Predicting MLB Career Success from Rookie Season Performance

Prepared by Jeremy Kozlowski and Hailey Terry

Sunday, December 10th 2023

Table of Contents

Introduction

This research project aims to investigate the predictive value of a baseball player's performance during their rookie season in relation to their overall career success. We will focus on using Major League Baseball data to assess performance. Baseball data has been kept for decades and is publicly available.

Problem Statement

Our primary research question is: Does a baseball player's performance during their rookie season predict their overall career success in the MLB? This is a useful question, as teams and general managers could leverage this information to provide cheaper, long-term contracts to young players who have a high indication of performing well throughout their career. Our secondary research question is: What statistics are the best predictors of career success? This project is personally interesting to us because we are fans of the sport, there is an abundance of data, and we have lots of players we can make predictions on since new rookies come into the league each year. If models can be made with low error, teams could save millions of dollars.

Data Collection

In this section, we shall discuss the details of data collection for this project.

Observational Units and Data Collection Process

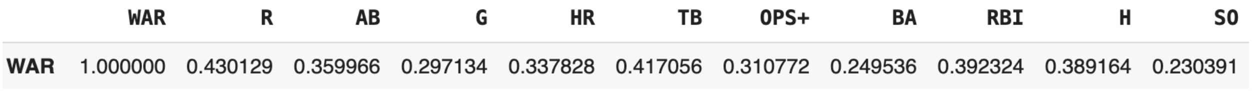
The observational unit in our primary data file is the rookie season statistics of a hitter who finished their careers between 1940 and 2023. We collected the data by scraping Baseball Reference [1], a popular and reliable website that hosts an abundance of baseball statistics. We first scraped the links of all eligible players for our dataset, and then requested each of those pages to extract their rookie season statistics. We employed column transformations, row filtering, and concatenation techniques to collect and preprocess our data. Scraping the data took 10+ hours and multiple attempts to run overnight. After multiple iterations, we were able to successfully combine our data in a CSV file.

Primary and Secondary Variables of Interest

Our primary variable of interest is Wins Above Replacement, more commonly referred to as WAR. WAR is a widely accepted metric that is used to evaluate a player’s performance. It calculates how many wins they bring to a team over the backup option to them otherwise known as their “replacement.” Secondary variables of interest include hitting statistics such as BA (Batting Average), OPS+ (adjusted On-Base Plus Slugging Percentage), R (Runs), RBI (Runs Batted In), and more. While we didn’t use all of these for our machine learning, these are ones that we determined were worth exploring in our investigation.

Compelling Data Visualizations

In this section, we shall detail our most compelling data visualizations.



	WAR	R	AB	G	HR	TB	OPS+	BA	RBI	H	SO
WAR	1.000000	0.430129	0.359966	0.297134	0.337828	0.417056	0.310772	0.249536	0.392324	0.389164	0.230391

Figure 1: The correlation coefficient between WAR and variables of interest

Figure 1 depicts one of the first descriptions we made, which was measuring the correlation between WAR and some critical variables of interest. The visualizations demonstrate the linear relationship between WAR and our chosen quantitative variables. While none of the correlations are high, there were some surprising relationships, including that between WAR and Runs (R) and WAR and Strikeouts (SO).

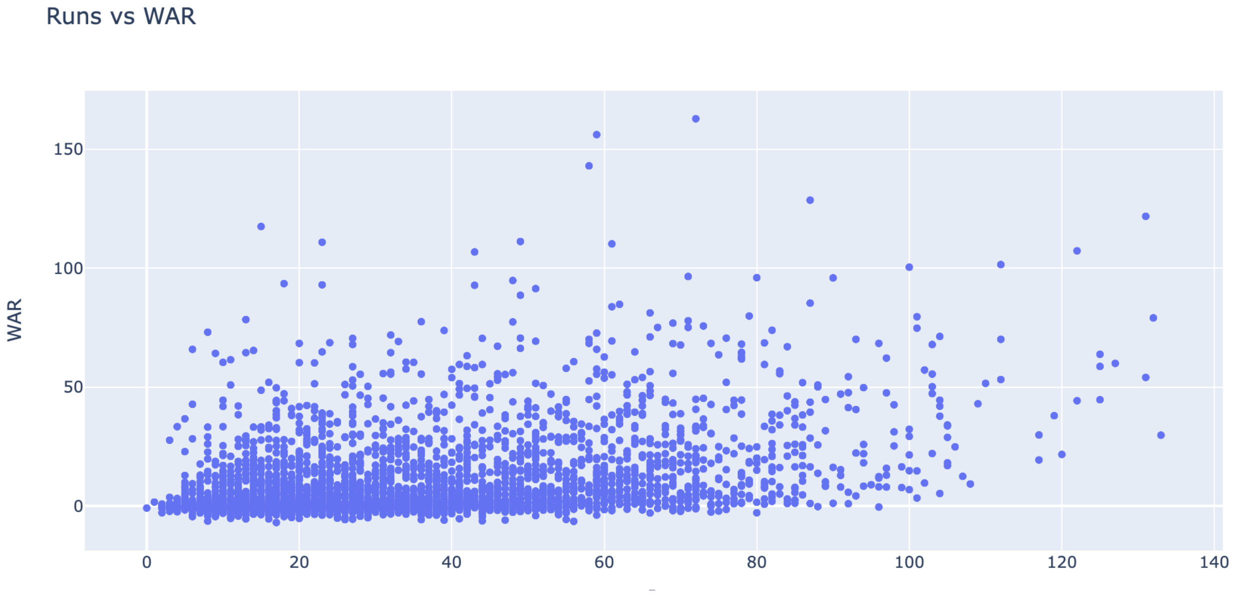


Figure 2: A scatterplot of WAR vs Runs

Figure 2 displays a scatterplot of WAR and R, which is an interesting relationship to visualize since it was the highest correlation among measured variables. Although WAR and R had the highest correlation, there isn't a clear visual relationship visible in the scatterplot.

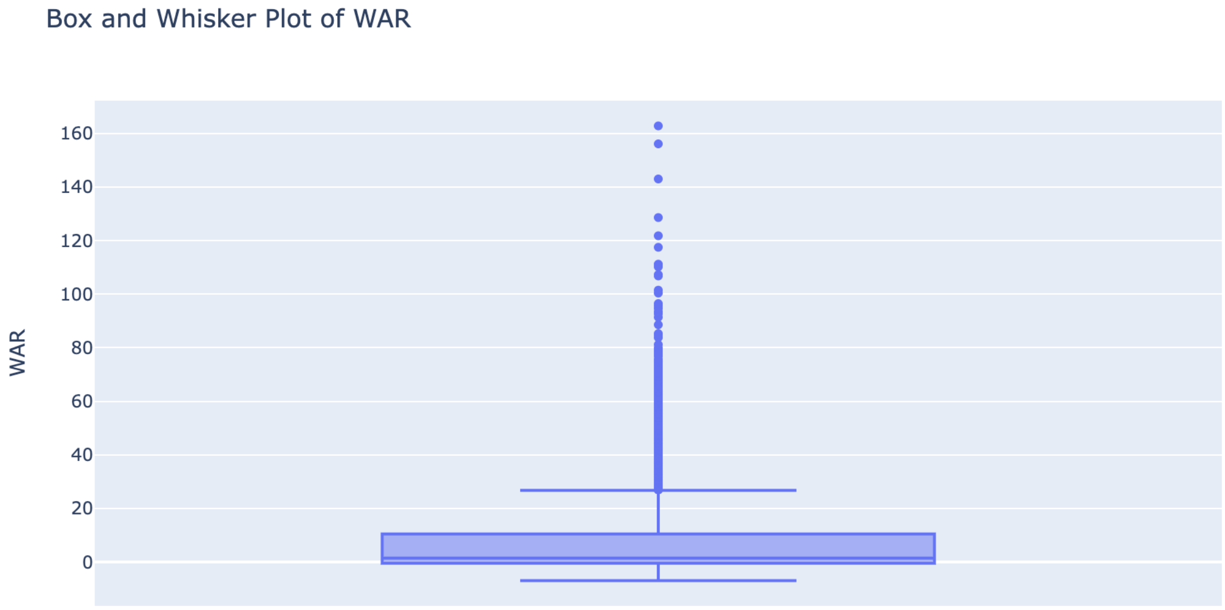


Figure 3: A box and whisker plot of WAR

Figure 3 shows the distribution of WAR in a box and whisker plot. There are many outliers on the upper end of the distribution, which we decided not to remove since they are significant. If we decided to exclude them, then we would almost never predict that anyone would reach over 26.8 WAR, which is unrealistic, as numerous players exceed that. With this in mind, we attempted to create a model that excluded outliers and found little to no improvement in said model. Since the dataset we curated is composed of reliable data, we know no measurement was falsified or mistakenly measured. However, approximately 10% of our data being outliers is still incredibly interesting.

Final Machine Learning Model

In this section, we shall explore the details of our final machine learning model.

Model Exploration

We explored various machine learning models, such as k-nearest neighbors, linear regression, a voting ensemble model, and a stacking ensemble model between the two algorithms. We decided to explore each of these models in order to account for potential complex relationships in the data. In the end, we decided our best model was a linear regression model. While the linear regression model did not yield the best RMSE, it provided simplicity. The simplicity of the linear regression model made up for the small difference in RMSE between that and our best model, the stacking ensemble model that combined our linear regression model and kNN model.

Feature Decision

In preparation for developing our model, we experimented with multiple feature sets. Our experimentation indicated that a blend of both power and contact features consistently yielded the most promising results. The final feature set we decided on was R, TB, BA, RBI, AB, G, and OPS+.

Summary of Our Final Model

The final model is a linear regression model with coefficients 0.327 (R), 0.071 (TB), -31.278 (BA), 0.024 (RBI), -0.016 (AB), -0.109 (G), 0.061 (OPS+) and an intercept of -10.679.
Our final estimated test RMSE is 14.266 WAR.

Conclusion

Our estimated test RMSE is not significantly different from the test RMSE if we predicted the mean every time (16.093 WAR). This shows us that our model was not necessarily effective in predicting player success based on their rookie season statistics. This could be caused by a couple of reasons. First, a player’s rookie season may not be the best indicator of how well they will perform in their overall career. Or, the k-nearest neighbors and linear regression algorithms might not be the best machine learning algorithms for this research question.

Future Work

There are numerous possibilities we can explore in the future. First, we plan to explore other machine learning algorithms as well, such as decision trees and neural networks to see if they yield a lower RMSE. Additionally, we could gather even more data, as our current dataset only includes players since 1940. We could also extend this to pitchers, in which a whole new set of features would need to be explored.

Annotated Bibliography

[1] "Baseball-Reference.com Players," Baseball Reference. Available: https://www.baseball-reference.com/players/. Accessed on: Nov. 10, 2023.

Baseball Reference hosts a thoroughly populated database of baseball players, offering an extensive range of statistics and historical data. This resource is important for research that delves into the analysis of baseball player performance and career trajectories, such as our project. The specific section of the website, "baseball-reference.com/players," contains links to player-specific data, encompassing a wide array of statistics that include career batting and pitching records, fielding statistics, and advanced metrics like Wins Above Replacement (WAR).