**Data Analysis of the**

**Arbitration Process in the MLB**

By

Alex Riles

Advised by

Dr. Tisha Hooks

Abstract

The arbitration process in Major League Baseball is used to prevent holdouts and prolonged disputes. The ability to predict the outcome of the hearing will help the players and teams know where they stand. This could change the amount a player or team submits, so the outcome could turn in their favor. This project used Neural Network and Random Forests models to investigate whether a predictive model is worth using to predict the outcome of arbitration. Results indicated that Neural Network and Random Forests models can be used to predict the outcome of the arbitration process with a low misclassification rate.

Introduction

Major League Baseball (MLB) has been called America’s Past Time due to it connects to America’s culture and its popularity. The MLB has led the way in cultural change with having the first African America player in Jackie Robinson, the MLB player union being established in the 1950s, and being the leader in changes in the statistical world when it comes to sports. The changes in baseball statistic started in the 1950s with statistics being put on the back of baseball cards, then the use of statistic to find player and predicts win with Money Ball in the early 2000s, and then to sabermetrics in the 2010s. The MLB has been ever changing in the way they use statistics to evaluate a prospect’s talent, current talent around to the league, and develop game plans throughout the course of a season.

For my research I wanted to investigate statistics of players and see if it is possible to predict the outcome of the arbitration process. The ability to see if it is possible to accurately predict the outcome of the arbitration process will allow teams and players to see if they are going to win the process or lose. This could influence the teams or players in maybe trying find a new deal with each other or change what they think they are worth. I will be using Neural Network and Random Forests to see which one performs better in predicting the outcome of the arbitration process.

Background

The Major League Baseball’s Collective Bargaining Agreement (CBA) is the contract that is signed between The MLB as the organization and the Major League Baseball Players Association (Players Union). The contract that is signed between these two-different organization can be view as the same contract that is sign between the Coal companies and the Labor unions that represent the coal miners. According to “Determinants of Major League Baseball Player Salaries” the product of a CBA has an agreed upon wage, working conditions, and many other things that are specific to the field of work. For the CBA that is signed between The MLB and the Players Union it contains the following: season length, minimum salary, revenue-sharing, luxury tax, the draft set up, arbitration, and many other parts. With making sure that the different parts of the CBA are up to data there is a length to the CBA (Wasserman2013). The length of the CBA could be anywhere from 5 to 7 years long. When the CBA is up The MLB and The Player Union renegotiate on updating the CBA. There has been time where The MLB and The Players Union have had major disagreements which have resulted in a lockout or strike. The most recent time this has happen in the MLB was in the 1970s and the lockout lasted half of the regular season. When the lockout happened in the 1970s one of the things that was on that CBA was Arbitration.

According to “Baseball Arbitration: An ADR Success” the arbitration process was established in 1973. This was a big deal because it gave teams ownership of a player for the early part of the player’s career and yet gave the player the freedom to find player somewhere different in the later part of their career. There were secret agreements were called the “reserve rule”. It was found that this rule was started in the 1879 and was that each team could have a list of five players that each team would agree upon would not touch even if that player’s contact was to expire (Monhait2013). This is a big deal because it didn’t allow player to test out the free agent market like you would see players do today. So, if Babe Ruth was on this list for the Yankees and Babe Ruth wanted to play for a different team like the Brooklyn Dodgers, he wouldn’t be able to do so because of this reserve rule. This also allowed owner to keep the player salaries down because there wouldn’t be a bidding war like you would see today. This reserve rule became a very big deal when it was made public. This became such a big deal the according to “Baseball Arbitration: ADR Success” there were three different Supreme Court cases about the issues. The Supreme Court cases happened in 1922, 1953, and 1972 (Monhait2013). Because of owner allowing player to increase their rights the Major League Baseball Players Association was established in 1954 to improve player right in MLB.

Now after the third Supreme Court case the arbitration process was established which was a middle ground for the MLB and the Players Union. Other than keeping a player with a team for the first six year of the MLB career according to “Baseball Arbitration: An ADR Success” some of the goals of arbitration was to have the player and team mutually agree upon the player’s salary for the next season and avoid lockouts (Monhait2013). This is a good thing because this help avoid situation like you see in the NFL of players sitting out to get more money. This still happens in baseball, but it only happens with players who have more than six years of experience and have been through the arbitration process. The arbitration process only happens to a player when they have three to six year of MLB service. Just because a player has three to six years of service doesn’t mean he has to go through the process. The player and team could agree on a contract to bypass the arbitration process. If a player and team wants to go through the arbitration process, then at the beginning of the free agency period the team would offer arbitration to the player and the player would then accept arbitration. Then if the player and team can’t come up with their own agreed upon contract then it goes to a third party and the arbitrator will look at what the player thinks he is worth and what the team thinks the player is worth. After looking over what both sides think the player is worth the arbitrator will decide what the player is worth between the player’s and team’s salary. When the decision is made it is final and that the player’s salary for one year.

DATA

For the data that is used for predicting the outcome of which side the arbitration will decide with is a mixture of traditional statistics, sabermetrics, and arbitration salaries. To process of collecting the traditional stats, sabermetrics, and arbitration salaries was through a combination of pull data of different data sets and manually make the data of different formulas. The databases and websites where the data and formulas for the data were found was “Lahman’s Baseball Database”, “Fangraph.com”, “Baseball-reference.com”, and “mlbtraderumors.com”. It was necessary to divide the data by position players and pitchers because of how different the statistics are for each group of players. With the MLB being around since 1879 and arbitration only starting in 1973 there is a lot of data that could be used in making a predictive model. To narrow down the data and to use the most recent data the data was set between 2011 and 2018.

When looking at the position players data the predictors that were used were Wins Above Replacement (or WAR), On Base Plus Slugging Plus (or OPS+), Defensive Runs Saved (or DRS), Batting Average (or AVG), Runs Batted In (or RBI), Runs Created (or RC), the amount the player put up for arbitration, the amount that the team put up for arbitration, and the midpoint between the player’s amount and the team’s amount. According to “What is WAR?” is that WAR is a summary of the total contributions to the team that a player plays for. WAR was not meant to be the perfect indicator of a player’s contribution but it should be used more as a comparison tool (Slowinski2012). This stat was taken from fangraph.com but if you want to look at all the formulas, position adjustments, and the good rule-of-thumb chart look at Appendix A. OPS+ is the On Base Percentage plus the Slugging Percentage with adjustments for league average and baseball park factors. On base Percentage (OBP) is the amount of time the player gets on base safely regardless of hit, walk, or hit by pitch over the amount at bats a player has. Slugging percentage (SLG) is the total bases over the amount of at bat a player has. So, for total bases a single is 1, a double is 2, a triple is 3, and a homerun is 4. This statistic was calculated manually using the formula found on baseball-reference.com and can be found in Appendix B. According to “DRS” DRS is “How many runs better or worse that player has been relative to the average player at his position.” (Slowinski2010) The DRS statistic was found at fangraphs.com and there is a table that breaks down how defensive ability relates to DRS in Appendix C. The batting average of a player was found in the Lahman’s Baseball Database. The way batting average is calculated is based on how many times the player reacts base safely when the ball was hit in play and there aren’t any errors on the play over the amount of at bats a player has. A good player would have a batting average over .300. The Runs Batted In statistic was found in the Lahman’s Baseball Database. RBI is the total amount of runs a player drove in with getting a hit, walk, or getting hit by a pitch. The Runs Created statistic is the total amount of runs that a player created with looking at more than getting a hit like the base running of the player. The RC was calculated manually with the formula from baseball-reference.com and the formula can be found in Appendix D. All the salaries that the players and teams submitted to the arbitrator along with the midpoint was founded at mlbtraderumors.com.

When looking at the pitcher data the predictors that were used were Wins Above Replacement (or WAR), Field Independent Pitching (or FIP), Walks Hits per Innings Pitched (or WHIP), Innings Pitched (or IP), Strikeouts (or SO), Earned Run Average (or ERA), the amount the player put up for arbitration, the amount that the team put up for arbitration, and the midpoint between the player’s amount and the team’s amount. WAR is the only statistic in baseball that can be used for both position players and pitchers. There is a difference is the way it is calculated due to the difference in the stats for each set of players that is used in the formula. The statistic was found at fangraph.com. If you are interested in the formula for WAR for pitchers in can be found in Appendix E. According to “FIP” Fielding Independent Pitching is based on the outcome that are due to result that come from strikeouts, walks, hit by pitch, and homeruns (Slowinski2010). The reason this statistic is used in the comparison of pitchers is that pitchers can’t get negatively impacted by their team that they pitch for having a bad defense. The statistic only accounts for balls that the pitch can control. FIP was manual calculated with that formula and table show was a good and bad FIP could be can be found in Appendix F if you are interested. Walks Hits Per Innings Pitched is a statistic that show the rate at which a pitcher gives a hit, walk, or hit by pitch per inning that the pitcher would pitch. This statistic was manual calculated by taking the total sum of hits, walks, and hit by pitch than divided by the number of innings a pitcher has thrown. Innings pitched is a statistic that shows how many times a pitcher got a hitter out throughout the course of a season. IP is recorded in thirds because there are three outs in an innings so innings can be record for example of 30 2/3 innings. That means the pitcher got 92 hitters out. This statistic in the Lahman’s Baseball Database didn’t come with IP it came with the number of outs that the pitcher recorded so it had to be manually recorded by taking the total amount of outs and divided by three. Strikeouts is a statistic that show how many times a pitcher got a hitter out with having the ball hit in play. This is done by getting to three strikes before the hitter can put the ball in play or before the pitcher hit 4 ball which would result in a walk. To get a strike there is a zone that is called “the strike zone” the ball must travel through is zone to be called a strike if the batter elects not to swing the bat. If the batter decides to swing the bat, were the ball ends doesn’t matter as long at contact isn’t made with the baseball. Earned Run Average is a statistic that shows how many earned runs a pitcher will give up for every 9 innings of work. An earned run happened when a runner scores a run without a field error or throwing error made by the defense. The difference between ERA and FIP is that a pitcher’s ERA can be positive or negative be affected by their team’s defense, whereas FIP statistics would have no such case happening. ERA was found in the Lahman’s Baseball Database.

The response variable for the predictive models is a categorical variable that shows the outcome of the arbitrator’s decision at the end of the arbitration process. When the arbitrator makes his or her decision, they don’t have to choose the exact amount that the player or team gave them. They could choose a number that falls in between those two numbers. With the misclassification rate being very high when trying to predict the exact salary the arbitrators will settle with. I thought I would be a better idea to try and predict which side the arbitrator will go with. So, the outcome variable was divided in to three categories; the arbitrator side with the player, team, or the midpoint between the player’s and team’s amount. This had to be manually created and that was done by assigning a 1 to any player who settled amount was between what they said they wanted and the midpoint, a 2 when the settled amount equaled the midpoint between the player’s and team’s amount, and a 3 when the settled amount was between the midpoint and the team’s amount. There wasn’t and shouldn’t be any cases where the player’s amount is lower than the team’s amount because then the player and team should have just settled with a contract outside of the arbitration process.

When reading in all the data I used R coding language to edit, clean, calculate, join, stack, and run the models that were done for this paper. All the code that I did for this project to run the models, and edit the data can be found here on my GitHub <https://github.com/Riles567/Capstone>.

Model Development

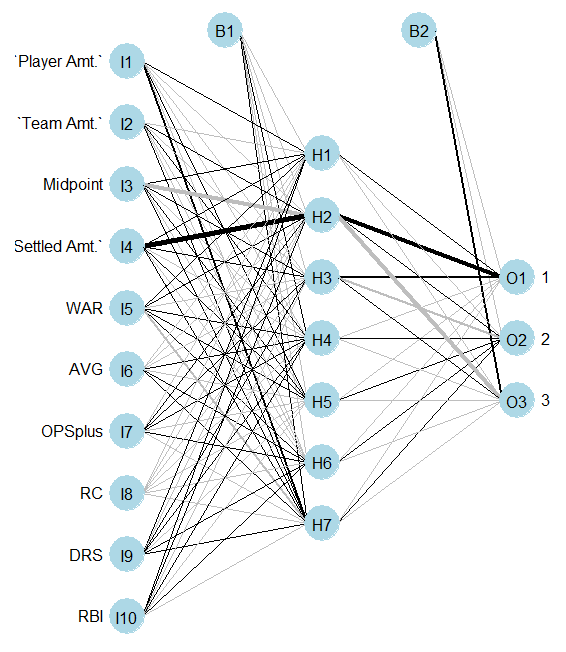
Neural Network

Before starting to run the Neural Network model the data should be transformed to have a normal distribution so that if there are any outliers they won’t affect the accuracy of the prediction that are being made. In determining what transformations that need to be made the “BoxCox” method was used. I did transform all the predictors make sure they were all had a normal distribution.

After the transformations were made for both the position players and pitching data, running and making adjusts to the Neural Network can be made. With having two sets of data each one need to be adjusted on their own to make sure that the Neural Network can make the best predictions. For both the position players and pitching data having just one layer on nodes showed the best result in the misclassification rates.

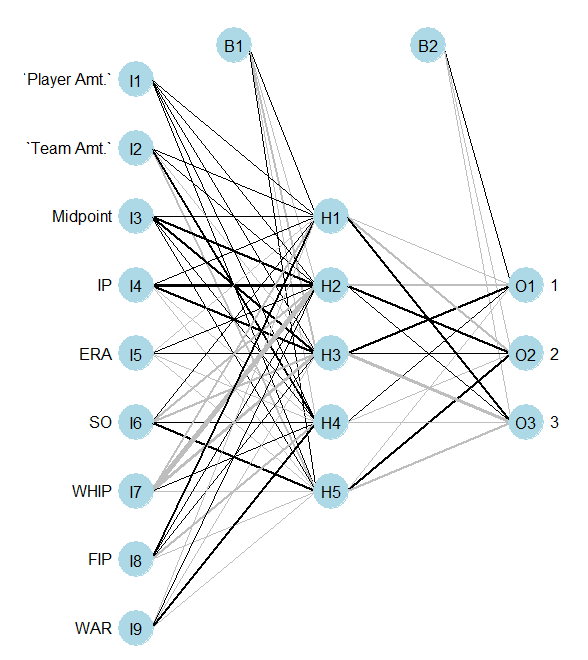
For the position players Figure 1 below shows a graphic of what the Neural Network looks like in its final model. As you can see there are 7 hidden nodes that receive the data from each predictor and then decide what categorical data it should predict. Neural Networks only travel in a one-way line, like our brains. The nodes only can receive data from the predictors and only send data to the response. The number of hidden nodes that can be used in the model can be adjust and through trial and error I found that 7 hidden nodes did the best job.

Figure 1 graphic showing the Neural Network model for the position players data



For pitchers Figure 2 below shows a graphic of what the Neural Network looks in its final model. As you can see there are 5 hidden nodes that receive the data from each predictor and then decide what categorical data it should predict. Through trial and error, I found that 5 hidden nodes did the best job at predicting the outcome.

Figure 2 graphic showing the Neural Network model for the pitchers data

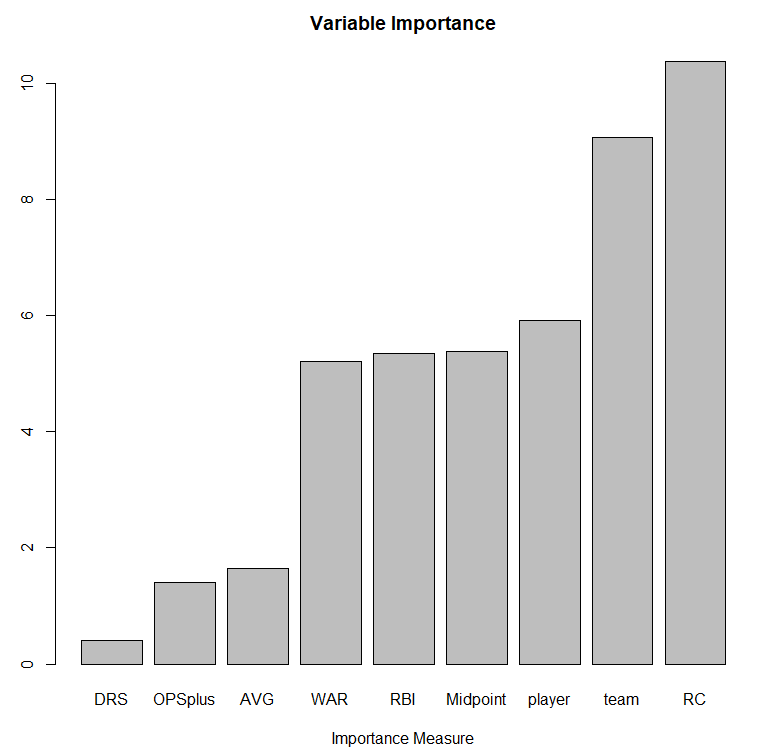


Random Forests

For Random Forests transformation don’t need to be made to the data because tree-based model is invariant to transformations of the predictors. Random Forests is a bootstrap-based tree model that incorporates randomness into each tree being made and take the overall average of all the trees being made. The number of trees that were being made in the Random Forests was 500 trees and that was the same for both the position player data and the pitching data. There was two options that were adjust in the improvement of the Random Forests model was the number of variables to randomly pick at each stage (mtry) and the maximum number of terminal nodes to use in the bootstrap trees (maxnodes).

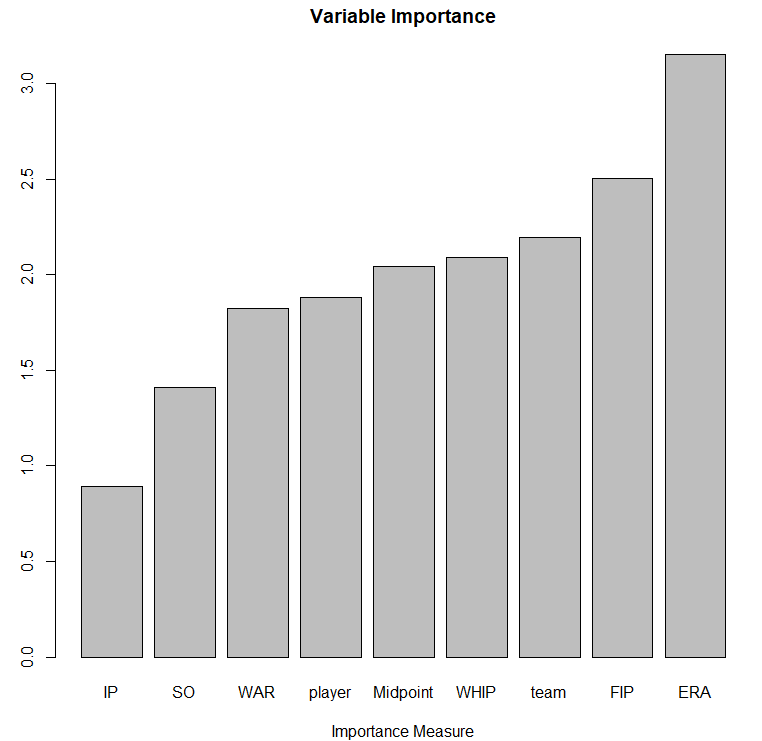
When adjusting the mtry and maxnodes for the position player it was a trial and error process. To start the trial and error process, mtry and maxnodes started at 5. After setting them both to 5 the misclassification rate was found to have a base line to compare to the changes that were going to be made to the mtry and maxnodes options. It started with keep maxnodes constant and changing the mtry option. After several changes to the mtry it was found that setting mtry to 5 would be the best option. Then having mtry constant the same process was done for maxnodes and it was found that setting maxnodes to 5 was the best options. One of the nice features of a Random Forests is that the variable of importance can be found quite easily. Looking at Figure 3 you can see that for position player the variable with the most importance was Runs Created followed by the amount of money each side gave to the arbitrator. The one thing that is interest is that OPS+ wasn’t more important due to how much homeruns are talk about and shown in the sports media world.

Figure 3 graphic showing the importance of each variable in the Random Forests with the position players data



The process in finding the final Random Forests model for pitcher was done in the same way as the position players model. The results were a little different when it came to what mtry and maxnodes were set at. The final Random Forests model for pitchers had mtry set to 3 and maxnodes set to 5. When looking at the variable of importance in Figure 4 you can see that ERA and FIP was at the top of the importance follow by the team’s amount of money and then WHIP. It could have been predicted that ERA and FIP would have the biggest importance in determining on the outcome of the arbitration process.

Figure 4 graphic showing the importance of each variable in the Random Forests with the pitchers data



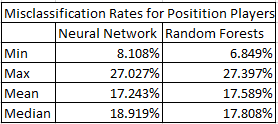
Cross Validation

To look at the accuracy of the Neural Network and Random Forests model that were made with the position players and pitching data a cross-validation must be done. The cross validation that was done in this paper was a Monte Carlo split sample cross validation. The cross-validation functions that were used were taken from Dr. Deppa class handouts. The functions found the minimum, maximum, mean, and median of the misclassification rate for all the iteration that the cross-validation went through. The misclassification rate is a percent of how many times that the model predicted wrong. With trying to find if Neural Network or Random Forests did a better job of predicting the outcome of the arbitration process, they will be compared to each other with the position players and pitching data.

Results

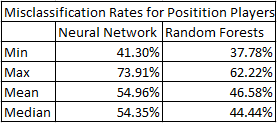
The cross-validation result for the position players data was acceptable for both Neural Network and Random Forests models. Looking at the misclassification rates for Neural Network the minimum misclassification rate was 8.108%, maximum was 27.027%, mean was 17.243%, and median was 18.919%. For Random Forests the misclassification minimum was 6.849%, maximum was 27.397%, mean was 17.589%, and median was 17.808%. When comparing these two models the most important misclassification rate to look is the mean and when looking at the mean they are close together. The Neural Network model does narrowly inch out the Random Forests model by .3%. Now .3% isn’t that big a difference to say Neural Network does a better job than Random Forests. If you are someone who doesn’t like transforming their data a lot, then you might prefer Random Forests over Neural Network and be ok with taking that .3% hit to being able to predict the correct outcome. I am choosing the Neural Network over Random Forests when predicting the outcome for position players.

Figure 5 Table to show the comparison between Neural Network and Random Forests misclassification rates for posisiton players



Now looking at the pitching data the results of the cross-validation for both Neural Network and Random Forests did a lot worse at predicting the outcome of the arbitration process. Looking at the misclassification rates for Neural Network the minimum misclassification rate was 41.3%, maximum was 73.91%, mean was 54.96%, and median was 54.35%. For Random Forests the misclassification minimum was 37.78%, maximum was 62.22%, mean was 46.58%, and median was 44.44%. Looking at the mean for both model you can see that the Random Forests model did 8.38% better at predict the outcome than the Neural Network model. I am choosing Random Forests to predict the outcome of the arbitration process, but I wouldn’t be confident at all getting the prediction right. The main issue that might cause the misclassification rate to be this high is the difficulty to tell if a pitcher is a starter or reliever based on statistics. Also, the could be the issue of what the player views themselves as compared to the team could also cause high of a misclassification rate.

Figure 6 Table to show the comparison between Neural Network and Random Forests misclassification rates for pitchers



Conclusion

When making a predictive model there is not guaranty that the model that is being create would be worth using. When looking the Neural Network and Random Forests models that were made for position players and pitchers some of the models were not worth using. The models that were created for the position player data did a good job predicting the outcome of the arbitration process with Neural Network having a mean misclassification rate of 17.2% with did .3% better than the Random Forest Model. The models that were made for the pitching data didn’t do so good at predicting the outcome of arbitration process with Random Forests having a mean misclassification rate of 46.58% which did 8.38% better than the Neural Network model. Now this poor performance could be explained to the difficulty of knowing how a team see which the pitcher or how the pitcher see themselves when it comes to be a starter or reliever.

Future Research

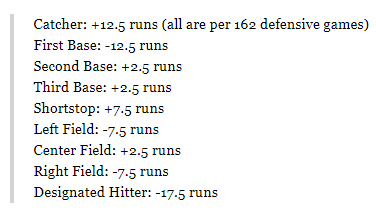
In the sport of baseball there are many areas where the use of models to predict outcome of different event could be created. For example, one might be a good predictive model to predict the winners of different awards like Most Valuable Player and Cy young award, expand past the arbitration process and predict how much a certain player should get in the open free agent market. Also, one might investigate how a player will perform better going through the arbitration process each year or sign a long-term contract to bypass the process.

References

* Monhait, Jeff. “Baseball Arbitration: An ADR Success.” 2013.
* Slowinski, Steve. “FIP.” *FIP | Sabermetrics Library*, 2010, library.fangraphs.com/pitching/fip/.
* Slowinski, Steve. “What Is WAR?” *What Is WAR? | Sabermetrics Library*, 2010, library.fangraphs.com/misc/war/.
* Slowinski, Steve. “WAR for Position Players.” *WAR for Position Players | Sabermetrics Library*, 2012, library.fangraphs.com/war/war-position-players/.
* Slowinski, Steve. “WAR for Pitchers.” *WAR for Pitchers | Sabermetrics Library*, 2012, library.fangraphs.com/war/calculating-war-pitchers/.
* “Batting Stats Glossary.” *Baseball*, www.baseball-reference.com/about/bat\_glossary.shtml.
* Slowinski, Steve. “DRS.” *DRS | Sabermetrics Library*, 2010, library.fangraphs.com/defense/drs/.
* Wasserman, Tyler. “Determinants of Major League Baseball Player Salaries.” 2013.
* “RStudio: Open Source & Professional Software for Data Science Teams.” *RStudio, PBC*, 3 Apr. 2020, rstudio.com/.

Appendix A

WAR = (Batting Runs + Base Running Runs + Fielding Runs + Positional Adjustment + League Adjustment +Replacement Runs) / (Runs Per Win)

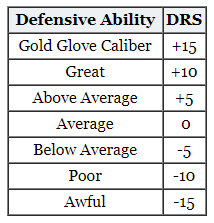
When it comes to calculating WAR by hand it is not possible because there are ceritian data that was develop by different people base on video clips of the player. More can be explained in the article “WAR for Position Players” that can be found at fangraph.com.

Appendix B

OPS+ = PRO+ = 100 \* ( OBP/lgOBP + SLG/lgSLG - 1)/BPF

BPF stands for Ball Park Factors. That statistic can be found at fangraph.com. The statistics lgOBP and lgSLG mean the league average for a given year for On Base Percentage and Slugging Percentage.

Appendix C



The table above shows how DRS relates to the defensive ability of a player.

Appendix D

RC = (H+BB-CS+HBP-GIDP)\*(TB+(.26\*(BB-IBB+HBP))+(.52\*(SH+SF+SB))))(AB+BB+HBP+SH+SF)

The formula above was used in the calculating of Runs Create during the project. It you want to make the Runs Created statistic and don’t have all these different stats there are different formulas at the baseball reference website.

Appendix E

WAR = [[([(League “FIP” – “FIP”) / Pitcher Specific Runs Per Win] + Replacement Level) \* (IP/9)] \* Leverage Multiplier for Relievers] + League Correction

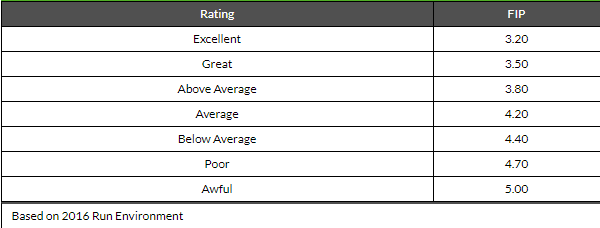


A more in-depth view on the WAR formula can be found at fangraph.com.

Appendix F

FIP = ((13\*HR)+(3\*(BB+HBP))-(2\*K))/IP + constant

FIP Constant = lgERA – (((13\*lgHR)+(3\*(lgBB+lgHBP))-(2\*lgK))/lgIP)



Sense FIP can account for league averages these rating and FIP that can be seen above can be a little different on a year by year bases.