UNIVERSITY of VIRGINIA | McINTIRE SCHOOL of COMMERCE
Center for Business Analytics

UVA-CBA-0121
Revised December 1, 2019
DRAFT

Salary Arbitration at MLB

In 2015, Major League Baseball (MLB) was planning for the collective bargaining agreement (CBA) with the MLB players' association (MLBPA). The old agreement which had been in effect since 2012 was set to expire on November 30, 2016. The new CBA would tackle many of the standard topics, including: revenue sharing, minimum salaries, scheduling, postseason play, all-star game format, clubhouse operations and amenities, player welfare and benefits, and drug prevention and treatment programs. MLB leadership was curious if it might make sense to enhance their salary arbitration process through advanced analytics and machine learning. If so, the league's negotiations with MLBPA around the new CBA presented the perfect opportunity to introduce a more rigorous, efficient analytics-driven process. However, it was unclear if the use of analytics/machine learning models did indeed constitute a more "rigorous" and "efficient" alternative. MLB's senior vice president (SVP) for operations felt that analysis was needed to explore the efficacy of analytical models in the arbitration process.

Background on Salary Arbitration Process

The MLB salary arbitration process dates back to 1970. At the time, owners thought an arbitration process might be a better alternative than free agency – although by 1976, the league had adopted both. In 2015, as noted on MLB.com at the time, the arbitration system worked as follows:

> *Players who have three or more years of Major League service but less than six years of Major League service become eligible for salary arbitration if they do not already have a contract for the next season. Players who have less than three but more than two years of service time can also become arbitration eligible if they meet certain criteria; these are known as "Super Two" players. Players and clubs negotiate over appropriate salaries, primarily based on comparable players who have signed contracts in recent seasons…*
>
> *If the club and player have not agreed on a salary by a deadline in mid-January, the club and player must exchange salary figures for the upcoming season. After the figures are exchanged, a hearing is scheduled in February. If no one-year or multi-year settlement can be reached by the hearing date, the case is brought before a panel of arbitrators. After hearing arguments from both sides, the panel selects either the salary figure of either the player or the club (but not one in between) as the player's salary for the upcoming season.*
>
> *The week prior to the exchange of arbitration figures is when the vast majority of arbitration cases are avoided, either by agreeing to a one- or multi-year contract. Multi-year deals, in these instances, serve as a means to avoid arbitration for each season that is covered under the new contract.*
>
> *Once a player becomes eligible for salary arbitration, he is eligible each offseason (assuming he is tendered a contract) until he reaches six years of Major League service. At that point, the player becomes eligible for free agency.*

There were three key stages in the "arbitration funnel." First, players with 3-6 years of MLB service (and in some cases, those with 2 years) were eligible for the arbitration process. The MLB defined any "year of service" as one where a player accrued at least 172 days on a major league roster. For example, in 2014 a

This case was prepared by Ahmed Abbasi, Associate Dean and Murray Research Professor & Brent Kitchens, Assistant Professor and Associate Director, Center for Business Analytics total of 185 players were eligible for the arbitration process. The second stage in the funnel was whether or not an eligible player and his club failed to agree upon a salary by mid-January. If so, the two sides would have to file for arbitration and submit their respective salary numbers. For the 10 year period spanning 2005-2014, about 30% of eligible player cases resulted in arbitration filings. The third and final stage in the funnel was the hearing and outcome determined by an arbitration panel (scheduled for February). A very small proportion of cases reached this point, with players and teams typically electing to settle beforehand. Between 2005 and 2014, only 55 cases were decided by a panel – merely 4% of all eligible players and 11% of total filed cases. Figure 1 shows the arbitration panel decision trends over the years (i.e., the number of annual player versus club "wins").

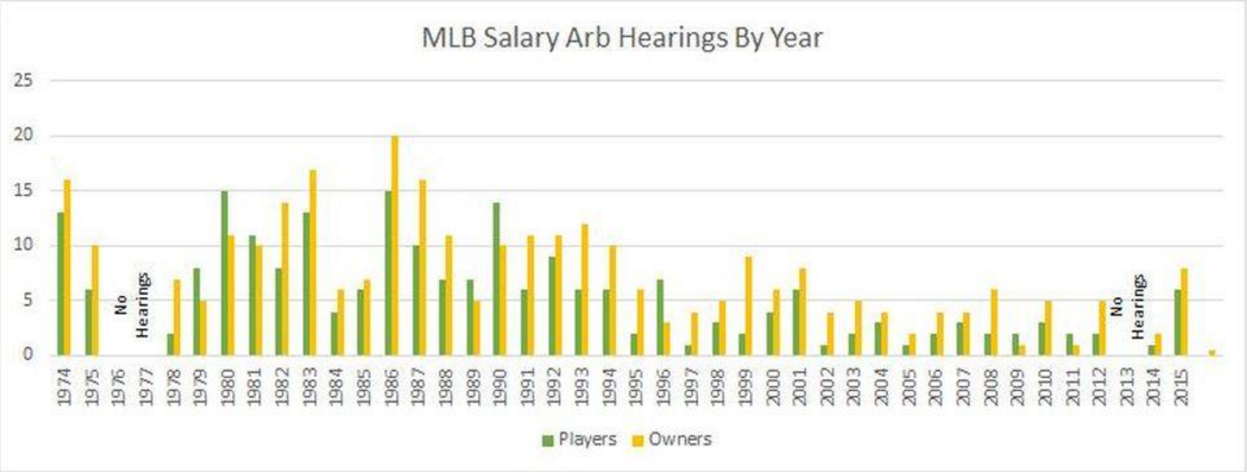

Figure 1: Salary Arbitration Hearing Outcomes over the Years (Source: Forbes Magazine, 2015)

Leveraging Machine Learning to Improve the Process

The UVA team was asked to explore the feasibility of using machine learning models for the second and third stages of the arbitration process. As the MLB SVP for Operations noted in a meeting:

> *"The ideal scenario would be a model that can accurately predict the outcome of an arbitration hearing before there is even a filing. That being said, there are lower hanging fruit. The process of figuring out how many eligible cases might result in filings, how many (and which) filings will need a panel, and assembling a panel of impartial judges is non-trivial, expensive, and time-consuming. Any mechanisms that shed light on information asymmetries throughout the process will be highly advantageous from a planning and process improvement perspective. If there is a case to be made for use of advanced models in the process, this would be a good time to know."*

MLB provided the UVA team with data for all players eligible for arbitration in the ten seasons spanning 2005 through 2014. The data included 1,358 total players: starting pitchers, relief pitchers, and positional players. The UVA team quickly realized that while many of their other industry and academic research projects could be classified as big data initiatives, this one entailed a "small N" situation – the set of feasible machine learning prediction tasks would be limited by the available data. Further, the relevant statistics for pitchers differed significantly from those for positional players (e.g, 1-3 basemen, short stops, outfielders, catchers). Hence, the team initially decided to explore machine learning predictive models for whether an eligible case resulted in a formal filing. This would allow the MLB to better anticipate demand for hearing dates in February, as well potential demand for judges panels. Admittedly, predicting filings was not considered to be as valuable as predicting arbitration outcomes (e.g., who would win or final salary) – particular given most filings did not result in hearings – but seemed most feasible as a starting point.

Exhibit 1

Salary Arbitration at MLB

Table A1: Fields in Dataset for Starting and Relief Pitchers

Category	Variables	Description
Org and Year	Org	Dummy variable for which organization a particular player played for during that year
	Year	The player year, ranging from 2005 thru 2014
	TE	Times eligible for arbitration. This represents the class of players that this player should be compared to.
	MLS	Major League service time. This is how many years.days served in the major leagues, and drives the comparison group for the player as well as when each player is eligible for arbitration.
	Filed	One of the DVs of interest – whether the player-year case resulted in a formal filing an arbitration hear. As noted, this happens when the two parties are unable to agree on a contract by the mid-January deadline.
	PY LRD Sal	Player labor relations department salary
	Salary Multiplier	Salary relative to the MLB minimum salary for that year. This was done to adjust for inflation in MLB salaries.
Player Year Stats	IP	Total innings pitched
	ERA	Earned run average
	G	Games played
	GS	Games started
	W	Wins
	L	Losses
	SV	Saves
	SVOP	Save opportunities
	HLD	Holds
	HLD 1+	Holds for 1+ innings (i.e. 3+ outs)
	DL	Days on the disabled list
	WAR	Wins above replacement
	WS	Win shares
	SO/BB	Strikeouts per bases on balls
	SO/9	Strikeouts per nine innings
	AVG	Batting average allowed
	OBP	On base percentage
	SLG	Slugging percentage allowed
	OPS	On base percentage + slugging percentage allowed
Player Career Stats	Same as player year stats	Aggregated and/or averaged across the player's career

For further details on any column, see: http://www.espn.com/gen/editors/mlb/glossary.html