Evaluating Major League Baseball Arbitration Awards Using a Hybrid Learning Model

Jonathan Gordon, Nick Aswad, Miguel Betances, Phil Kudryavtsev

Abstract

Major league baseball players' salaries are initially determined by the contract they sign after being drafted. They are not free to negotiate a new salary until they reach three years of major league service time. After the third year, they can negotiate their next contract with their current team. If they cannot reach an agreement with their club, their salary will be determined by an independent arbiter. These players are identified as "arbitration eligible". In this paper we will use both clustering techniques (such as DBSCAN, K-Means, TSNE) and regression techniques (such as KNN, Bagged Tree, and Random Forest) to predict the salary the arbitration players might expect to receive, using publicly available 2023 player statistics. The predicted salaries will then be compared to the actual amounts negotiated or awarded by the independent arbitration panel. The results show that certain categories of players are underpaid, but not all categories. The inability of the arbitration players to negotiate with any club does not appear to be as large of a detriment as expected.

INTRODUCTION

Major League Baseball (MLB) players are some of the highest paid professional athletes in the world. Star players get large, 9-figure multi-year deals while newcomers or minor contributors sign league minimum entry-level contracts. Every year a draft is held for the 30 MLB teams to select high school and college players to join their team. A second draft is held to select international players. After being selected, players are offered a contract. In most cases the offer will be within a few percentage points of what the player drafted at the same slot was paid the year before. This initial contract will dictate their salary as they progress through the minor baseball leagues and enter the major leagues. Once players reach three years of accrued time at the major league level, they are eligible to negotiate their next contract. However, they can only negotiate with the team that drafted them. If they are unable to come to an agreement with the ball club, they can request to have a panel of arbitrators hear their dispute. The panel will listen as both sides plead their case. The panel then selects which side wins. The panel cannot split the difference or select a different salary than what the two sides proposed. Many cases are settled before the panel hears the case as neither side wants to lose. However, every year many cases go to arbitration. Players that reached six years of major league accrued time and have fulfilled their existing contract become free agents and can negotiate with any of the 30 teams. This study will help us understand if the arbitration process results in players with limited leverage (can only negotiate with one team) are paid similarly to players who already reached free agent status (can negotiate with any team).

We first looked at position player 2023 statistics, excluding pitchers since their statistics are not comparable to position players. After cleaning our data and prior to clustering, we ran regression analysis and predicted the arbitration player salaries. This served as our baseline results. We then used various clustering methods to group players with similar performance statistics and salaries. We reviewed and interpreted the results and chose the clustering algorithm and number of clusters that made the most sense to us contextually. We ran multiple regression methods for each cluster based on salary (the response) and their statistics (the predictors). For each cluster we selected the method with the lowest RMSE score and used it to predict the salaries for each of the arbitration eligible players in that cluster. We evaluated our results, comparing the predicted salary vs. the actual salary of the arbitration eligible players. We measured the mean difference and standard deviation between actual and predicted salaries for each cluster. We concluded that clustering players based on their statistics and salary allowed us to reduce the mean difference and standard deviation significantly compared to our baseline results. We were able to identify which clusters of players were underpaid, overpaid, and paid fairly. We also noted that there can be future work done on improving the model, including adding data from prior seasons.

DATA SET

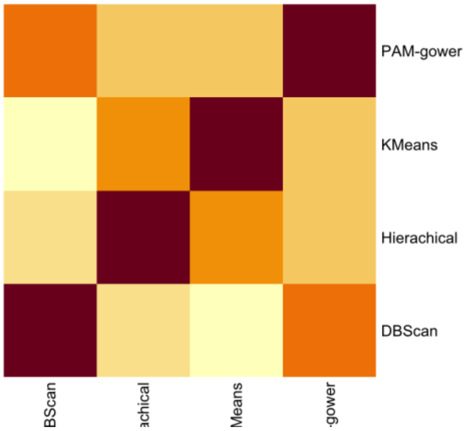
Our data set is comprised of 583 position players who played in MLB for the 2023 season which we found on spotrac.com. The data includes stats such as at-bats, hits, RBIs, home runs, stolen bases, batting average, on-base percentage, slugging percentage, age, salary, and whether they are eligible for arbitration for 2023 or not. We filtered out the players who played less than 60 games during the season, whether it was due to injury or going back and forth between the minor and major leagues. We noticed many players in their late 20s were making the league minimum of $720,000.00 or less, where some only had a few game appearances while others played most of the season. Salaries ranged from as little as $200,000 to $40,000,000 (Aaron Judge, NYY). Players' ages also had a wide range, from 20 to 43 years old. One of the main limitations with our dataset is that we were only using one season's worth of statistics, limiting our sample size for each player. Some players who were among the top earners played less than players making the league minimum or had worse statistics, potentially creating many outliers.

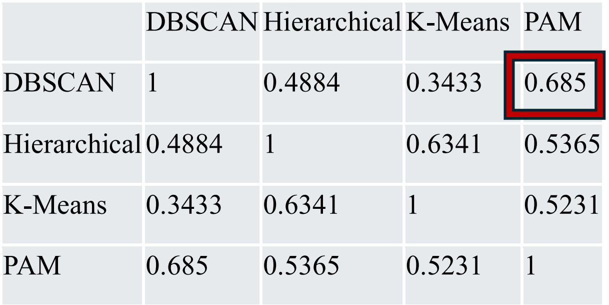
THEORY AND METHODS

Our assumption is that not all players' salaries are evaluated on the same metrics. For example, some players are paid to hit home runs while others are paid for getting on base. Therefore, different players have different variables that predict their salary. Due to this, we wanted to cluster the players to create multiple subsets of the data with statistically similar players in each cluster. Once the clusters were created, we tested multiple regression models for each cluster and utilized cross validation to pick our best model. We understood that each cluster might have a different best model.

Prior to clustering, we created a regression model using all the data and predicted the arbitration player salaries. The mean difference between predicted and actual salary was 55%. The standard deviation of the differences was 97%. This became our baseline model. The next step was to apply a variety of clustering methods to segment the players by their statistics and salary. We used DBSCAN, Euclidean Hierarchy, K-Means, and PAM-TSNE to cluster the data. We selected the number of clusters for each method using elbow curves, and silhouette plots.

We ran each clustering method with the selected number of clusters, created a similarity matrix using RAND-index scores, and plotted a heat map. We ended up choosing the PAM method for its high similarity score and its ability to make the clusters easily interpretable via the medoids.





	DBSCAN	Hierarchical	K-Means	PAM
DBSCAN	1	0.4884	0.3433	0.685
Hierarchical	0.4884	1	0.6341	0.5365
K-Means	0.3433	0.6341	1	0.5231
PAM	0.685	0.5365	0.5231	1

ANALYSIS

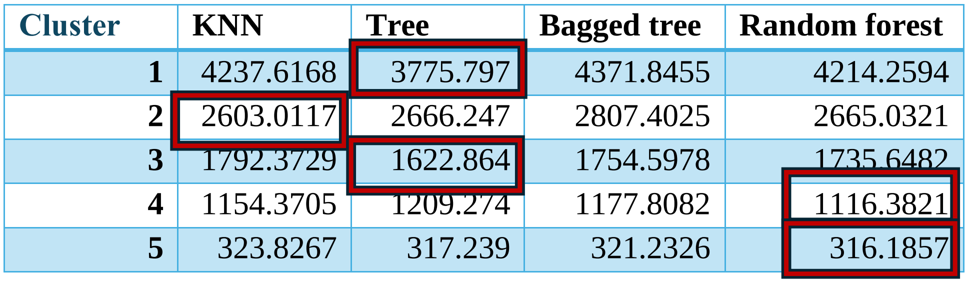
After applying the PAM method, we conducted an analysis on the cluster medoids, to understand what each cluster contained, how the players were grouped, and how their salaries were justified. The medoids show that the players are clustered in a logical way. Cluster one, has the 'Superstars.' The salary, batting average, on-base-percentage, slugging percentage, and OPS are the highest. Slugging is one of the most important variables for players, as it equates to high entertainment, which in turn, is an important variable for salary justification. As you go down the clusters the batting average, slugging percentage, and on base percentages are reduced, as are the salaries.

PAM Medoids

#	Age	G	AB	R	H	X2B	X3B	HR	RBI	SB	CS	BB	SO	SH	SF	HBP	AVG	OBP	SLG	OPS	Salaries ($K)
12	33	90	360	76	112	21	2	17	51	14	2	44	71	0	1	5	0.311	0.393	0.522	0.915	29000
39	29	138	515	75	147	31	6	20	74	6	7	59	130	0	7	2	0.285	0.357	0.485	0.842	18000
76	32	142	478	64	122	24	3	21	74	3	4	34	122	1	2	1	0.255	0.305	0.45	0.755	9000
129	33	123	422	42	106	18	1	17	61	0	0	31	132	0	2	2	0.251	0.304	0.419	0.724	4200
385	28	72	238	27	55	14	0	9	21	5	0	28	67	0	0	0	0.231	0.312	0.403	0.715	770

We ran multiple regression models with cross validation for each cluster and, within each cluster, evaluated the RMSE's to select which regression model worked best for that cluster. We ran the KNN, Decision Tree, Bagged Tree, and Random Forest methods for each cluster.

RMSE Table



Cluster	KNN	Tree	Bagged tree	Random forest
1	4237.6168	3775.797	4371.8455	4214.2594
2	2603.0117	2666.247	2807.4025	2665.0321
3	1792.3729	1622.864	1754.5978	1735.6482
4	1154.3705	1209.274	1177.8082	1116.3821
5	323.8267	317.239	321.2326	316.1857

We used the best model per cluster (lowest RMSE) from the process described above to make salary predictions of the arbitration players that fell into each cluster. Looking at the figure below this text, there is an “Actual Salary” column and a “Predicted Salary” column that contains the predictions. To the right is the Delta column, which tells whether the model overestimated or underestimated the salary prediction. There were no arbitration players in cluster 1.

Actual vs. Predicted (first 10 players)

First	Last	Cluster	Actual Salary	Predicted Salary	Delta
Juan	Soto	2	$23,000,000	$17,989,899	22%
Pete	Alonso	2	$14,500,000	$16,590,909	-14%
Gleyber	Torres	3	$9,950,000	$9,226,795	7%
Anthony	Santander	3	$7,400,000	$9,260,281	-25%
Christian	Walker	4	$6,500,000	$4,700,289	28%
Alex	Verdugo	4	$6,300,000	$4,572,668	27%
Luis	Arraez	4	$6,100,000	$4,606,039	24%
Mike	Yastrzemski	4	$6,100,000	$4,483,754	26%
Kyle	Farmer	4	$5,585,000	$4,466,519	20%
Kyle	Tucker	4	$5,000,000	$4,554,702	9%

Looking at the summary results below, we can see that clustering and choosing the best model per cluster greatly reduced the variability between predicted and actual salaries. The mean delta ranged from 1.5% to 23%. The standard deviation of the delta was reduced from 97% in our baseline results to 31% or less in our proposed model. The model predicted that the lowest tier players were overpaid, but keep in mind their average salaries are close to the minimum. The model predicted that mid tiers players were underpaid. There were only two higher tier players, and the model predicted that they were paid fairly.

Statistics for the difference (delta) in predicted vs. actual salaries

Cluster	min	Q1	median	Q3	max	mean	sd	n
2	-0.16	-0.0725	0.015	0.1025	0.19	1.5%	24.7%	2
3	-0.28	-0.1975	-0.115	-0.0325	0.05	-11.5%	23.3%	2
4	-0.82	-0.4025	-0.14	0.055	0.26	-18.8%	30.9%	28
5	-0.29	0.0575	0.275	0.4625	0.59	23.3%	25.5%	28

CONCLUSION AND FUTURE WORK

We felt the approach of using clustering to group players and select optimal regression models per cluster greatly improved our ability to predict salaries as compared to the baseline method with no clusters. Nonetheless, something that we want to change during future revisions of this project is the fact that we only had 1 year of data. We feel that to make better predictions, it would be crucial to have multiple years of data, especially the years prior to players signing their contracts, as this would help us understand salary justification a lot better. There are several examples of star players who signed large contracts, but after earning those contracts, started to perform poorly, and these players impact our predictions. With more years of data, we hope to reduce the impact of these players on our predictions, which will make this analysis more helpful to player agents before salary negotiations.

REFERENCES

Spotrac. (n.d.-a). *2023 MLB Arbitration*. Spotrac. https://www.spotrac.com/mlb/arbitraion/_/year/2023

Spotrac. (n.d.-b). *2023 MLB Salary Rankings*. Spotrac. https://www.spotrac.com/mlb/rankings/player/_/year/2023/sort/cap_total