ignored in maximization situations. The raw data (Table 1 shows samples) available to us is nearly complete (Figure 1) and only minor adjustments need to be made for several players WAR and salary values. The column with remaining missing values is `OpposingBattingAverage` because this field only applies to pitchers.

Table 1: Sample Data

	playerID	Name	MainPos	SecondPos	ThirdPos	GamesPlayed	AtBats	RunsScored	Hits	...	InningsPitched	WARPG
0	alcanar01	Arismendy Alcantara	cf	2b	3b	40.5	152	18	29.5		0	0.004535
1	arriieja01	Jake Arrieta	p	N/A	N/A	23.8	28	1.6	4.4		139	0.119075
2	baezja01	Javier Baez	2b	ss	3b	40	144.5	14.5	29		0	-0.004159
3	baxtemi01	Mike Baxter	rf	lf	1b	44.6	81.8	10.4	18.8		0	0.003695
4	beelda01	Dallas Beeler	p	N/A	N/A	2.5	3.5	0	1			

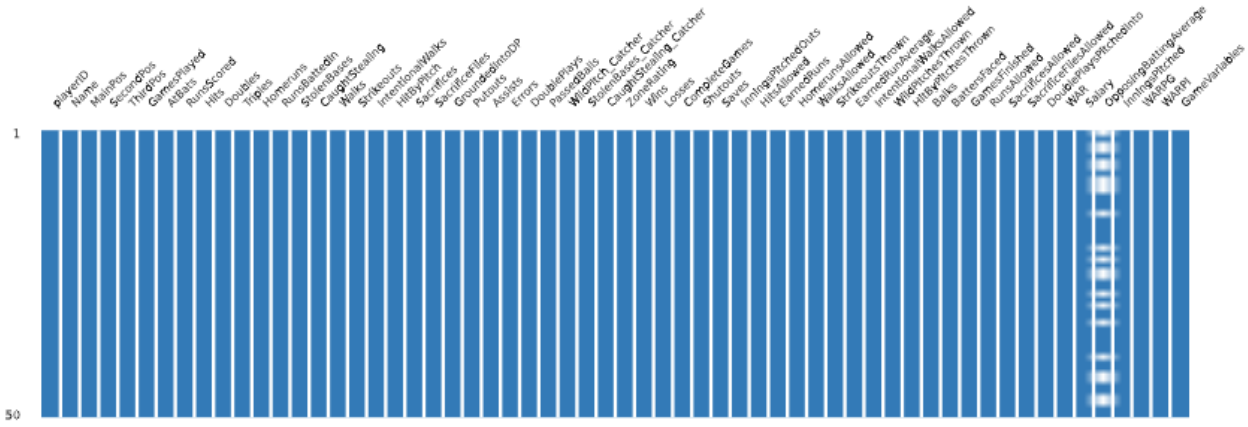

Figure 1: Nullity Matrix

Assumptions

- We assume that 40 players must be selected to the roster for the 2016 season. In the MLB, 26 man Active Rosters are typically set before each series. Since our data does not break down games into series, we will cap the roster size at 40. This 40 man roster accounts for players on the active 26 man as well as those who may be on injured reserve

or available to be called up from minor league teams associated with the major league club.

- We assume that pitchers can only accumulate innings in integer multiples; that is if a pitcher starts an inning they will also complete it instead of potentially being replaced after one or two outs have been recorded.
- We assume that pitchers will need to be relieved in each game; while complete games do occur, we will require each game that a reliever must pitch at least one inning.
- We assume that fielders will play all nine innings of each game they start; pinch hitters and runners are not considered for this program.

Objective

The aim of this linear program is to maximize the team WAR for the Chicago Cubs 2016 season. As player WAR is calculated on a season basis, per-game and per-inning WAR is calculated for fielders and pitchers, respectively to allow for scheduling of players on a day-to-day basis. Once these metrics are calculated, the objective function is defined as the sum of each player's WAR multiplied by their contribution to each game over the course of the 162 game season; for fielders a binary variable representing whether or not they started and for pitchers an integer variable representing the number of innings pitched.

Constraints

Constraints, detailed in **Appendix A**, for this problem fall into two categories: those which limit player usage given their historical playing time and those that ensure a full team is fielded in each game. These are implemented in slightly different ways for pitchers and fielders, though achieve the same goal. For the former category, fielders are limited to a total number of

games played determined by their average number of games played over the past 5 years and a threshold variable that allows them to exceed this by a given percentage. Pitchers are limited instead to innings with a similar percentage excess allowed over their average yearly innings pitched.

Further differences between pitchers and fielders must be accounted for when considering the latter. Fielders are assigned binary variables for whether or not they play a position in a game and which position (if any) they play in a game. Players can only play one position in each game and it must be one of their three most played positions historically. Helper constraints are put in place to ensure players do not accumulate games played in particular positions if they didn't actually play a position as well as the converse scenario. Then, for each game across the Cubs' roster exactly one player must be starting in each position i.e. the sum of variables across the team for a particular position is one. To account for player fatigue, each player is required to have at least one game off out of every fourteen to account for fatigue -- this is achieved by spanning across the 162 game season with a width of 14 and forcing each games-played sum to be less than or equal to 13. Finally, to maintain that players' "main position" as determined by their historical play is maintained in this simulated season, at least 70% of a fielder's play time must come from playing this position.

Pitchers introduce a different type of complexity since only one position is in play instead of 8. This group is broken out into starters and relievers based on their average innings pitched per year. As with fielders, helper constraints are put in place to restrict pitchers from accumulating innings in games they didn't play and vice versa. Each game must have a starting pitcher that throws at least five innings and at least one relief pitcher that throws one inning. Then, like with fielders, each sub-group must account for rest taken between games played.

Starting pitchers take a four game break between starts while relievers can pitch every other game.

Ultimately, we consider the team's salary as a constraint on our solution. Major League Baseball has a Luxury Tax in place that requires teams that spend over a certain amount of money on player salaries in a given year to pay a tax on money spent over this cutoff to disincentivize teams in the country's largest markets from assembling 'super teams' and leaving mid-market clubs with the scraps. As player salaries are only paid to those on the roster, one final group of binary variables that indicate whether a player made the team (made at least one appearance during the season) are multiplied by their salary and summed.

Computational Experiment and Results

Using the PuLP python package (Mitchell, 2011) the objective function and each of the constraints were added to the problem. Due to our large number of constraints (more than 22,000), the longer the program is allowed to run, the closer to an optimal solution it approaches. Two minutes provides enough time for a nearly optimal solution without a great negative impact to productivity. For sensitivity analysis, we tested simulating player "injury" (or other unavailability). Our tests included:

- "high value" players
    - Kris Bryant
    - Jon Lester
- A "medium value" player:
    - Jorge Soler
- A "low value" player:

    - Jake Arrieta
- And an unused, negative value player:
    - Donn Roach

When completely removing a “high value” fielder such as Kris Bryant, the program was unable to converge to a solution. This would indicate that the team’s current roster is insufficient & not robust to losing certain players. The recommendation an analyst would provide the manager is to ensure all positions can be backfilled. We then tested taking out a mix of values of players for half of the season and Table 2 provides the results. As expected, an injury to a high value player has a more significant impact on final season WAR. In this small sample set, loss of player season WAR is trending to correlate with that player’s WAR; section 7 discusses how we would like to expand on this analysis.

Table 2: Sensitivity Analysis

playerID	Name	Pos	Player WAR	Time injured	Season WAR
-	Baseline - all available		-	-	53.66
bryankr01	Kris Bryant	3B	6.1	Full Season	No Solution
bryankr01	Kris Bryant	3B	6.1	Half Season	49.33
lestejo01	Jon Lester	P	4.8	Half Season	52.6
solerjo01	Jorge Soler	RF	.50	Half Season	53.55
roachdo01	Donn Roach	P	-.14	Half Season	53.66

Without accounting for the sensitivity analysis of extended player injuries, an optimal solution was found within the constraints defined. The team season WAR was calculated to be 53.66. This result is found with the following configuration shown in the stacked bar chart of

players and positions played per game. The stacked bar chart (Figure B.1) represents how many games each player played by position. Teams will often have players play their secondary position due to changes in rotation and players starting that day. This is reflected in the results. Figure B.2 shows each player's predicted games played for the 2016 season vs the actual games played during that season. Considering that this model fails to incorporate mid-season trades as well as previous statistics for rookies, the model does a relatively good job. The model fails to correctly predict some major contributors such as Javier Baez, Jason Heyward, and Ben Zobrist. Players who suffered major injuries during the season will also be falsely represented in the model such as Miguel Montero. One other consideration that the model does not account for is pinch hitters such as Kyle Schwarber who may not have actually started in a lot of games, but had a lot of at bats due to substitution were not represented well. Rather than predict the correct amount of games played by each player, the optimization provides a metric of value that each player contributes to the team. (Figures shown in appendix B)

Discussion and Conclusions

The program was able to converge on an optimal solution for the season, meeting all constraints and outputting lineups for all 162 games. A replacement team with a WAR of 0 would produce 52 wins in a 162 game season (Sports Reference 2021). Using the calculated WAR from the optimal solution, the 2016 Cubs are predicted to end the season with a team record of 105.66 wins. The actual 2016 Cubs won 103 games that season so the results provide an accurate estimate for a team's record. This would be further tested with every team and varying seasons to prove the legitimacy of this method.

While the problem solved here attempts to optimize the roster of the 2016 Cubs and position played by each player, it does not take into account possible trades and rookies as it uses historical data. Possible solutions to account for this issue would be to have an adjusted constant for all rookies. This would be a linear constant that would decrease minor league statistics to be on par with major league stats. Doing this would allow for a better optimization of the roster due to the more realistic consideration of minor league players/rookies new to the team.

To further this study, there are many additional factors or variables that can be incorporated. To create a better model requires increasing the complexity of the problem. Incorporating minor league prospects and trade options would provide an increased overview for roster optimization. This would allow consideration of all potentially available players instead of restricting to the players already in the organization. Along with this, considering player decline due to age is a real life consideration that must be calculated in. For example, a player reaching his 40s will typically not provide the same production he has in his younger years. Our current sensitivity analysis could be more thorough by wrapping our solution into methods capable of iterating over the entire team and varying parameters. Sensitivity analysis could also be performed to develop cost-benefit analysis for making trades or offering higher salaries to retain high-performing players (WAR/cost tradeoff).

References

Duarte A.R., Ribeiro C.C., Urrutia S., and Haeusler E.H. 2007. “Referee Assignment in Sports Leagues.” In: Burke E.K., Rudová H. (eds) Practice and Theory of Automated Timetabling VI. PATAT 2006. Lecture Notes in Computer Science, vol 3867, pp 158-173. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-77345-0_11

Greene, Michael J and Adam J. Hirsch. “Workforce Analytics in Baseball Player Management.” (2011).

Lewis, Michael M. 2003. *Moneyball: The Art of Winning an Unfair Game.* New York: W. W. Norton. ISBN 0-393-05765-8.

Mitchell, Stuart, Michael J. O'Sullivan and Iain Dunning. “PuLP : A Linear Programming Toolkit for Python.” (2011).

Null, Brad. 2009. Stochastic modeling and optimization in baseball. Ph.D. diss., Stanford University, http://turing.library.northwestern.edu/login?url=https://www.proquest.com/dissertations-theses/stochastic-modeling-optimization-baseball/docview/305012994/se-2?accountid=12861 (accessed December 4, 2021).

Schorsch, Emanuel & Valera, A. (2015). Baseball Lineup Optimization. 10.13140/RG.2.2.27186.04800.

“The Relationship between War and Team Wins.” Sports. Accessed December 5, 2021. https://www.sports-reference.com/blog/2012/08/the-relationship-between-war-and-team-wins/.

Appendix A. Parameters and Constraints

A.1 Sets

s - indexed member of starting pitcher list

r - indexed member of relieving pitcher list

p - indexed member of pitcher list (inclusive of PS, PR)

f - indexed member of fielder list

l - indexed member of all players list (inclusive of PS, PR, & F)

o - indexed item from position list

g - index of game number of 162

A.2 Parameters

GV_{lg} - GameVariables **binary** for whether player participates in game

PO_{fo} - PositionVariables **binary** for which position fielder fills in each game

IV_{pg} - InningVariables **integer** records how many innings each pitcher plays in each game

MR_{l} - MadeRoster **binary** for whether a player makes the forty-man roster

$WARPI_{p}$ - WAR Per Inning **float** for each pitcher

$WARPG_{f}$ - WAR Per Game **float** for each fielder

SAL - Salary **float** for each player

MP - Main Position **categorical** for each fielder

A.3 Objective

$$\max Z = \Sigma(GV_{f} * WARPG_{f}) + \Sigma(IV_{p} * WARPI_{l})$$

A.4 Constraints

A.4.1 Pitchers

$$\sum_{G} IV_{pg} \leq 1.1 * Average\ Innings/yr \quad \text{(A1)}$$

$$1000 * GV_{pg} \geq IV_{pg} \quad \text{(A2)}$$

$$GV_{pg} * 5 \leq IV_{pg} \quad \text{(A3)}$$

$$\sum_{G=i,\, i+5} GV_{pg} <= 1,\ i \in \{1, 2, \ldots 157\} \quad \text{(A4)}$$

$$\sum_{s} GV_{pg} = 1,\ g \in \{1, 2, \ldots 162\} \quad \text{(A5)}$$

Constraint (A1) limits a pitcher to 10% more than his average innings per year. (A2) A pitcher cannot accumulate innings during a game he did not participate in; arbitrarily large number for weighting GV_{pg}. (A3) Starting pitchers must pitch at least five innings in a start. (A4) A starting pitcher cannot pitch in more than one game per five-game stretch. (A5) Exactly one starting pitcher must play in every game of the season.

$$GV_{rg} \leq IV_{rg} \quad \text{(A6)}$$

$$\sum_{r} GV_{rg} \geq 1,\ g \in \{1, 2, \ldots 162\} \quad \text{(A7)}$$

$$\sum_{G=i,\, i+1} GV_{rg} <= 1,\ i \in \{1, 2, \ldots 161\} \quad \text{(A8)}$$

$$\sum_{r} IV_{rg} + \sum_{s} IV_{sg} = 9,\ g \in \{1, 2, \ldots 162\} \quad \text{(A9)}$$

Constraint (A6) forces relief pitchers to throw at least one inning per appearance. (A7) Relief pitchers as a group must pitch at least one inning per game. (A8) Relievers must take at least one day off after making an appearance in a game. (A9) Between relievers and starters, a full nine innings must be pitched each game.

A.4.2 Fielders

$$\sum_{G} GV_{fg} \leq 1.1 * Average\ Games/yr \quad \text{(A10)}$$

$$\sum_{f} \sum_{o=1,8} PO_{fog} = 8,\ g \in \{1, 2, \ldots 162\} \quad \text{(A11)}$$

$$\sum_{o=1,8} PO_{fog} \leq GV_{fg},\ g \in \{1, 2, \ldots 162\} \quad \text{(A12)}$$

$$\sum_{G=i,\, i+14} GV_{fg} \geq 13,\ i \in \{1, 2, \ldots 139\} \quad \text{(A13)}$$

$$\sum_{G=1,162 | o=f_{MP}} PO_{fog} \geq \sum_{G=1,162} GV_{fg} * 0.7 \quad \text{(A14)}$$

Constraint (A10) limits a fielder to at most 10% more than his average games played per year. (A11) forces eight fielders to be play each game with exactly one fielder at each position. (A12) restricts fielders from accumulating games played at a position from games in which they do not enter. (A13) Fielders must take at least one game off for rest for each fourteen game stretch. (A14) Players must spend at least 70% of their games playing their main position.

A.4.3 Roster Size and Salary Cap

$$\sum_{l} MR_{l} = 40 \quad \text{(A16)}$$

$$\sum_{l} MR_{l} * SAL_{l} \leq 189,000,000 \quad \text{(A17)}$$

Constraint (A16) forces 40 players to be chosen to make the 2016 roster. (A17) restricts the total salary of the team to $189MM, the threshold at which the MLB luxury tax would come into effect.

Appendix B. Figures and Tables

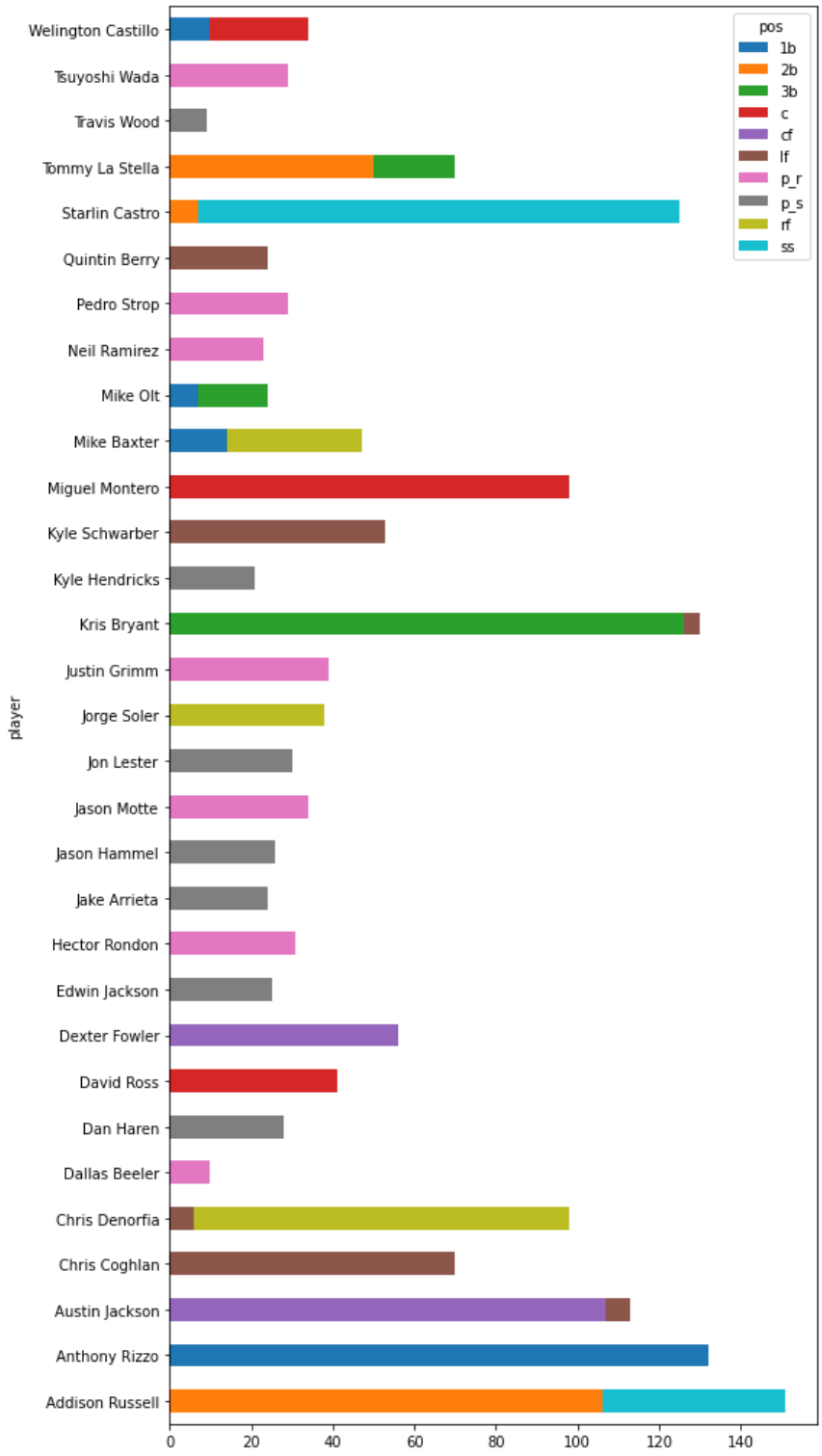

Figure B.1. Bar chart of games played and position by player

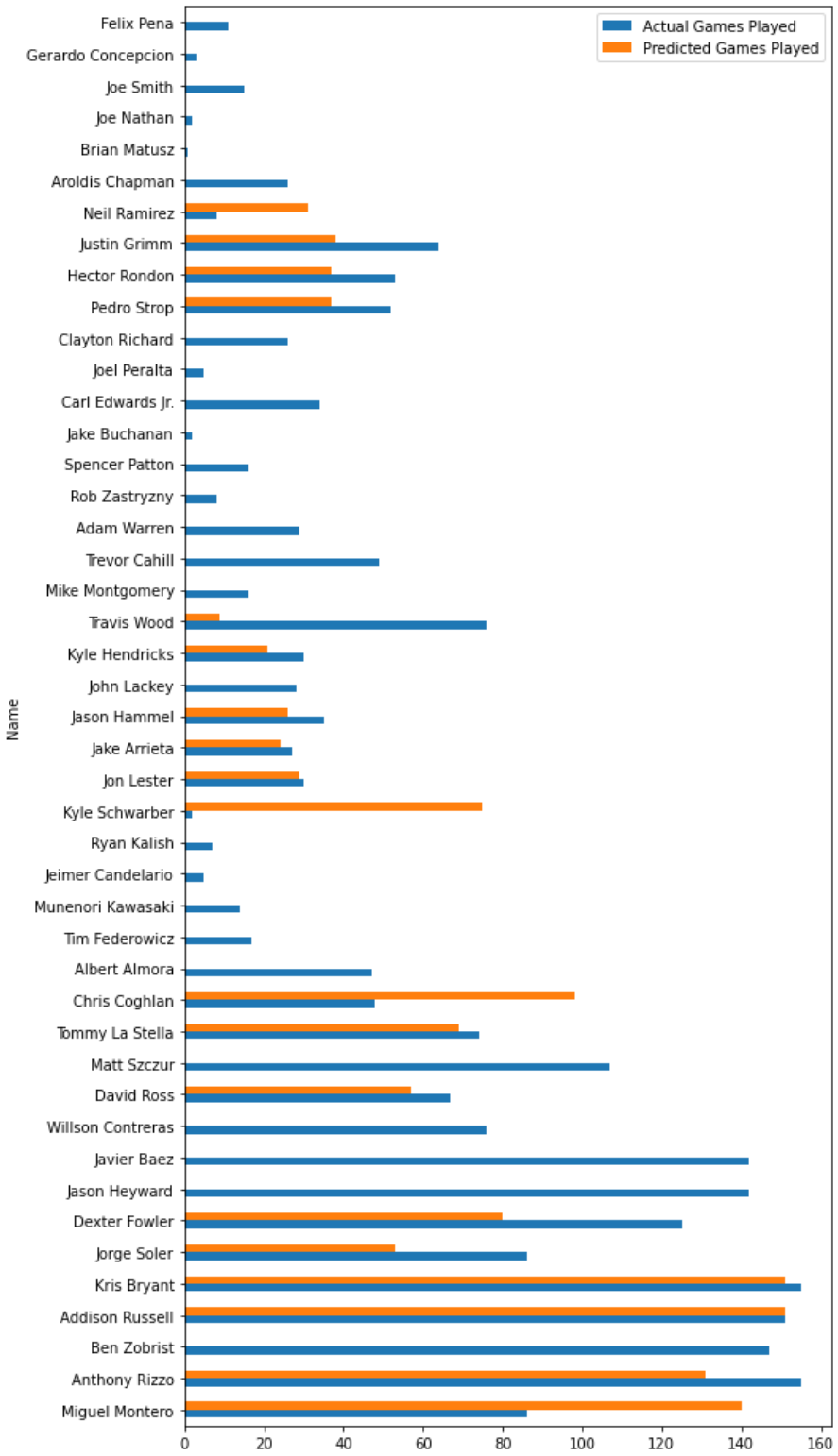

Figure B.2. Bar chart comparing predicted games played and actual games played by player

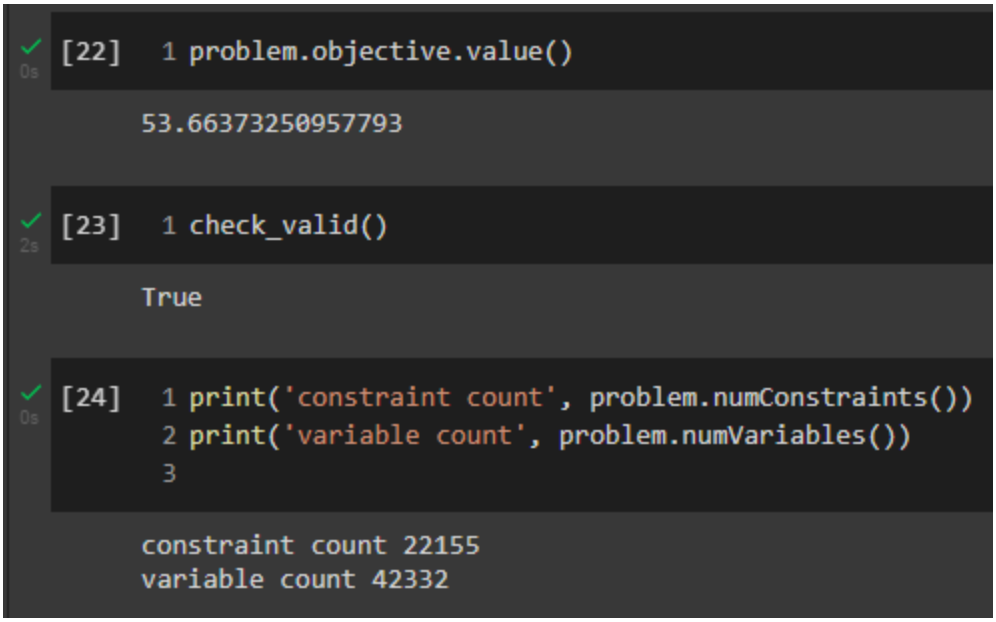

Figure B.3. Solution Validation

Table B.2. Optimization Lineup Results - Player Sample

playerID	Name	WAR	GameCount
alcanar01	Arismendy Alcantara	0.183657	0
baezja01	Javier Baez	-0.16636	0
baxtemi01	Mike Baxter	0.164776	47
berryqu01	Quintin Berry	0.353005	34
bryankr01	Kris Bryant	6.098138	151
castiwe01	Welington Castillo	0.740666	34
castrst01	Starlin Castro	1.929982	151

Table B.3. Optimization Lineup Results - Game Sample

	Position in lineup										
game	**c**	**1b**	**2b**	**ss**	**3b**	**lf**	**rf**	**cf**	**p_s**	**p_r**	**p_r2**
1	**Miguel Montero**	**Anthony Rizzo**	**Addison Russell**	**Starlin Castro**	**Kris Bryant**	**Chris Coghlan**	**Jorge Soler**	**Austin Jackson**	**Dan Haren**	**Jason Motte**	**NaN**
2	**Welington Castillo**	**Anthony Rizzo**	**Addison Russell**	**Starlin Castro**	**Kris Bryant**	**Chris Coghlan**	**Chris Denorfia**	**Dexter Fowler**	**Jason Hammel**	**Dallas Beeler**	**Justin Grimm**
3	**Miguel Montero**	**Anthony Rizzo**	**Addison Russell**	**Starlin Castro**	**Kris Bryant**	**Chris Denorfia**	**Mike Baxter**	**Dexter Fowler**	**Jake Arrieta**	**Neil Ramirez**	**NaN**
4	**Welington Castillo**	**Mike Baxter**	**Addison Russell**	**Starlin Castro**	**Kris Bryant**	**Chris Coghlan**	**Chris Denorfia**	**Austin Jackson**	**Edwin Jackson**	**Pedro Strop**	**NaN**
5	**Miguel Montero**	**Anthony Rizzo**	**Addison Russell**	**Starlin Castro**	**Kris Bryant**	**Austin Jackson**	**Mike Baxter**	**Dexter Fowler**	**Kyle Hendricks**	**Hector Rondon**	**NaN**