Predicting wRC+ with Baseball Savant

Perry Chenkin

Western Governors University

Table of Contents

A. Project Highlights

Research Question:

The research question addressed by this project was "Which grouping of offensive Baseball Savant metrics is the most accurate and stable when predicting a player's wRC+?"

Project Scope:

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.

Overview of Solution:

The solution involved training multiple Random Forrest Regression models and determining which are the best-performing models. The tools used for this project were Python and Jupyter Notebooks. Various Python packages and libraries were used. Pandas and Numpy were used for data collection and analysis. Matplotlib and Seaborn were used to create charts. Scikit-learn was used for training and testing the Random Forest models.

B. Project Execution

The proposed plan was to identify the best-performing metric groups by training Random Forest regression models.

Since this project aimed to collect data, create, and evaluate models, the CRISP-DM methodology was the most suitable approach for this project.

The project execution did not differ from the initial plan. All the objectives outlined in the proposal were met, and all the deliverables were successfully produced. The CRISP-DM methodology was chosen for this project and followed. First, the business problem was identified. Next, the data was explored and understood, then prepared for modeling. Then, each model was trained and subsequently evaluated. The project was completed within the proposed timeline, with milestones meeting their deadlines.

C. Data Collection Process

The data selection and collection did not differ from the original plan. The Baseball Savant metrics were acquired using the requests library in Python, and pybaseball was used to download Fangraphs data. No obstacles were encountered during the data collection process. Since the data is entirely free and publicly available, there were no governance issues.

C.1 Advantages and Limitations of Data Set

Advantages:

- Since the Baseball Savant metrics are given in percentiles, and wRC+ is normalized to the league average, the data doesn't need to be normalized.
- The data is easily separable by year.
- All the data is entirely free and publicly available.

Limitations

- Using more years can utilize a significant amount of computing power.
- Baseball Savant metrics only date back to 2015, and some of the metrics have only become available in the last five years.

D. Data Extraction and Preparation

Python and Jupyter Notebook were used for both the data extraction and preparation processes. The data was extracted with two different Python libraries. The requests library was used for the Baseball Savant data, and pybaseball was used for Fangraphs. This was appropriate as it allows easy changes to data collection within the code. Various Python methods were used to prepare the data. The name columns were one of the columns to be used for merging, so the names in the Savant dataframe needed to be reformatted to match those in the wRC+ dataframe. Then, only the required columns were selected, and the two dataframes were merged into a single one.

E. Data Analysis Process

E.1 Data Analysis Methods

The first step of the analysis was training random forest regression models. A model was trained for each different combination of Baseball Savant metrics. The models were trained on the data from the 2015 to 2022 seasons, excluding 2020, and tested individually on each season from 2023 to 2025. For each test year, 50 bootstraps were performed with replacement to create 50 different test datasets. For each bootstrap, a weighted accuracy score was calculated. Following the completion of the bootstrap for each year, the following metrics were calculated: the mean accuracy score, standard deviation, and confidence interval. These metrics were saved as a dictionary within a dictionary, under the key corresponding to the combination of Baseball Savant metrics for which they are intended.

The next step was to determine the group with the best accuracy score for each year and identify which groups are not statistically significantly different from the best, as determined using a paired bootstrap test. The difference was calculated between the individual bootstrap scores of the top group and each of the other groups, respectively. For each group, the mean difference and confidence interval were calculated. If the confidence interval contained 0, then the two groups were not significantly different from each other. Groups that were either the top group or tied with the top in at least two of the three years were selected.

The final step before analysis was to calculate a stability score for each group. This score was calculated using the standard deviation, confidence interval, and the mean accuracy score. The score for each group was compared to the median, and if the group's score was better than the median, it was selected.

E.2 Advantages and Limitations of Tools and Techniques

Random Forest:

Advantages:

- Highly accurate
- Handles data with outliers and non-linear relationships
- Lower risk of overfitting
- Can handle large amounts of data

Limitations:

- Can take a long time and computing resources to train
- While individual trees are easy to interpret, a Random Forest can be more complex

Paired bootstrap test:

Advantages:

- Controls for variability since both samples use the same dataset
- Works well with smaller datasets

Limitations:

- Requires data to fit three assumptions: normal distribution, matched pairs, continuous data

E.3 Application of Analytical Methods

Random Forest Regression:

1. Split data into training (2015-2022) and testing (2023-2025) sets
2. Train the random forest model
3. Perform 50 tests on each testing year using bootstraps with replacement, calculating a weighted score for each bootstrap

4. Calculate the mean score, standard deviation, and confidence interval for each test year, and store the results
5. Repeat the process, looping through each metric group

Paired Bootstrap Test:

1. Determine the group with the best mean weighted score
2. Compare bootstrap scores against the top group
3. Calculate the mean and confidence intervals to determine if the differences are significant or not, and store the results

Validation:

1. Histograms are used to demonstrate that the data is generally normally distributed
2. The data used in the weighted score, which is continuous
3. The data is used from the same dataset to predict the same thing

F Data Analysis Results

F.1 Statistical Significance

The models were first evaluated using three metrics: mean absolute error, r^2, and root mean square error. These three were combined into a single weighted score using the formula (0.6 * MAE) - (0.3 * (R^2* 10)) + (0.1 * RMSE). MAE is given the highest weight since it represents the actual average error of the model. R^2 is next, as it represents the proportion of change in wRC+ that each metric group is responsible for. Finally, RMSE is used as essentially a tiebreaker. This score is calculated for each bootstrap for each year. A paired test is then used to determine which groups are not significantly different from the group with the best score in each year. Groups that appear as the top or tied for the top at least two times over the three years will be considered for the top groups, and five groups met this requirement.

Next, a stability score was calculated using the total standard deviation and confidence interval for each group across all three test years, with the formula (0.7 * (STD/Max STD)) + (0.3 * (CI/Max CI)). To be one of the top groups, the group must have a stability score above the median. When this was used to filter the top groups, only four remained.

F.2 Practical Significance

This project aimed to find which groupings of Baseball Savant metrics are the most stable and accurate when predicting a player's wRC+. Prioritizing the use of the metrics identified by this project could give teams an edge when performing player transactions. They could use it to identify players due for positive or negative regression in ways that other teams won't.

F.3 Overall Success

Based on the criteria in the project proposal, this project was a success. The criteria for success were defined as the output of a dataframe with all groups that met both requirements, which was produced. There was also the hypothesis stating that there would be at least two subsets of the total group of metrics, and the final results produced four different subsets.

G. Conclusion

G.1 Summary of Conclusions

The final four groups were (brl_percent, k_percent, bb_percent, chase_percent), (brl_percent, k_percent, bb_percent), (brl_percent, exit_velocity, k_percent, bb_percent), (brl_percent, exit_velocity, k_percent, bb_percent, whiff_percent). The first thing that stands out is that all four groups contain the same three metrics (brl_percent, k_percent, bb_percent), which is also the only group of less than four metrics in the top. In baseball, the three true outcomes are home runs, walks, and strikeouts because these are the three outcomes of an at-bat that the defense has no impact on. These three metrics(brl_percent, k_percent, bb_percent) could be argued to be the three true outcomes of Baseball Savant. These metrics represent the frequency with which a player walks, strikes out, or hits the ball hard, all of which are at-bat outcomes that the defense has no impact on. The brl_percent metric is more in-depth than other batted ball metrics.

The other standout group was (brl_percent, k_percent, bb_percent, chase_percent), as it was the only group to appear at the top or tied for the top in all three years. This group takes the three metrics and adds chase_percent, which represents the frequency at which a batter swings at pitches outside the strike zone. The combination of these four metrics paints a good picture of how well a player understands and controls the strike zone.

G.2 Effective Storytelling

Different charts and tables are used to help with effective storytelling. First, a heatmap is used to illustrate the correlation between all the metrics used. Individual box plots were generated to demonstrate the spread in wRC+ for each season used in the project. Histograms are used to display the distribution of scores generated during bootstrapping. 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.

G.3 Recommended Courses of Action

Test how predictive changes in these metrics are: Rather than just looking at how predictive each metric is on its own, they could be tested for how predictive they are against themselves. Perform the same experiment, but instead of using a player's metrics, see how well their year-over-year changes predict changes in wRC+. Then compare the results to the original models and determine which one is a better predictor.

Add more statistics and metrics: While this is a good start, incorporating additional statistics and metrics could yield even better models. Run a similar experiment, using the top models with more metrics from sources such as Baseball Savant and Fangraphs.