COMPSCI 216: Everything Data

Baseball Analysis: Final Report

Eli Baker, Jason Granato, Jason Levine, Michael McCullom, Jack Morgan

Contents

Introduction and Research Questions

In recent years, professional sports leagues have become more reliant upon statistical modeling and machine learning to guide their decisions in player acquisition and contract management. One league, in particular, stands out in its use of data science for team operation: Major League Baseball. Baseball has a staggering number of metrics; a fact evidenced by legendary sportswriter Jim Murray, who wrote "I don't know whether you know it, but baseball's appeal is decimal points. No other sport relies as totally on continuity, statistics, orderliness of these." Given the prevalence of statistics in baseball and our group's interest in the sport, we chose to pursue the sport as the subject matter of this report, primarily focusing on contract negotiation relative player value. From the perspective of a casual observer, we noticed that certain players tend to be underpaid or overpaid by a significant margin relative to their production. We wanted to investigate how a player's production from a perspective of their statistical output can predict whether they are currently over or undervalued, and what their next contract should look like based on their production.

Our research question is as follows: are players paid based primarily on their overall ability? Furthermore, can we identify which predictors beyond a player's overall ability will increase the contract they receive and how much will they get paid? The last question we had was could a general manager use the information we gathered for the benefit of their team, potentially finding players who the market will undervalue given their overall ability? In this analysis, we will be focusing on position players due to the lack of compatibility between statistics for pitchers and position players.

One of the greatest challenges for a baseball front office is to assess a player's value on the field, and understanding why a player is over or undervalued is essential to that goal. Baseball teams are worth anywhere from 900 million to 5 billion, so there is a massive market for predictive data analytics. Our model will aim to provide value to general managers and develop a baseline understanding of the challenges that MLB front offices face in evaluating talent.

Summary of Results

In carrying out our analysis we were able to answer our original research questions. First, unsurprisingly, MLB players' overall ability is correlated with how much they receive in free agency. From the linear regression model constructed on just WAR, it is determined that 59.1% (r^2) of the variability in AAV is due to players' WAR. We were then able to isolate which other predictors were statistically significant in predicting the residuals from our original model. We found that position, home runs, RBIs, and age all were statistically significant in predicting salary beyond a player's WAR. We added those predictors to our simple WAR-based model and improved the r^2 by .07. Lastly, we were able to create a model that relatively accurately predicted which players would receive contracts either above, below, or commensurate with their overall ability.

Data Sources

For our project, our data was sourced from two primary locations. The first was Sportrac.com, which provided us with year-by-year lists of the MLB free agents and the resulting contracts they signed. We were able to copy the Sportrac data into CSVs and read them in python. We looped through the files line by line and eliminated all players with “p” in their position column, which stands for pitcher. We also isolated players who signed MLB contracts in free agency, designated by a number greater than 0 for salary and without “(minor)” in the line. We then joined each year's data together into one comprehensive DataFrame containing all of the MLB batters who signed a free-agent contract over the last 9 years. The other data source was the pybaseball package in python. This package provided us with 319 columns worth of data for each MLB player by season. This data originally came from Fangraphs, a popular site that tracks various MLB statistics. We initially struggled to use pybaseball because the default method filtered out non-qualifying players, but we were eventually able to include all players in the DataFrame by adding an argument to pybaseball’s function. We performed a merge on our two DataFrames by name and year to create the final DataFrame, which contained all of the free agents with their statistics from the previous year.

Results and Methods

Our first step was to develop a preliminary regression model based entirely upon the statistic of wins above replacement, or WAR, to set a baseline for our players' worth. In baseball, WAR is fundamentally the most important statistic for evaluating a player’s worth by determining how many wins he is worth to the team more than a replacement player, which in the case of the MLB, is a minor league talent or a readily available free agent. The calculation for WAR takes into account many different player efficiency metrics to provide a comprehensive value that indicates how many additional wins a team would have with this player in the lineup. We will use WAR as our baseline statistic because winning is definitively the most effective way to determine team success from a financial standpoint. Winning teams engage fans to a greater extent, sell more tickets and merchandise, receive more viewership on large networks, and get to play in playoffs games and contend for a championship. Of course, there are other factors in determining a team’s relative earnings like market size and history, but from a general manager’s perspective alone, building a winning team is the best way to ensure revenue.

With these details in mind, we chose WAR as our effective baseline to determine a player’s value, because their value comes in the form of bringing wins to the team. The first step in doing this was to compile our player data and ensure that we had access to each player’s most recent contract and their WAR, in addition to other statistics we might need for the rest of the project like runs, positions, etc. We could obtain all in-game player metrics from pybaseball, which requires a single line of code to produce our 8-year DataFrame of players and their numbers. The issue was that pybaseball does not include contract data, which is why we chose to use Sportrac to access all of the information that we needed. After importing this data into our notebook (the

process is explained in the Data Sources section), we then merged this data by name and season with our pybaseball DataFrame, which created a row for every player in every season from 2011 to 2019. From there we made some simple conversions from the string-based contract values into integers, and we were ready to make our first model.

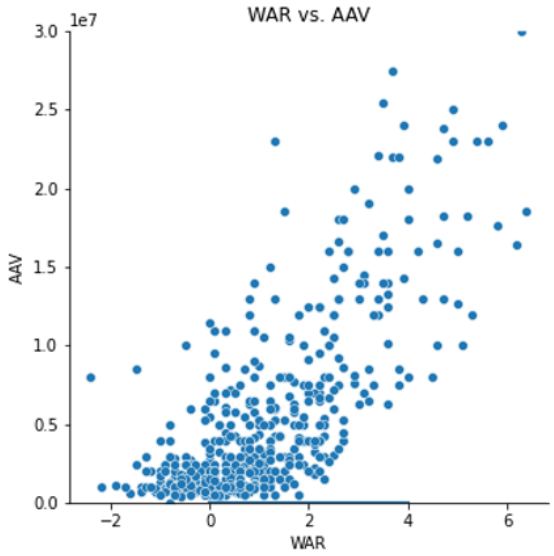

Figure 1: Relationship between WAR and AAV

To confirm our expectations we made a plot of WAR and AAV, and after seeing that there was clearly some form of a positive correlation, we created our model. Our first model used WAR and season as the only variables for the model, with average annual value, or AAV, as the target. We chose to include season as a variable because salaries have on average increased from year to year, so we did not want the data to be skewed by the higher salaries of more recent years. This model possessed an r^2 value of 0.591 and a mean squared error of 1.28×10^{13}, which is high but makes sense given our contract values in the millions. This r^2 value shows that there is a relationship between WAR and AAV, but there are certainly several other factors that can explain this relationship, leading us to our next step.

We first decided to account for home runs, a rather flashy player metric that we predicted would be found in overvalued players. Of course, home runs are essential to winning in baseball, but given their ubiquity in the media, we predicted that they would be a statistic that is overvalued relative to a player's WAR. We divided our players into overvalued and undervalued players, meaning those with contracts larger than our WAR model predicted are overvalued and those with contracts smaller than predicted are undervalued. We then created the below visualization comparing these player's home runs:

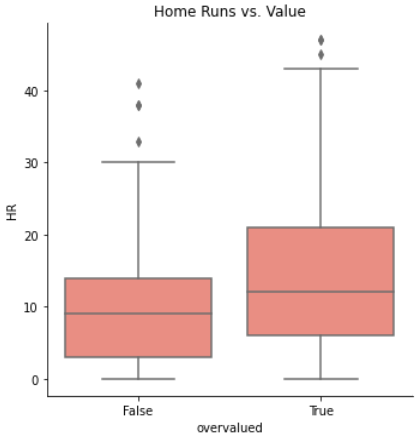

Figure 2: Player Value and Home Runs

As we predicted, overvalued players tend to hit more home runs than undervalued players.

We then looked at player age, a factor that we recognized would likely have an impact upon contract value in MLB players.

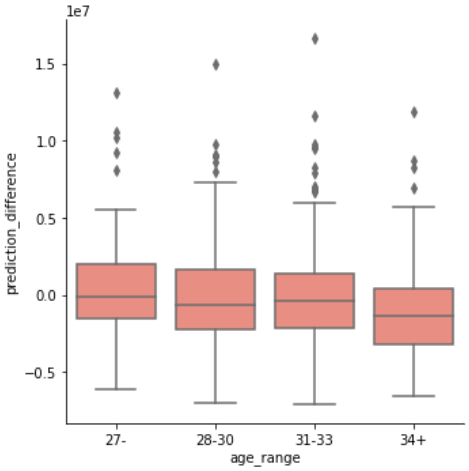

Figure 3: Player Value and Age

The above plot highlights four different player age brackets on the x-axis and the difference from our predictive model on the y axis. This difference represents a player's AAV minus their predicted AAV from our initial WAR model, meaning that positive numbers reflect an overvalued player and negative values an undervalued one. As we can see, younger players are generally overvalued and older players are usually undervalued. This makes sense, given that teams are willing to invest a significant amount into younger prospects in hopes that their longevity will pay off, and are likely unwilling to pay a veteran who has a higher likelihood of injury.

The next metric we looked at was position. While we have avoided pitchers in this report, there are still nine different positions among batters. We chose to reduce this number down to five different position groups to obtain more data per group, combining similar positions like “left fielder,” “center fielder,” and “right fielder” into a single “outfielder” category.

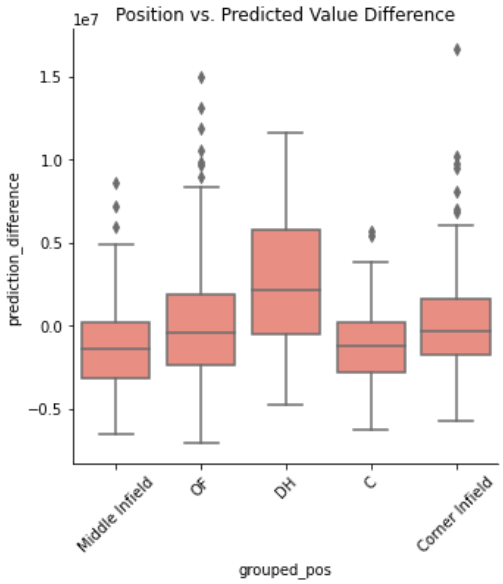

Figure 4: Player Value and Position

As we can see, designated hitters tend to be overvalued. Outfielders and corner infielders appear to be valued as expected, and catchers and middle infielders tend to be undervalued.

Finally, we looked at handedness when it comes to batting. Left-handed batters traditionally tend to do better against right-handed pitchers, who are the majority. We wanted to see if they were potentially undervalued or overvalued when it came to their contracts.

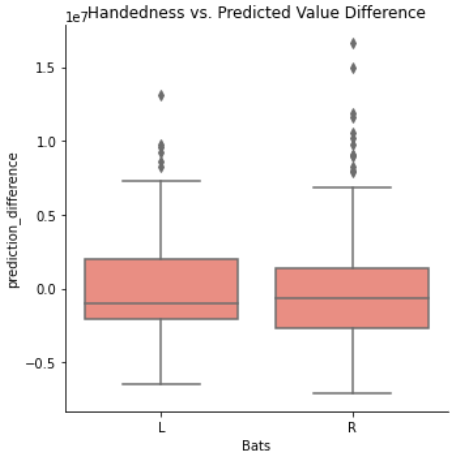

Figure 5: Player Value and Handedness

While the median values were similar, we found that the first and third quartile of the left-handed box plot was higher than that of the right, potentially suggesting that left-handed batters are overvalued, so we decided to take a further look at this metric as well.

We looked at other statistical categories and found nothing noteworthy in our visualizations, so we chose not to include such results in this report, moving forward to hypothesis testing our metrics of interest.

To confirm our hypotheses about the significance of these player statistics towards their contract value, we conducted four two-sided t-tests. We chose not to do a hypothesis test for the position statistic because it would be impossible to make two distinct groups to carry out the test (would need to use an ANOVA test, which we haven't studied in this class). The table below gives a summary of the results of the tests along with a brief description of the values that were being tested.

Table 1: Hypotheses tests for meaningful variables

Null Hypothesis	**Alternative Hypothesis**	**P-Value**	**Reject or Fail to Reject**
Overvalued players do not hit more home runs than undervalued players	Overvalued players do hit more home runs than undervalued players	1.06×10^{-6}	Reject
Lefties are not valued differently than righties	Lefties are valued differently than righties	3.76×10^{-1}	Fail to reject
Overvalued players are different ages than undervalued players	Overvalued players are not different ages than undervalued players	4.31×10^{-3}	Reject
Overvalued players have a different number of RBIs than undervalued players	Overvalued players do not have a different number of RBIs than undervalued players	1.16×10^{-8}	Reject

From the results, it is evident that many variables outside of WAR are significant when looking at contract value. We saw a significant difference in our predicted value difference statistic (actual AAV - predicted AAV) for three of the four tests run, specifically home runs, RBIs, and age. In addition, it is evident from Figure 4 that there is a significant difference between each position. Therefore, we decided to add these significant variables to our refined linear regression model in an attempt to increase its precision. To include our grouped position variable, we used a

One Hot Encoder to convert these groups into usable data for the model. After running this updated linear regression, the resulting r^2 was found to be 0.664, a .07 increase from the simple WAR model. Although this increase is not huge, it does prove the impact adding other significant variables has on the strength of the regression.

Lastly, we set out to create a K-NN model that would predict which players would sign contracts that either undervalued, correctly valued, or overvalued their overall ability. We first had to create a column in our DataFrame for each player that stated if their actual salary was less than 70% of their predicted value (Undervalued), between 70% and 130% of their predicted value (Properly Valued), or more than 130% of their predicted value (Overvalued). We then performed a train-test-split on our data. From there, we executed a GridSearchCV to find the number of nearest neighbors that would result in the highest accuracy score for our KNN model. In the end, we had a 25-NN Model with an accuracy score of 52.05% and a confusion matrix as pictured below.

Figure 6: Confusion matrix for K-NN Model

The 25-NN could predict whether the player would be overvalued, properly valued, or undervalued, with 52.05% accuracy based on the player's age, WAR, HRs, RBIs, and position. From the confusion matrix, the 25-NN model predicts when a player is undervalued at around a 69% rate but only predicts the other two categories at around a 39% rate.

These results, although better than random guessing, indicate that there are factors outside of a player's on-field production that affect their contract value. For instance, Mookie Betts led the league in 2021 for jersey sales. Therefore, he adds a great deal of value to the Dodgers outside of his all-star production. This example shows that we could increase the accuracy of our model if we were to gather data about financial stats like jersey sales, ticket sales, TV ratings, and many others due to a player's involvement with an MLB franchise.

Limitations and Future Work

Despite our project being very successful, some limitations hindered our model's success. The main dilemma we had to deal with during the early stages of the project was finding enough data to make meaningful conclusions. Since we were analyzing salary relative to how individuals performed, only the free agents of each year that were signed were included in the data set. This is because these individuals received a new contract, and that contract value was an integral part of the analysis performed. In addition, it is very hard to compare pitching and positional player stats, so we had to limit the data set to just fielders (removing pitchers). Therefore, trimming the data to the useful players, we only had around 60 individuals per free-agent class in the data set. To combat this issue, we were able to purchase access to 9 years of free agent salary data, increasing the data set to over 500 players. Although this isn't a massive amount of data compared to other fields of data analytics, we felt this was definitely enough substance to create a useful model and conduct a thorough analysis.

Another limitation the group faced was the significant impact COVID-19 had on Major League Baseball. The financial crisis that ensued due to this pandemic also affected sports teams, as they are a business just like any other industry. Therefore, new salaries were much more stringent during the 2020 season especially. In addition, some players sat out due to coronavirus concerns, and the season was shortened significantly. This greatly increased the variability in WAR and decreased the credibility of statistics when analyzing a player's actual ability. Subsequently, this variation decreased predictability and added randomness to our model, preventing us from creating a very refined final product.

With these issues come areas to grow in the future. While we did unsuccessfully search for more years of data regarding free agent signings, this additional data could most likely be found on other sites outside of Spotrac simply due to the vast amount of information available in today's world. Although we could not directly export the free agent salary info, if we had more time for this project, we could create a more advanced web scraper that would take online data and format it to our liking. In addition, another area we would like to explore in the future is to develop a model that takes data from multiple years, applying varying weights to each year based on recency and career maturity. This would be helpful because the model would have additional inputs that eliminate the ability of fluke seasons, either bad or good, from severely impacting a player's contract value. Lastly, factors in the contract itself affect the value outside of just yearly salary (AAV). For example, players often are willing to take contracts with a slightly lower AAV if it is for more years, as the total value is often higher and the guaranteed income lasts longer into an individual's career. In addition, there are many more complicated clauses involving backloading contracts, including performance bonuses, and more related terms that make value far from a black and white term. Adding these additional aspects of contracts into our model would refine it further and theoretically increase its effectiveness.

Appendix

A link to our DeepNote codebook is provided below (all of the data is imported into CSV's and can be viewed from DeepNote):
https://deepnote.com/project/Salary-Predictor-Final-Project-ytW36lAvQtGHYUDE390i6w/%2Fnotebook.ipynb