Predicting Batting Performance with Ensemble Neural Nets and Data Augmentation in Baseball Simulator

Hubert Luo [1]

Abstract

Developing a metric to evaluate batting performance, identifying the drivers of increased offensive ability, and using ensemble neural net architecture with data augmentation to improve prediction of batting performance for 20,868 seasons in a baseball simulator.

1. Introduction

BrokenBat is an online baseball simulator consisting of 756 total teams. There are more than 600 active users who each manage a baseball team consisting of fictional, computer-generated players. These fictional teams then compete against one another in games simulated on a pitch-by-pitch basis.

This paper aims to develop a metric to evaluate batting performance, identify the drivers of increased offensive ability, and improve batting models using an ensemble neural network with data augmentation approach.

2. Previous Work

This paper takes inspiration from not only the prevalence of sabermetrics in professional baseball, but also a wide variety of batting projection systems in MLB such as PECOTA and ZiPS. In real-life baseball, it is possible to obtain much more granular data than is available in a baseball simulator like BrokenBat.

For example, in BrokenBat, there is no way to determine something as straightforward as a batter's average exit velocity or hard-hit rate. Therefore, the applicability of this research done for MLB is somewhat limited when working with only the more-elementary data that is available for BrokenBat.

This paper also builds on previous work done to predict wOBA for BrokenBat players by introducing historical player performance and creating a different metric that adjusts for level of competition and ballpark. (Luo, 2018)

3. Approach

3.1. Data Scraping

In BrokenBat, there are six league levels with teams promoting up or down this pyramid of leagues depending on their team performance. The initial universe of players scraped for this paper is all position players on the roster of a team at League Level 4 or higher (LL1-4) as of the start of the 2047 season.

Scraping was performed with permissions from game administrators prior to commencing and was structured to minimize the impact on server load. A total of 3,641 unique players were scraped, in total encompassing 20,868 different seasons.

3.2. Data Description

In BrokenBat, ever player has demographic information such as age and salary. Experience was derived by taking the number of seasons each player has had with at least 300 plate appearances (PA).

Each batter has a collection of batting attributes, ranging from 0 to 20 inclusive. The game manual (Muller, 2020) lists them as such:

- Hitting: ability to hit the ball and put it in play.
- Bat Control: ability to make contact with the baseball, especially important for avoiding strike outs and successful bunting.
- Plate Discipline: ability to discern strikes from ball. Batters with good plate discipline will walk more and make the pitcher throw more pitches.
- Power: ability to drive the ball, albeit not necessarily in a high trajectory.
- Speed: ability to run fast around the bases.

The median player in the training dataset was age 28, had 4 years of experience, and earned a salary of $1.6 million.

[*]Equal contribution [1]Department of Statistics, University of California, Berkeley.

The mean batting attributes for the players were 15.8 hitting, 13.6 plate discipline, 13.4 bat control, and 13.5 power. An example player card is provided in Figure 1, and the empirical CDF of these features is provided in Figure 2.

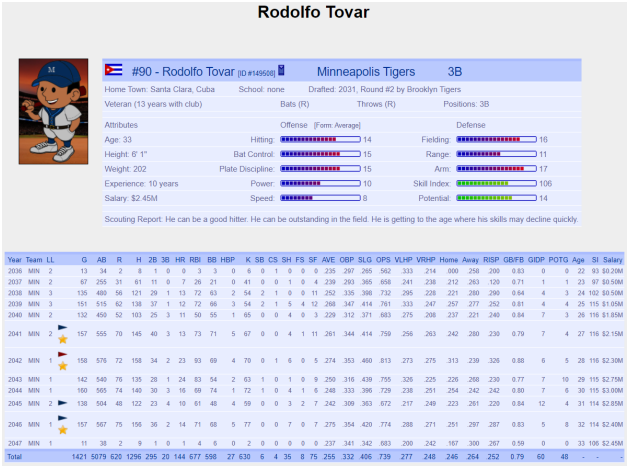

Figure 1. Example player card with batting attributes and other statistics.

Previous work for MLB found that wOBA was observed to reach a Cronbach's alpha of about 0.57 at 400 PA. (Pemstein & Dolinar, 2015) Therefore, in order to reduce the amount of random variance in the dataset while still maintaining a large-enough sample of players to evaluate, a cutoff of 400 PA was used, resulting in 6,040 seasons for the training dataset after splitting.

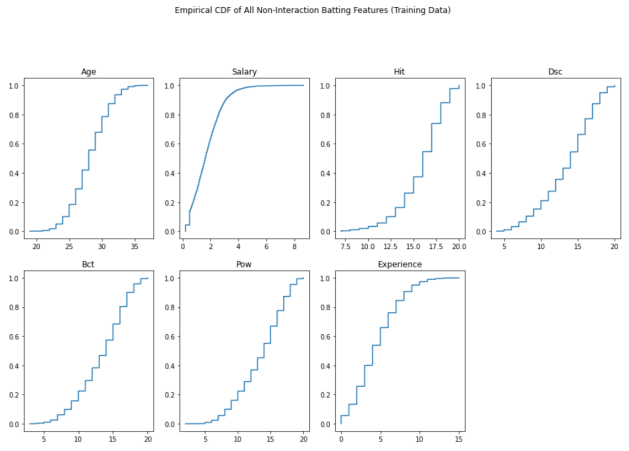

Figure 2. Empirical CDF of batting features in training dataset.

3.3. Developing a Batting Performance Metric

This paper takes inspiration primarily from the batting component of the Wins-Above-Replacement (WAR) metric for MLB, which encompasses batting, base running, and fielding adjusted for player position and league. In MLB, the batting component is calculated using a weighted Runs Above Average (wRAA) adjusted for park and league. The formula for wRAA in MLB is based primarily on weighted on-base average (wOBA) with some scaling for league average.

Table 1. BrokenBat League Level Factors to convert wOBA from previous LL to LL-1.

LL	PREVIOUS LL	LLF
1	1	1.0
1	2	0.979
1	3	0.971
1	4	0.957
1	5	0.935
1	6	0.894

wOBA is a weighted linear sum of a player's offensive contribution, placing greater weight on contributions that lead to higher run expectancy - a home run is much more valuable than a single for example. For BrokenBat, wOBA can be directly calculated for each player without modification and will form the basis of the batting performance metric for BrokenBat, Batting Runs (BR), which will also be adjusted for league level and ballpark.

There is patently a stark difference in league levels - a player who does well in LL-6 will almost certainly see a steep drop in performance in LL-1 as the difference in the quality of competition is simply drastic. In order to calculate the batting component of WAR, wOBA needs to account for the league level.

Therefore, League Level Factor (LLF) was created as the factor by which a player's wOBA should be scaled so it would represent that player's wOBA if they were to play in LL-1. The higher the LLF, the closer that league's level of play to LL-1.

To determine appropriate LLF, a player's performance at pairs of different league levels was compared as long as they had a sufficiently large sample of at bats at each league level and a comparable Skill Index (SI). Specifically, the criteria used was at least 200 AB in a season and a change in SI of five or fewer. LLF coefficients are displayed in Table 1.

A home ballpark can also play a drastic role in a player's offensive performance given the high number of at-bats a batter may have in an environment that is either conducive to generating runs or run-suppressing.

Therefore, a Park Factor (PF) was created to represent how friendly a team's home ballpark is for run-scoring. It is the average runs scored in the home ballpark by both teams divided by the average runs scored in that team's road games

by both teams. The higher the PF, the more run-friendly that ballpark is.

The distribution of all park factors is shown in Figure 3. It is roughly Gaussian with a mean at 1 and SD of 0.12. Park factors ranged from 0.63 (Saginaw in 2043) being the most pitcher-friendly ballpark to 1.49 (Arlington in 2038) being the most batter-friendly ballpark.

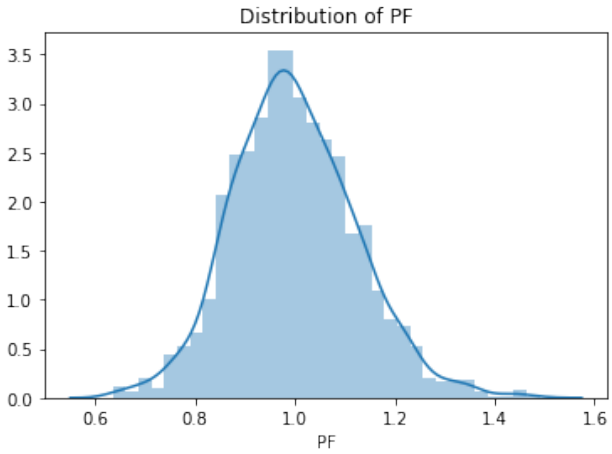

Figure 3. Distribution of all Park Factors.

A scaled wOBA was then derived from the wOBA to account for league level and home ballpark. The unscaled wOBA was first multiplied by LLF to get the projected wOBA if they were to play in LL-1. Due to the wide range in PF, a scaled PF was used to penalize the PF and prevent it from having too significant of a result on the findings. The higher the scaling constant α the more the PF is penalized in the calculation of the scaled wOBA. A scaling constant $\alpha = 0.75$ was empirically found to have reasonable effects on the scaled wOBA.

$$Scaled\, wOBA = \frac{wOBA \cdot LLF}{\alpha \cdot (1 - PF) + PF}$$

A Batting Runs (BR) metric was then derived by first standardizing wOBA to have mean 0 and standard deviation 1. This standardized metric was then centred at 50 and scaled to a range of roughly 100 to create the BR metric and make it more intuitive when comparing two BR values.

$$BR = 50 + 15 \cdot \frac{Scaled\, wOBA - 0.3147}{0.0367}$$

Figure 4 shows the distribution of BR in the training dataset. The distribution is roughly Gaussian with a mean of 50 and standard deviation of 15. The highest BR belonged to Jimmy Bryant, who posted a 102.34 BR in 2045, and the lowest BR belonged to Van Martini, who posted a -13.10 BR in 2043.

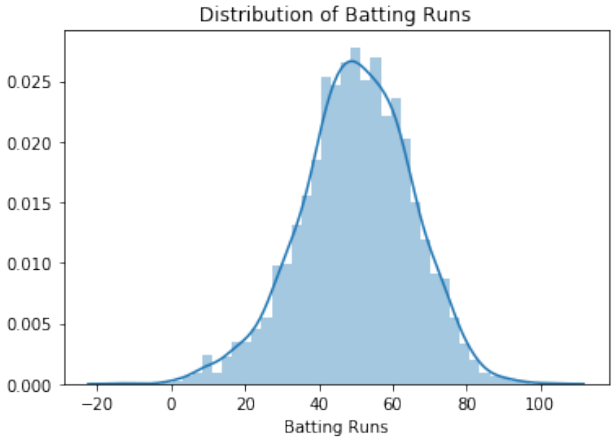

Figure 4. Distribution of BR values in training dataset.

3.4. Baseline Model

A baseline model was first trained, using a player's BR in their previous season (BR-1) as the baseline prediction for the BR in the current season. This resulted in a training RMSE of 13.65 and a validation RMSE of 13.36.

3.5. Neural Net Model

The training data was first scaled using the median to reduce the effect of outliers. Cross-feature interaction terms were then created, and finally highly correlated features were removed. The optimal neural net model was found to have the following architecture:

- Fully connected layer with ReLU activation and output size of 32
  - Batchnorm layer
  - Dropout layer with dropout rate of 0.1
- Fully connected layer with ReLU activation and output size of 16
  - Batchnorm layer
  - Dropout layer with dropout rate of 0.1
- Fully connected layer with output size 1

The resulting model was then trained with a batch size of 16 for 20 epochs.

3.6. Neural Net Model with Data Augmentation

Data augmentation was performed using the scaled training set with selected features to increase the robustness of the models by presenting some variation to individual components of each datapoint.

A small perturbation $\varepsilon \sim N(0, 0.1^2)$ was introduced to augment the data. The training dataset was sampled 100,000 times with replacement and the perturbations were added to the value of each variable for the sampled datapoints.

3.7. Ensemble Neural Nets

An ensemble neural net was then created using five models with AdaBoost. Three models were trained on the full training set and two models were trained on specific subsets of the augmented data, either developing players under age 27 or declining players over age 30.

This was due to the fact that these subsets of players have distinctive characteristics in BrokenBat. Players under 27 are still developing, with batting attributes expected to gradually increase each year and thus yet to hit their peak. This group of developing players accounted for 29.07% of seasons in the training set.

On the other hand, players over age 30 usually already show signs of decline in their batting attributes and performance. These declining players represented 21.36% of seasons in the training set.

These cutoff points were selected not only based on previous experience with player attribute changes in the simulator, but also to ensure a relatively even split between developing and declining players.

See Section 5 for further discussion about these subgroups of players.

4. Results

4.1. Model Results

The datasets used to evaluate the models were the training and validation sets which encompassed 80% and 10% of the original data respectively. The remaining 10% of the original dataset went to a holdout testing dataset which was patently unused throughout this process.

All models outperformed the baseline model, with incremental improvements also seen after adding data augmentation. The ensemble model outperformed all other models on the basis of validation RMSE and generalized better than the model with data augmentation. A summary of the results is shown in Table 2.

On the final testing dataset, the ensemble neural net had a RMSE of 9.745.

Table 2. Summary of results for neural net models. Note that 'DA' refers to Data Augmentation.

MODEL	TRAIN RMSE	VAL RMSE
BASELINE	13.648	13.356
NN	9.692	10.109
NN WITH DA	9.121	9.966
ENSEMBLE NN	9.269	9.945

4.2. Principal Component Analysis

Using PCA on the scaled training set with selected features, it was found that five principal components were sufficient to account for 87.16% of the variation in the data, as demonstrated in Figure 5.

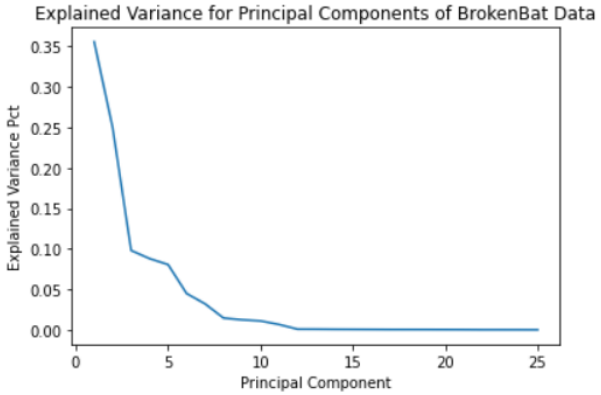

Figure 5. Proportions of explained variance in principal components of training data.

The first principal component relies heavily on previous batting performance to project future batting performance - the component has large positive weights on the BR in the previous season (BR-1) and the season prior to that (BR-2).

The second principal component is largely driven by a negative factor on the momentum in previous batting performance, i.e., an increase from BR-2 to BR-1 was found to actually lead to a correction in the opposite direction for the BR in the current season.

4.3. Drivers of Offensive Performance

Based on a random forest model using the scaled training set with selected features, interaction terms involving a player's hitting and plate discipline, in addition to a player's hitting and power, were found to be the most useful in predicting batting performance. Other factors found to be important were the player's BR-1 and BR-2.

Hitting was the most important individual non-interaction

feature. This aligns well with anecdotal historical evidence, where managers emphasize hitting the most out of the batting attributes when evaluating players.

More noteworthy is the importance of hitting and plate discipline, as most managers would instead prioritize a different secondary attribute such as power or bat control. The importance of plate discipline is under-valued among managers and represents an inefficiency in player evaluation among BrokenBat managers today. The feature importance plot of the most important features for predicting BR is shown in Figure 6.

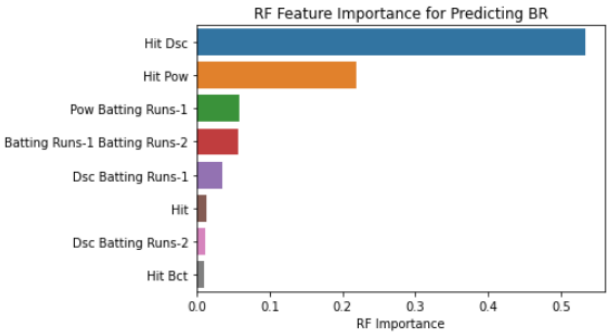

Figure 6. Feature importance of selected features from random forest model on training data.

5. Limitations

There is survivorship bias involved in the data among older player as it is unlikely that managers would give a large number of at bats to an older player who is no longer providing useful value to their team. Therefore, there is not only a relative lack of players who are in their declining phase, but there is also an over-estimation of the abilities of these older players since only the best ones still received regular at bats.

In addition, there is a limitation for younger players who are still seeing increases in their player attributes. The data for the attributes scraped into the dataset are the attributes at the beginning of the season. Therefore, by the end of that season, a player's attributes may have increased by a substantial amount, usually 1-3 points, and the value at the beginning of the season is no longer reflective of the actual value.

Therefore, given these limitations on both developing and declining players, separate models were trained for these subgroups of players, as described previously in Section 3.7, to partially accommodate the limitations described here.

6. Future Work

Minor league performance could be incorporated into the models to improve the predictions especially for younger players or those with a more limited track record at the major league level.

A different approach based on PECOTA for MLB would be the natural next step in this progression of player modelling work. Specifically, the goal would be to find comparable players to the one in question and use the career trajectories of those comparables to guide the prediction. An ensemble model of this player comparables model with the current neural net models could also lead to improved performance.

7. Acknowledgements

I would like to thank Steve Muller, the creator and administrator of BrokenBat, without whom none of this would have been possible.

References

Luo, H. Brokenbat player performance report, 2018.

Muller, S. Game manuel, 2020.

Pemstein, J. and Dolinar, S. A new way to look at sample size, 2015.