Project Proposal: Predicting MLB Player Performance

Mike Ogrysko

Loyola University Maryland

DS 795: Data Science Project Design

Dr. Bu Hyoung Lee

December 1, 2022

Project Background

Major League Baseball (MLB) is a multi-billion dollar-business, nearing $11 billion in gross revenues in 2022 (Brown, 2022). Its 30 teams have player payrolls that range from $61 million to $282 million with an average of $163 million (Spotrac, 2022). The average value of an MLB team is $2 billion (Ozanian & Teitelbaum, 2022). Given the enormity of these revenues, payrolls, and team values, MLB and its teams are relying on data now more than ever to build their on-field product.

How are these teams using the data they collect? Some teams are using this data as a development tool to measure performance and adjust technique as necessary. For example, pitching coaches for the Boston Red Sox used video and data collected from high-speed cameras and ball trackers to diagnose biomechanical delivery flaws of two of their relief pitchers (Schrage, 2019). Players are using analytics to market themselves during free agency. In the winter leading up to the 2020 season, outfielder Marcell Ozuna sent a 43-page booklet to teams highlighting metrics where he excelled like the exit velocity of his hits and his Expected Weighted On-base Average (xwOBA) (Davidoff, 2019), which takes exit velocity, launch angle, and sprint speed into account to create a measure of a player's offensive contributions (MLB, 2022). Teams use data about where a player hits the ball to position their defensive players on the field. In 2013, the Tampa Bay Rays and Milwaukee Brewers were estimated to have saved their teams between 10 and 15 runs over the course of the season just by shifting their players to specific areas on the field based on data they evaluated for each hitter (Common, 2014). Teams are using data to project the results of individual pitcher and hitter matchups. The Yankees estimate hitters' performances against their pitchers' specific pitches – using measured pitcher velocity and spin as well as a hitter's typical bat path – to simulate an expected result. This allows the team to map out specific scenarios and plans for each game (Carig, 2018). Teams also use this data to generating metrics to project player performance and facilitate the creation of their season rosters.

This project will focus on the player performance aspect of data analytics in Major League Baseball. Specifically, it will use available player data to answer the following research questions:

- What is the estimated peak player age at each position?
- Of the players who are playing at their estimated peak age in 2023, which will exceed their 2022 Wins Above Replacement (WAR) metric?
- Of the players who played at their estimated peak age in 2022, which will exceed their 2022 WAR metric?
- How will these players perform in 2023?

Source and Nature of Data

MLB player data is publicly available from several sources. This project will rely on the data compiled at Stathead Baseball (https://stathead.com/baseball/). This site guarantees full batting, pitching, and fielding statistics, as well as play-by-play data from 1976 onward (Baseball-Reference, 2022). This project's research questions will rely on player data from 1980 onwards. All data will be exported from Stathead Baseball in a series of CSV files.

A crucial metric for measuring player performance is Wins Above Replacement (WAR). The general idea behind WAR as a metric is placing a value on a player that accounts for the number of team wins that he alone represents when compared to a replacement player at the same position (Castrovince, 2020, p. 189). As an example, a team comprised of only Minor League players is estimated to win 48 games. If a player from that team was replaced with a free agent, and the team's win total jumped to 54 games, the new player would have a 6 WAR (Castrovince, 2020, p. 191). Interestingly, there is no single calculation for a player's WAR. The individual teams have their own WAR calculations, factoring in different public inputs and other considerations such as scouting reports, budget, and other market-specific information (Castrovince, 2019). Public facing statistical reference sites, Baseball-Reference, FanGraphs, and Baseball Prospectus have their own calculations for WAR that use different metrics in their computations (Baseball-Reference, 2022). This project will use the Baseball-Reference calculation for WAR, which is available for each player through Stathead Baseball.

Tools and Methods

Estimated Peak Age at Each Position

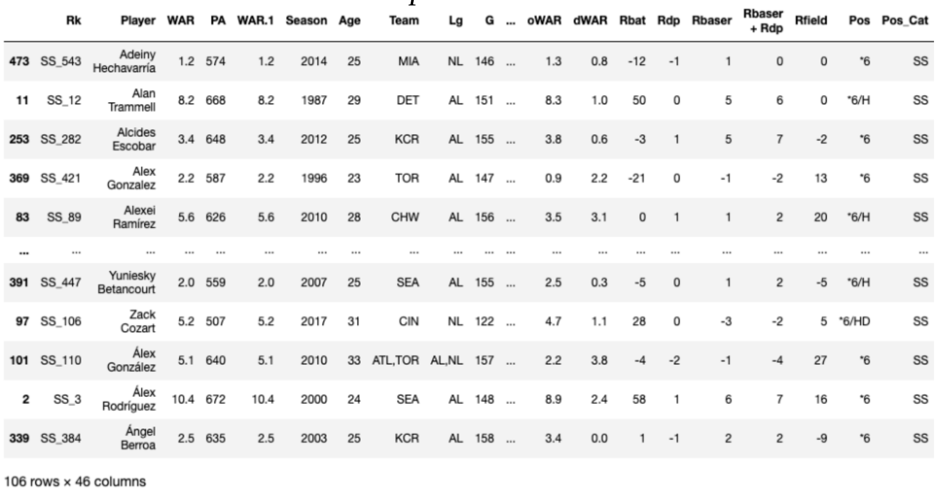
To determine estimated player peak ages based on position, this project will focus on single-season player data from 1980 to present matching the following criteria:

- Regular season only
- Played specific position for at least 51% of games (positional players)
- Plate appearances are greater than or equal to 500 (positional players)
- Started 60% of games pitched in (starting pitchers)
- Started at least 20 games (starting pitchers)

This represents approximately 5500 positional player seasons and 3800 starting pitcher seasons. From there, the data will be filtered to only include season data for players who have played 3 or more seasons at their position. The peak will be determined by identifying the season for a particular player where the maximum WAR was achieved (see Figure 1).

Figure 1

Individual Career-Best Shortstop Seasons 1980-2022



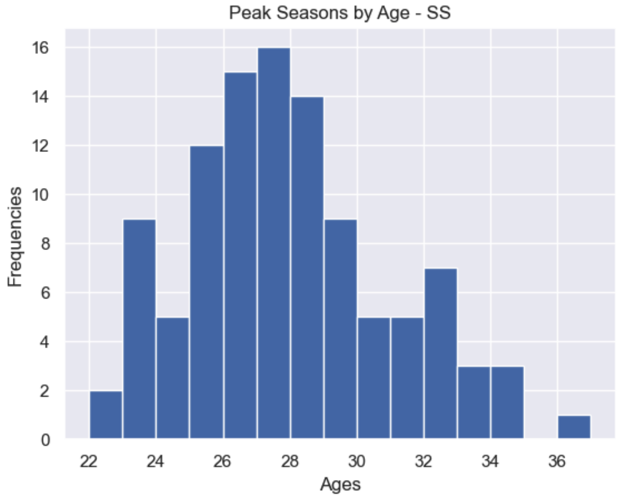
	Rk	Player	WAR	PA	WAR.1	Season	Age	Team	Lg	G	...	oWAR	dWAR	Rbat	Rdp	Rbaser	Rbaser + Rdp	Rfield	Pos	Pos_Cat
473	SS_543	Adeiny Hechavarría	1.2	574	1.2	2014	25	MIA	NL	146	...	1.3	0.8	-12	-1	1	0	0	*6	SS
11	SS_12	Alan Trammell	8.2	668	8.2	1987	29	DET	AL	151	...	8.3	1.0	50	0	5	6	0	*6/H	SS
253	SS_282	Alcides Escobar	3.4	648	3.4	2012	25	KCR	AL	155	...	3.8	0.6	-3	1	5	7	-2	*6	SS
369	SS_421	Alex Gonzalez	2.2	587	2.2	1996	23	TOR	AL	147	...	0.9	2.2	-21	0	-1	-2	13	*6	SS
83	SS_89	Alexei Ramírez	5.6	626	5.6	2010	28	CHW	AL	156	...	3.5	3.1	0	1	1	2	20	*6/H	SS
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
391	SS_447	Yuniesky Betancourt	2.0	559	2.0	2007	25	SEA	AL	155	...	2.5	0.3	-5	0	1	2	-5	*6/H	SS
97	SS_106	Zack Cozart	5.2	507	5.2	2017	31	CIN	NL	122	...	4.7	1.1	28	0	-3	-2	5	*6/HD	SS
101	SS_110	Álex González	5.1	640	5.1	2010	33	ATL,TOR	AL,NL	157	...	2.2	3.8	-4	-2	-1	-4	27	*6	SS
2	SS_3	Álex Rodríguez	10.4	672	10.4	2000	24	SEA	AL	148	...	8.9	2.4	58	1	6	7	16	*6	SS
339	SS_384	Ángel Berroa	2.5	635	2.5	2003	25	KCR	AL	158	...	3.4	0.0	1	-1	2	2	-9	*6	SS

106 rows × 46 columns

This will provide a frequency for each age at each position. The age with the highest frequency will represent the estimated peak age for a given position (see Figure 2).

Figure 2

Frequency of Career-Best Seasons based on Age



Exceeding Previous WAR Metrics in 2023

The analysis of whether a player will exceed their 2022 WAR metric is a classification problem. In this project, to classify whether a player will exceed their 2022 WAR in 2023, training data will be compiled to include the following:

- Averages of basic statistics (e.g., batting average, hits, stolen bases, etc.)
- Average WAR
- Pre-peak, peak, and post-peak WAR
- 0/1 classifier indicating whether or not the player exceeded WAR in peak and post-peak seasons

This training data will consist of approximately 3000 position player records and 1000 starting pitcher records. After compiling this data in CSV exports, Pandas dataframes will be used to create averages, engineer features (as necessary), and clean the data. The Support Vector Machine (SVM) algorithm will be used to classify whether or not the player exceeded his previous season’s WAR metric. The goal with SVM is to maximize the margin between the separating hyperplane, or decision boundary, and the support vectors, or the training examples closest to the hyperplane while providing a binary output – 0 or 1 (Raschka & Mirjalili, 2019, p. 166). For this classification, the model will return a “0” if the player is not expected to exceed his 2022 WAR in 2023 and a “1” if he is expected to exceed his 2022 WAR.

Player Performance in 2023

Despite the different calculations of WAR, its scale is widely agreed upon (see Table 1). This project will attempt to project 2023 player performance by first establishing a player’s expected WAR category using this scale as a basis for categorization.

Table 1

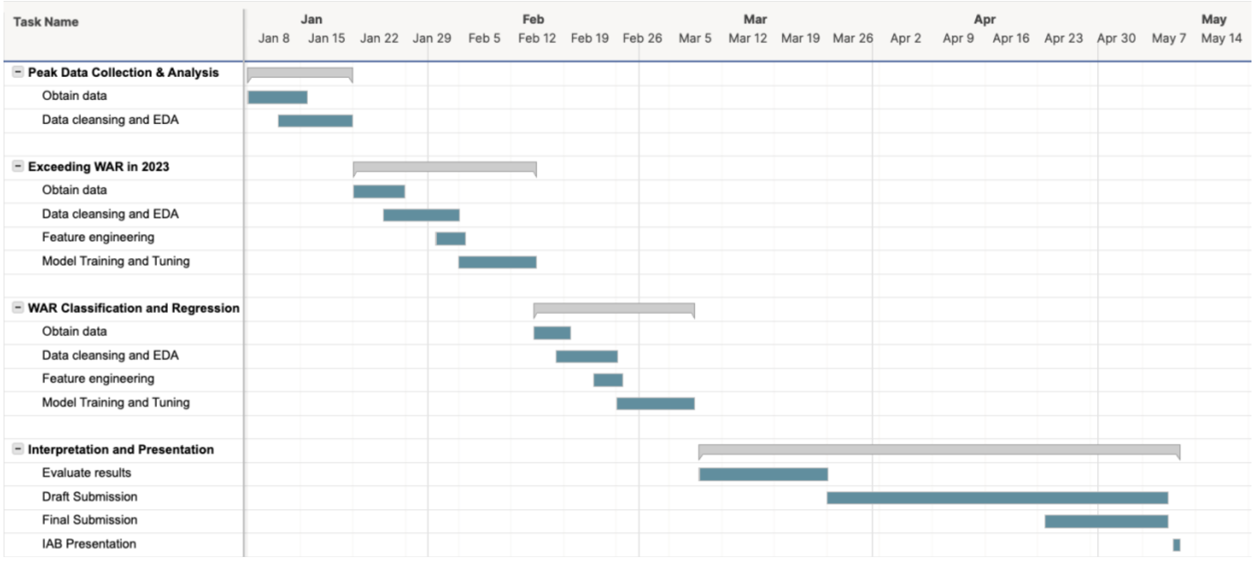
WAR Categorization for Positional Players and Starting Pitchers

Quality of Player	WAR
MVP Material	8 or higher
Superstar	6-8
All-Star	4-6
Solid Regular	2-4
Role Player	1-2
Bench Material	0-1
Triple A Material	0 or below

Note. Reprinted from *A Fan's Guide to Baseball Analytics* (p. 192) by A. Castrovince, 2020, Sports Publishing. Copyright 2020 by Anthony Castrovince.

A training dataset will be compiled to include much of the same data as in the previous classification with a 0-6 WAR categorization. Again, the training data should consist of approximately 3000 position player records and 1000 starting pitcher records. The Random Forest algorithm will be used to classify the player's position on the WAR scale. The Random Forest algorithm is an ensemble of decision trees, which are averaged to build a robust model with good generalization performance while generating a categorization – 0 to 6 in this case (Raschka & Mirjalili, 2019, p. 198). After classification, a multiple linear regression model will be created to predict the WAR metric for each player.

Project Plan



Assumptions and Risks

For this project, the following risks and assumptions have been identified:

- Data available from Stathead Baseball is sufficient for analysis. As the project continues, there is a risk that additional data could be necessary and would need to pulled from other sources to be merged with an existing dataset.
- Currently identified models will be sufficient to answer the research questions. As the project continues, alternative models may be necessary to achieve the desired results.
- With any project, scope creep is a possibility – will additional questions be introduced that need to be answered in this analysis?
- Overall, risks associated with data availability are low because the data is publicly available.

References

Baseball-Reference. (2022). *Data Coverage.* https://www.baseball-reference.com/about/coverage.shtml

Baseball-Reference. (2022). *WAR comparison chart*. https://www.baseball-reference.com/about/war_explained_comparison.shtml

Brown, M. (2022, Apr 7). *How Major League Baseball could crack $11 billion in revenues in 2022*. Forbes. https://www.forbes.com/sites/maurybrown/2022/04/07/how-major-league-baseball-could-crack-11-billion-in-revenues-in-2022/?sh=68098b77f63a

Carig, M. (2018, Sep 27). *The Yankees have paired the best bullpen ever with cutting-edge tools to optimize its usage. Will it be enough in the playoffs?* The Athletic. https://theathletic.com/552336/2018/09/27/the-yankees-have-paired-the-best-bullpen-ever-with-cutting-edge-tools-to-optimize-its-usage-will-it-be-enough-in-the-playoffs/?redirected=1

Castrovince, A. (2019, Feb 3). *The influence of WAR on modern front offices*. MLB.com. https://www.mlb.com/news/war-embraced-by-mlb-front-offices-c303484670

Castrovince, A. (2020). *A fan's guide to baseball analytics*. Sports Publishing.

Common, D. (2014, Aug 18). *How the defensive shift and big data are changing the game*. CBC. https://www.cbc.ca/news/world/how-the-defensive-shift-and-big-data-are-changing-baseball-1.2739619

Davidoff, K. (2019, Nov 19). *Marcell Ozuna among players smartly embracing analytics in MLB free agency*. New York Post. https://nypost.com/2019/11/19/marcell-ozuna-among-players-smartly-embracing-analytics-in-mlb-free-agency/

MLB. (2022). *Expected Weighted On-base Average (xwOBA)*. https://www.mlb.com/glossary/statcast/expected-woba

Ozanian, M. & Teitelbaum, J. (2022, Mar 24). *Baseball's most valuable teams 2022: Yankees hit $6 billion as new CBA creates new revenue streams*. Forbes. https://www.forbes.com/sites/mikeozanian/2022/03/24/baseballs-most-valuable-teams-2022-yankees-hit-6-billion-as-new-cba-creates-new-revenue-streams/?sh=244684d600a2

Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning*. Packt Publishing.

Schrage, M. (2019, July 15). *What baseball can teach you about using data to improve yourself.* Harvard Business Review. https://hbr.org/2019/07/what-baseball-can-teach-you-about-using-data-to-improve-yourself

Spotrac. (2022). *MLB team payroll tracker*. https://www.spotrac.com/mlb/payroll/2022/?ref=trending-pages