Combinatorial and Markovian Baseball Lineup Optimization

Ryunosuke Saito, Neel Mehta, Ashwin Pasupathy, Anton Dahbura, PhD

Johns Hopkins University | Whiting School of Engineering | Baltimore, MD
Design Day 2022

Introduction

In baseball, the universal goal is to score more runs than the opponent. There are numerous variables that can be changed in a baseball game to help a team score more runs i.e. studying the opposing team's pitching and fielding mechanics, choosing where a batter should hit a ball. However, optimizing the lineup order remains the most difficult factor to task for a manager to get right. We hope to shed some light on potential methods that can be used to come closer to building a lineup that will score the most runs. The ideal lineup may vary on a game-to-game basis.

Objectives

The objective of the project is to use two different approaches to obtain a consensus lineup sequence for the Baltimore Orioles offense. Development of the optimal lineup can help guide managers about gameday lineup formation, substitutions and pinch-hits

Materials and Methods

Two different approaches were taken to formulate optimal lineups for the Baltimore Orioles.

Combinatorial Method:

Using game results from the Baltimore Orioles games, we take the plate appearance results and reorder the lineup (9! permutations) and simulate the game to find the ideal lineup to maximize runs.

To simulate situations such as double plays, sacrifice flies, scoring a run on a single from second base, etc, we use a random forest model trained from 2015-2020 to simulate those situations.

Markovian method:

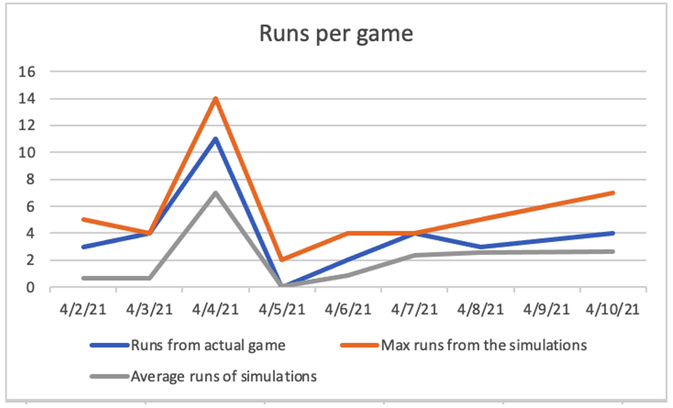
Simulation of run scoring can be performed with the use of a discrete state space time-homogeneous Markov Model, as the game transitions from state to state with conditional independence. There are 25 total states in a game, with 8 transient states per out and 1 absorbing state with 3 outs. Batted ball data for a player are obtained from Statcast stratified by pitcher type. Using a random forest to predict outcomes from batted ball data, transition matrices can be formed, and simulated run distributions can be formed

Results

Figure 1: Results of simulation vs. actual game

The orange line represents the maximum expected runs scored on any permutation for that specific game.

*Note: Since our algorithm outputs multiple optimal lineups, we currently do not have a graphical representation of specific lineups results throughout a span of games.



Model Results	Sac Fly	Double Play	Second to Home
Logistic Regression Accuracy	84.8%	71.8%	74.5%
Random Forest Accuracy	89.5%	79.1%	76.1%

Figure 2: Accuracy of Models Used In Simulation

In our game simulation, we needed a way to probabilistically determine the outcome of certain events such as sacrifice fly balls, double plays, and a second baseman going home. To accomplish this, we used batter and game data from Statcast and determined which outcomes would happen under a specific game scenario. The table shows the accuracy of the models. Random forests provided the highest accuracy amongst all the models we explored.

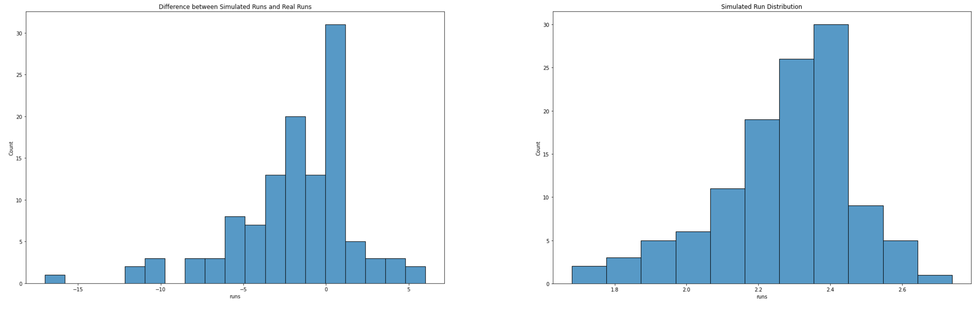


Figure 3—Markovian Simulated Run output

To validate the model, simulated run outputs for each real lineup were compared to the true run outputs for 100 of the 162 games in the Orioles 2021 season. The distribution of the difference between simulated runs and real runs is shown above for a single simulation (left), as well as the average simulated run distribution (right).

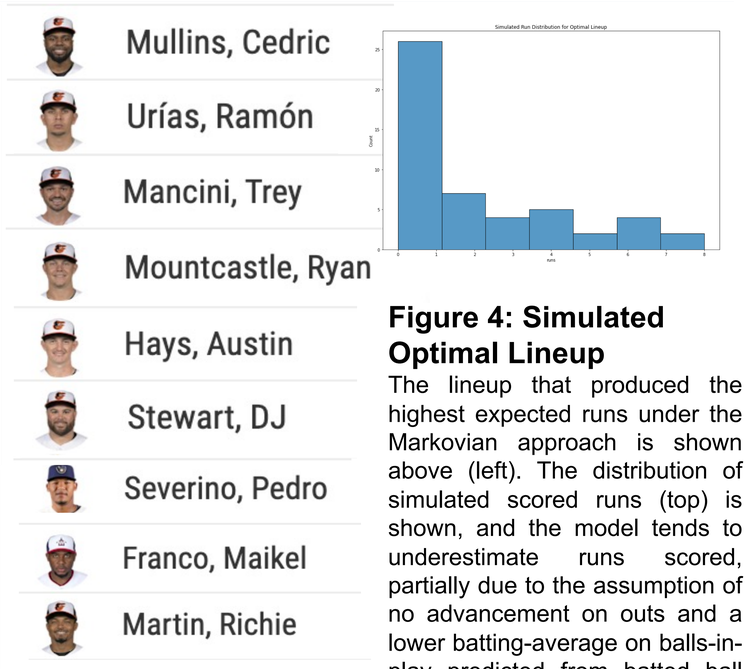


Figure 4: Simulated Optimal Lineup

The lineup that produced the highest expected runs under the Markovian approach is shown above (left). The distribution of simulated scored runs (top) is shown, and the model tends to underestimate runs scored, partially due to the assumption of no advancement on outs and a lower batting-average on balls-in-play predicted from batted ball profiles. Only one game was played with this lineup, and 3 runs were scored.

Conclusion

Through the combination of the combinatorial methods and the Markovian method, we were able to see some similarities. Both models underestimate the runs scored due to the simplification of the model. However, despite that, both models can find a lineup that would score 1-2 runs higher than the actual game runs on average, and also find a lineup that stood out in terms of its run output.

Despite the Markov model tending to underestimate the real runs produced, the model is useful in the sense that batted ball data should be a better predictor of future success that is defense and field independent, which are factors that the offensive team's manager cannot control. Future work may look at stratifying the model by pitcher handedness and using inherent lineup heuristics to more accurately reflect the decision-making process that MLB managers encounter