How Far Will It Go? Predicting Hit Distance for MLB Players

Ian Curtis

2022-10-19

Project Overview

Overall, this project aims to predict the distance a batted ball will travel when hit from a professional MLB player. This project falls into the Sports Analytics domain and joins a select number of projects that have been performed publicly on MLB data. All MLB teams have a private data analytics team devoted to running analyses and making predictions for that team; however, these analyses and results are not made public in an attempt to give the respective team an advantage.

One of the great aspects of baseball is the amount of random variables at play. One of these variables is how far a batter will hit a ball. With a model that can predict how far a ball will travel, teams may be better able to predict where to place their position players and how to prepare these players for a ball. For instance, if a player is more likely to hit a ball in the outfield than the infield, a team's outfielders can be more prepared for a fly ball.

I aim to split the baseball field into different distance segments and predict into which of these areas a ball will be hit. I plan to create three different models: one of multiple linear regression, one of a neural network, and one of a decision tree. The neural network and decision tree will predict the category automatically while the regression model will predict an exact hit distance and categorization will occur based on that number.

My main goal for this project is to complete it entirely using R. I also wish to create the regression model myself from scratch. I hope to determine an accurate way to compare these models, visualize results and initial variable distributions, communicate recommendations effectively for future work in the area, and to determine if this project will agree with previous studies that neural networks tend to outperform regression models (see below).

Related Work

Various other projects have been attempted in baseball analytics, usually with the goal of predicting pitch type or the outcome of a game. Huang and Li (2021) use various neural networks and Support Vector Machines (SVM) to predict wins and losses of MLB games. Cserepy et al. also attempted to predict game outcomes but through simulations rather than an hard algorithm. Lee (2022) uses neural networks but for predicting a pitch type and where the pitch will land in the strike zone. Everman (date unknown) uses a handfull of popular statistics to predict overall team performance and includes a custom metric that will be considered for use in this project. Hung (2012) decided to use Principal Component Analysis to reduce data dimensions followed by a k-means clustering algorithm in order to evalute certain batter metrics and skill set. Young et al. take a unique approach by using a neural network to predict which players might be inducted into the Hall of Fame. There appears to be support for using neural networks in baseball analysis as well as classifiers. Regression seems to be a fairly traditional approach to sports analysis and both Maszczyk et al. (javelin throws) and Karnuta et al. (baseball injuries) demonstrate the superiority of neural networks in this field.

Data Plan

The data for this project could come from multiple sources but I am choosing to start with the MLB Baseball Savant Database. If there is time, I may also include the Lahman Baseball Database or from the Retrosheet collection. The R package `baseballr` provides a way to connect to the Savant and Retrosheet websites and the `lahman` package imports the Lahman data.

The Savant website is maintained by MLB analysts for both casual and serious researchers to analyze MLB data. The Lahman database was created originally by Sean Lahamn "to create a [public] repository for baseball stats and historical information" and to increase access to MLB data (Lahman). The dataset and R package are now maintained by a small group of dedicated people and include summarized data per player by season. Retrosheet is a service that provides play-by-play and box score data almost as far back as baseball has existed. The challenge is that the data is in a format that requires a special step of extraction.

As the data is already available, I do not need to collect data myself. However, I do expect to have to preprocess data which will likely involve grabbing only 2022 data (for now), removing columns that are not needed in the analysis, standardizing data as to not bias results, and dealing with possible missing values. Part of the data acquisition will also involve filtering by pitches that ended in a hit or a putout (e.g., no swings and misses or foul balls).

Once I have a better idea of the data I am working with, I will decide on the category splits for hitting distance, defining regions of the baseball field where a ball could land and assigning each observation its respective category. (If time, I could build in a feature that determines whether a certain hit would be a home run in a certain ballpark.)

Implementation Plan

In order to make this project happen, I first plan to connect R to the Savant database, running a few test queries to make sure everything works. I then will look at the attributes provided and decide which variables I believe influence the distance of a batted ball and create a data frame of these variables. I will begin by analyzing the data anonymously; that is, without an associated player name. If time permits, I will build in a way to predict hit distance *per player* (or perhaps grouped by team).

Once the data has been created, I will run some exploratory data analysis (EDA) to identify any outliers, strange observations, and skewed data. Since regression requires linear data, some data transformation may be required as may dimensionality reduction. Once the data is ready to be analyzed, I will split the data into a train and test dataset, then I will build the algorithms.

The major piece of this project will be to build the linear regression model myself. I have taken a course in multivariate statistics and learned basic information on how a regression analysis is done with matrices and I plan to implement such matrix algebra myself including any necessary summary statistics needed to evaluate the regression model (such as R^2, the MSE, and the estimate for the regression line). I then plan to use the R `nnet` package (recommended by Mahdi et al.) to construct a neural network and the `rpart` package to build a decision tree to predict the region of the field a ball is predicted to land.

Evaluation Plan

To evaluate these algorithms, I plan to use the testing dataset as split in the beginning of the pipeline. Each algorithm will have used the same training set and will also be evaluated using the same testing set. I will create a confusion matrix determining which hits the algorithm incorrectly categorized. I then will assess the overall accuracy using the formula used in Huang and Li (2021):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

where TN = true negative, TP = true positive, FN = false negative, and FP = false positive. At the very least, I hope to create a model card (Mitchell et al.) for the regression model as that is the one I plan to implement myself but I would like to also create one for the decision tree and neural network.

Plan for Group Collaboration

Not applicable, although I will be managing my code and project files on GitHub.

Timeline

Week Beginning	Goal(s)	Notes
10/17	* Write proposal * Turn in proposal * Work on messing with importing packages and data	
10/24	* Run test queries * Have data imported, variables chosen, and data split * Begin working on preprocessing * Determine how to break up field into categories	* Proposal due 10/26 * Fall Break
10/31	* Incorporate proposal feedback * Have data cleaned and preprocessed	Receive proposal feedback
11/7	* Start exploring matrix algebra and regression techniques * Start working on progress report	
11/14	* Finish regression algorithm * Create necessary summary statistics to adjust model * Outline report draft	Progress report due 11/16
11/21	* Use packages to make neural network and decision tree * Start report draft	Thanksgiving Break
11/28	* Run model evaluation by testing algorithms on test data * Finish neural network and decision tree if not done * Finish report draft	Report draft due 11/30
12/5	* Incorporate draft feedback * Finish model evaluation and model card if not quite done, add extra features if time	Draft feedback
12/12	* Finish final report * Wrap up any loose ends, add extra features if time	Report due 12/14

References

"Baseball Savant: Trending MLB Players, Statcast and Visualizations." *Baseballsavant.com*, https://baseballsavant.mlb.com/. Accessed 17 Oct. 2022.

Cserepy, Nico, et al. "Predicting the Final Score of Major League Baseball Games." *Economics*, 2015, https://cs229.stanford.edu/proj2015/113_report.pdf.

Everman, Brad. *Analyzing Baseball Statistics Using Data Mining*. [date unknown]. https://truculent.org/papers/DB%20Paper.pdf

Friendly, Michael, et al. *Lahman: Sean "Lahman" Baseball Database*. 10.0-1, 26 Apr. 2022. *R-Packages*, https://CRAN.R-project.org/package=Lahman.

Huang, Mei-Ling, and Yun-Zhi Li. "Use of Machine Learning and Deep Learning to Predict

the Outcomes of Major League Baseball Matches." *Applied Sciences*, vol. 11, no. 10, 10, Jan. 2021, p. 4499. *www.mdpi.com*, https://doi.org/10.3390/app11104499.

Karnuta, Jaret M., et al. "Machine Learning Outperforms Regression Analysis to Predict Next-Season Major League Baseball Player Injuries: Epidemiology and Validation of 13,982 Player-Years From Performance and Injury Profile Trends, 2000-2017." *Orthopaedic Journal of Sports Medicine*, vol. 8, no. 11, Nov. 2020. *SAGE Journals*, https://doi.org/10.1177/2325967120963046.

Koseler, Kaan, and Matthew Stephan. "Machine Learning Applications in Baseball: A Systematic Literature Review." *Applied Artificial Intelligence*, vol. 31, no. 9–10, Nov. 2017, pp. 745–63. *Taylor and Francis+NEJM*, https://doi.org/10.1080/08839514.2018.1442991.

Lahman, Sean. "Baseball Archive." *SeanLahman.Com*, 29 Dec. 2010, https://www.seanlahman.com/baseball-archive/.

Lee, Jae Sik. "Prediction of Pitch Type and Location in Baseball Using Ensemble Model of Deep Neural Networks." *Journal of Sports Analytics*, vol. 8, no. 2, Jan. 2022, pp. 115–26. *content.iospress.com*, https://doi.org/10.3233/JSA-200559.

Mahdi, Salsabila, et al. A Review of R Neural Network Packages (with NNbenchmark): Accuracy and Ease of Use. *The R Journal*. https://www.inmodelia.com/exemples/2021-0103-RJournal-SM-AV-CD-PK-JN.pdf.

Maszczyk, Adam, et al. "Application of Neural and Regression Models in Sports Results Prediction." *Procedia - Social and Behavioral Sciences*, vol. 117, Mar. 2014, pp. 482–87. *ScienceDirect*, https://doi.org/10.1016/j.sbspro.2014.02.249.

Mitchell, Margaret, et al. "Model Cards for Model Reporting." *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, 2019, pp. 220–29. https://doi.org/10.1145/3287560.3287596.

Petti, Bill, et al. *Baseballr: Acquiring and Analyzing Baseball Data*. 1.3.0, 9 Sept. 2022. *R-Packages*, https://CRAN.R-project.org/package=baseballr.

Retrosheet. https://www.retrosheet.org/. Accessed 19 Oct. 2022.

Ripley, Brian, and William Venables. *Nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*. 7.3-18, 28 Sept. 2022. *R-Packages*, https://CRAN.R-project.org/package=nnet.

Therneau, Terry, et al. *Rpart: Recursive Partitioning and Regression Trees*. 4.1.16, 24 Jan. 2022. *R-Packages*, https://CRAN.R-project.org/package=rpart.

Tung, David D. Data Mining Career Batting Performances in Baseball (Preprint). *Journal of Data Science*. 2012. https://vixra.org/pdf/1205.0104v1.pdf.

Young, William A., et al. "Determining Hall of Fame Status for Major League Baseball Using an Artificial Neural Network." *Journal of Quantitative Analysis in Sports*, vol. 4, no. 4, Oct. 2008. *www.degruyter.com*, https://doi.org/10.2202/1559-0410.1131.