Predicting wRC+ with Baseball Savant

Perry Chenkin

Western Governors University

Contents

A. Proposal Overview

A.1 Research Question or Organizational Need

Which grouping of offensive Baseball Savant metrics is the most accurate and stable when predicting a player's wRC+?

A.2 Context and Background

The individual nature of Major League Baseball has always led to it being a more statistics-driven sport. Over the last twenty years, there has been a significant uptrend in statistics-based analytics. With advances in technology, teams can now track more metrics, also known as sabermetrics, than ever before. Metrics such as exit velocity and launch angle are used to analyze player performance and predict future performance.

Another effect of baseball's individuality is its inherent randomness. Not just pitch to pitch or game to game, but even year to year, it is tough to predict what will happen in baseball. There are also many different ways to be a good hitter in baseball; some players are driven by their barrel rates, while others are driven by their strikeout rates, and so on.

A.3 and A3A Summary of Published Works and Their Relation to the Project

Review of Work 1

In this paper, the author (Moorefield, 2021) discusses how the 2002 Oakland A's are largely responsible for the way teams employ sabermetrics and how this style of analytics began to infiltrate the business world. In 2002, Billy Beane, A's general manager, built a team that earned the second-best record in baseball that year while having the second-lowest payroll. He accomplished this by identifying players who were undervalued by traditional statistics, utilizing sabermetrics. "Sabermetrics is the statistical analysis of data in baseball which aims to quantify baseball players' performances based on objective statistical measurements..." (Moorefield, 2021). The success of that A's team led to sabermetrics being used by all teams league-wide. Technologies such as Statcast led to even more metrics. Statcast utilizes cameras in all stadiums to track new metrics, including exit velocity and launch angle.

Review of Work 2

The author of this study (Decesare, 2016) aims to investigate the correlation between winning percentage and runs scored using various metrics, including both traditional and advanced sabermetrics. They begin with a discussion about the history of statistics in baseball. They start with the first recorded box score of a baseball game, which only kept track of outs and runs, and follow all the way to modern times with more statistics than you could imagine. They then provide a breakdown of the statistics they will be using for analysis, including wRC+. For their analysis, they used multiple linear regression to determine the R value each metric has with winning percentage and runs scored. The results showed that sabermetrics generally correlated more with winning percentage and runs scored than traditional stats.

Review of Work 3

In this article for Fangraphs, the author (Slowinski, 2010) delves into the calculations for Weighted Runs Created (wRC) and Weighted Runs Created Plus (wRC+), explaining the purpose behind these statistics and how to utilize them. The original stat, Runs Created, attempts to calculate how many runs a player is worth to their team for a given season. wRC improves on Runs Created by taking the same idea, but basing it on the statistic Weighted On Base Average (wOBA). As stated in its name, wOBA assigns weights to individual stats used in the calculation to create a more accurate representation of the value each player generates. wRC+ takes a player's wRC and normalizes it against the league average. It also adjusts for the effects of the ballparks in which the player plays, since some stadiums are more challenging to hit in than others. This goes a long way to tell the true story of how well a hitter performed in the year.

A.4 Deliverables

The deliverables for the solution include dataframes, charts and a regression model. Dataframes will be created showing the different metrics calculated throughout the analysis, culminating with a dataframe that features the top groups. There will be a heatmap to show how the features correlate with one another and box plots showing the spread of wRC+ for each year. There will also be two sets of scatter plots comparing true wRC+ values to predicted ones. One set will feature the top groups, the other will feature random groups to show how much more accurate and stable the top groups are. There will also be the hyperparameters used for training the regression models, so models can be created using the top groups.

A.5 Benefits and Support of Decision-Making Process

The potential benefit of this analysis would be to give an organization a competitive edge over others. This project aims to identify which groups of metrics exhibit the best year over year stability and predictability. Using the identified groups should provide a team with better results when performing player predictions for roster construction.

B. Data Analytics Project Plan

B.1 Goals, Objectives, and Deliverables

The goal of this project is to determine which groups of Baseball Savant metrics are the most accurate and stable.

Objective 1: Collect the data and create a dataset comprising all necessary metrics and statistics.

Deliverable 1.1: A data set with each player's Baseball Savant metrics and their wRC+ for each individual year.

Deliverable 1.2: A heatmap displaying the correlation between metrics used in this project.

Deliverable 1.3: Box plots showing the spread of wRC+ for each year.

Objective 2: Train and test random forest regression models.

Deliverable 2.1: Trained regression models.

Deliverable 2.2: Three dataframes, one for weighted score, one for standard deviation, and one for confidence interval.

Deliverable 2.3: A dictionary with the results of each individual bootstrap

Objective 3: Perform paired bootstrap tests to determine statistical significance with the top performer for each year.

Deliverable 3.1: A Dataframe showing all groups that are tied with the top group in at least two of the three test years.

Objective 4: Calculate the stability score for each group.

Deliverable 4.1: A dataframe with all groups that have a stability score better than the median.

Objective 5: Determine which groups appear in both datasets.

Deliverable 5.1: A dataframe showing the top groups.

Objective 6: Create scatterplots to show comparisons between the top groups and the rest.

Deliverable 6.1: Two sets of scatterplots

B.2 Scope of Project

The scope of this project includes:

Collecting data and creating a dataframe for analysis.

Training Random Forest regression models.

Comparing models and choosing the best performers.

Determine the most predictive and consistent groups.

Create charts to visualize model metrics.

Outside the scope of this project:

Incorporating other metrics into the chosen group.

Use the model to predict a player's wRC+.

Seeing how these metrics relate to wins and losses.

B.3 Standard Methodology

The methodology that will be used on this project is the Cross-Industry Standard Process for Data Mining, CRISP-DM.

Business Understanding: This phase is when the business problem is developed. This project will attempt to find which grouping of Baseball Savant metrics is best at predicting a player's wRC+.

Data Understanding: In this phase, data collection and exploration take place. Use this phase to examine the features and determine which will be used for the analysis.

Data Preparation: This data requires minimal cleaning. Some columns need to be reformatted. Unnecessary columns will be dropped, and the two datasets will be merged.

Modeling: This phase involves training Random Forest Regression models on each distinct group.

Evaluation: The models will be ranked by weighted score, standard deviation, and confidence interval, with a composite ranking being calculated for final evaluation.

B.4 Timeline and Milestones

Milestone or deliverable	Duration	Projected start date	Anticipated end date
Collect Data and create dataframe	1 day	12/10/2025	12/11/2025
Train and test random forest models	1 day	12/11/2025	12/12/2025

Paired bootstrap tests	3 hours	12/12/2025	12/12/2025
Calculate stability scores	1 hour	12/12/2025	12/12/2025
Determine top groups	1 hour	12/12/2025	12/12/2025
Create scatterplots	1 hour	12/12/2025	12/12/2025

B.5 Resources and Costs

All tools and resources used are available free of charge.

1. Python
2. Jupyter
3. Pybaseball
4. Baseball Savant

B.6 Criteria for Success

For a group to be viewed as a stable predictor, it must be statistically tied to the top score in at least two out of three years and have a stability score below the median. A successful project will result in a dataframe listing all groups that meet both criteria.

C. Design of Data Analytics Solution

C.1 Hypothesis

There will be at least two subsets of the whole group of metrics that meet both criteria: tying the top score in at least two-thirds of the test years and having a stability score below the median.

C.2 and C.2.A Analytical Method

The model chosen for this project is a Random Forest Regression model. A different model will be trained on each group of metrics from the 2019 to 2022 seasons, and individually tested on each year from 2023 to 2025.

Random Forest models are generally highly accurate and easily interpretable. In Major League Baseball, a player's performance is influenced by numerous factors, including the opposing pitching and weather conditions. These factors result in player metrics generally not having linear relationships with performance; however, Random Forests handle non-linear relationships well. Random Forests also handle outliers well, which can appear fairly regularly in baseball. Since Random Forests are bootstrap-based models, they are naturally compatible with the bootstrapping done in this project.

C.3 Tools and Environments

The project was developed using Python and its various libraries, including pandas, matplotlib, and scikit-learn. The library pybaseball is used to collect player wRC+ data, and the library requests is used to scrape the web for Baseball Savant metrics. The project was completed entirely in Jupyter Notebook.

C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance

This project will use various metrics to evaluate statistical significance. During model testing, a weighted accuracy score for each model will be calculated using MAE, R-squared, and RMSE.

The score will be calculated for each bootstrap, and the standard deviation and confidence interval will be found for each metric group. The thirty scores found during bootstrapping will be used for paired bootstrap testing. For each year, every group will be tested against the group with the best accuracy score. The scores for each bootstrap will be compared, and the confidence interval of the differences will be calculated. If the confidence interval contains zero, then the group is statistically tied with the top group. Using the accuracy score, standard deviation, and confidence interval for each group, a weighted stability score will be calculated. For a group to be deemed significant, it must be tied with the top group in at least two-thirds of the years and have a stability score above the median.

As stated earlier, Random Forests are accurate and easily interpretable, and can handle data with non-linear relationships and outliers. For the accuracy metric, the highest weight was assigned to MAE, as it is a measure of the model's actual accuracy. The next highest weight is the R^2 score, since that details how much of a player's wRC+ that metric group accounts for. Finally, since it is more prone to being impacted by outliers, RMSE is given the smallest weight, essentially serving as a tiebreaker. For the stability score, the highest weight was given to standard deviation, as it is a measure of how spread out the data is. A lower standard deviation among the bootstrap scores indicates a more stable model. The next weight was given to the accuracy score because it doesn't matter how stable a model is if it isn't accurate. The confidence interval, another measure of spread, was given the lowest weight.

C.5 Practical Significance

The practical significance of these findings is the impact they could have on a team's analysis. Teams centering their predictive analysis around the top groups of metrics determined by this analysis should see better results.

C.6 Visual Communication

When the data is collected, there will be a heatmap to show the correlation between each metric. There will also be a box plot for each year to demonstrate the spread of the wRC+ for each year. Finally, there will be two sets of scatterplots showing the actual and predicted wRC+ values for the test years. One set will be of the top groups, the other will be a random sample of all group combinations. The purpose of these sets of charts is to visualize how much more accurate and consistent the top models are.

D. Description of Dataset

D.1 Source of Data

The data was acquired from two sources. The metrics were obtained from Baseball Savant, and the players' wRC+ values were sourced from Fangraphs.

D.2 Appropriateness of Dataset

This project aims to find the models that best predict a player's offensive output. This data is appropriate because one use of Baseball Savant metrics is to predict a player's offensive output, and wRC+ from Fangraphs is a measure of a player's overall offensive performance.

D.3 Data Collection Methods

The first dataset was downloaded from www.baseballsavant.mlb.com using the Python requests library. The Fangraphs data was acquired using the Python library pybaseball.

D.4 Observations on Quality and Completeness of Data

Both datasets were already clean and complete. The only issue was that the name column in each dataset was formatted differently, which was one of the columns required for merging. The Baseball Savant data had the player's last name first, so it was reformatted to list the first name first.

D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances

Data Governance - Baseball Savant metrics are taken from the data captured by cameras and radars at all MLB stadiums. Fangraphs is one of the most respected companies in baseball statistics.

Privacy - No personal identifiable information is used other than publicly available player names.

Security - There are minimal security risks since this data is all free and publicly available.

Ethical, legal, and regulatory compliance considerations - All data used is collected and released with the consent of Major League Baseball.

References

Decesare, V. (2016). *USING CONVENTIONAL AND SABERMETRIC BASEBALL STATISTICS FOR PREDICTING MAJOR LEAGUE BASEBALL WIN PERCENTAGE*. The Pennsylvania State University. https://honors.libraries.psu.edu/files/final_submissions/3354

Moorefield, J. (2021). *The Oakland Athletics use of sabermetrics and the rise of big data analytics in business*. University of Tennessee at Chattanooga. https://scholar.utc.edu/cgi/viewcontent.cgi?article=1319&context=honors-theses

Slowinski, P. (2010). *wRC and wRC+*. Fangraphs. https://library.fangraphs.com/offense/wrc/

Baseball savant: Statcast, trending MLB players and visualizations. baseballsavant.com. (n.d.). https://baseballsavant.mlb.com/

Fangraphs baseball. (n.d.). https://www.fangraphs.com/