Evaluating Run Scoring Efficiency in MLB Through Markov Chain and Machine Learning Processes

Brett Gustin
Department of Sport Analytics, Falk College of Sport, Syracuse University

Syracuse University
Falk College of Sport & Human Dynamics

Abstract

Baseball's unique aspect of a 162-game season is part of what makes the sport so special. The durability needed for a Major League Baseball year is hard to compare to any other sport. The length of the regular season leads to believe that the best teams are the ones that can come out on top in the standings after an enduring battle all year. However, once in the playoffs in baseball, anything can happen. In the past nine years, five wild card teams have made it to the World Series with three of them being crowned champions. If it is all about making it to the dance, then the goal is to perform with consistency in the regular season. Doing so over the long course of a season will set you up as a team for playoff standing. The goal of this research work is to quantify how teams set themselves up for playoff contention in terms of scoring runs efficiently. Through various modelling processes including Markov Chain simulation, understanding how previous teams succeed through consistent run production allows for accurate prediction for teams to come.

Introduction

In order to implement a simulation process for run efficiency, proper data collection of batting outcome events is needed. By turning the game into transition states, every moment in a baseball game is at a current state and has a probability to move to a succeeding state. Moving through transition states is defined through a probability of how frequent that outcome or move to another part of the game can occur. The probabilities necessary to define this movement can be obtained through batting outcome events at the team level. Data is collected from over seven years of MLB team offensive statistics and their postseason success, excluding the pandemic season.

Methodology

Using probabilities to determine base state combinations allows for movement through transitions in the chain. For instance, a movement from (0,0) to (1,0) is a runner on first and still no one out. Therefore, that transition is the probability of a single, hit by pitch, walk, or error (Asaro, 2015). The key to a Markov chain simulation is to determine the summation of runs one can expect a team to get at any state of the half-inning, **given the average run expectancy matrix at that state.**

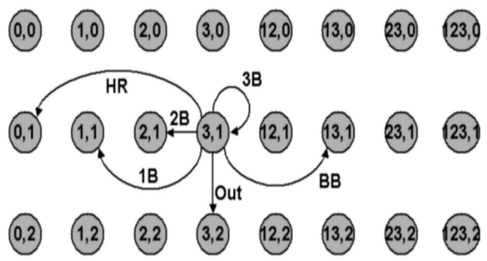
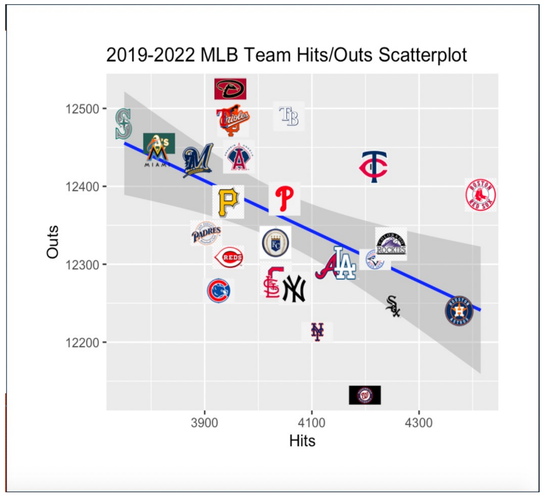


Figure 1: The 24 states with example transitions shown

Modelling through linear regression of team statistics on runs scored gains insight into the most influential variables that separate successful teams through the run scoring process. Then, XGBoost allows for machine learning classification to identify the strengths of postseason teams. These results create a glimpse into what is defined as most significant towards run efficiency.



Markov Chain Results

Scenario	Postseason RunsExp	Non-Postseason RunsExp
2 Out	0.690611833	0.644815709
Scoring Position 2 Out	0.780335243	0.738333695
1 out	2.528669438	2.45348435
Scoring position 1 Out	2.89531775	2.835999967
0 out	3.61855775	3.53045625
Scoring position 0 Out	4.134626167	4.077317333

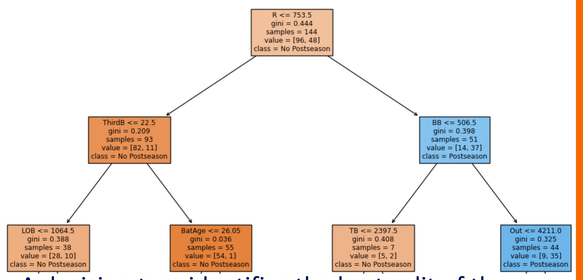
Low out situations with runners in scoring position are the highest expected run totals, with playoff teams scoring at just a .06 higher RE margin. In total, postseason teams consistently make more out of the positions with runners in scoring position than non-playoff teams.

Postseason Prediction Results

Coefficients	Estimate	Std Error	Pr(>|t|)
WalkHitByPitch	0.2898	0.03276	1.06e-15***
FirstB	0.34552	0.03341	<2e-16***
SecondB	0.82682	0.06751	<2e-16***
ThirdB	0.66583	0.2073	0.00157**
HR	1.28969	0.05845	<2e-16***
Out	-0.10729	0.03626	0.00352**
Postseason	14.05926	4.43237	0.00179**

Adjusted R-squared: .912

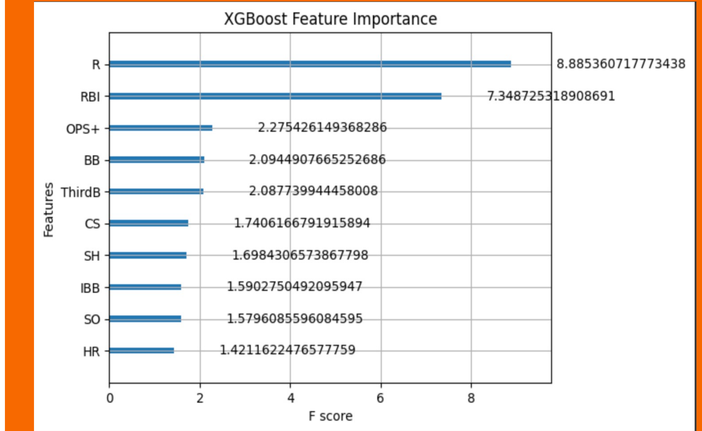
Only metrics used for transition states are used in linear regression for complete independence of batting outcomes on runs scored. The coefficient of the Postseason variable indicates that a unit increase of making the postseason leads to a 14 unit increase in runs scored. Given the length of the season, the difference between being in and out in terms of the postseason as far as efficient run production can be extremely slim (~14 runs).



A decision tree identifies the best split of the highest IG on our features towards making the postseason. In XGBoost, the errors from previous decision trees help the model's training(Ensemble learning).

XGBoost Model Weighted Accuracy: 72%

Classification Report	Precision	Recall	F1-Score
0	0.75	0.82	0.78
1	0.67	0.57	0.62



Conclusions

Our research question regarding scoring production was analyzed to create an interpretation of run efficiency in the modern-day MLB. This was done with Markov Chain modelling which identified all the specific states playoff teams performed better in. This was often in high stake situations with less outs so more opportunities to consistently put up runs. Classification and Regression techniques helped model which stats playoff teams differentiate themselves with. Understanding how the run scoring process operates at its highest efficiency helps lead to predictive analysis for future teams to model their performance on.

References

Albert, J., & Hu, J. (2020, July 30). *Probability and Bayesian Modeling*. Retrieved April 15, 2023, from https://bayesball.github.io/BOOK/probability-a-measurement-of-uncertainty.html

Asaro , V. J. (2015). *Markov League Baseball: Baseball Analysis using Markov chains*. Cal Poly. Retrieved April 16, 2023, from https://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1063&context=statsp

Beyder, E. (2015, June). *Simulation model using standardized lineup to evaluate player offensive player* . Retrieved April 15, 2023, from https://core.ac.uk/download/pdf/48501665.pdf

Bukiet, B. (n.d.). *A Markov Chain Approach to Baseball*. Retrieved April 14, 2023, from https://web.njit.edu/~bukiet/Papers/ball.pdf

Calestini, L. (2018, September 10). *The Elegance of Markov Chains in Baseball*. Medium. Retrieved April 15, 2023, from https://medium.com/sports-analytics/the-elegance-of-markov-chains-in-baseball-f0e8e02e7ac4

Cella, S. (n.d.). *Markov Chain Baseball*. Retrieved from https://www.lmtsd.org/cms/lib/PA01000427/Centricity/Domain/172/Markov%20Chain%20Baseball.pdf

Acknowledgments

I would like to thank Dr. Paul for all the help during my research endeavors.