MAST679 - Project 2

Francis Fregeau

October 2022

Introduction

Computing accurate probabilities tied to particular events is of crucial importance in the realm of sports betting. Developing overly complicated machine learning models which can account for a plethora of events becomes tedious whenever the sample space Ω becomes too large, and very few sports can be modeled as a series of events in a straightforward way. A notable exception is the sport of baseball, which can be described as a Markov chain which states account for the following situations:

- The number of batters that have been previously eliminated (0, 1 or 2).
- Whether or not the 1^{st}, 2^{nd} and 3^{rd} bases are occupied ($2^3 = 8$ possibilities).
- Whether or not the round has ended, which occurs whenever more than 2 batters have been eliminated.

The total number of states is thus $3 \cdot 8 + 1 = 25$, and transition matrices can be build for individual players according to any arbitrary batting order. If X_t denotes the chain's state at time t and M_t denotes the transition matrix of the player batting at time t, then

$$P(X_T = j) = \left(\prod_{t=1}^{T} M_t \right)_{1,j}$$

where it is understood that the 1^{st} state of the chain is the starting one. This framework is appealing because any and every probability tied to individually or mutually occurring event(s) in Ω can be computed via Monte-Carlo using $\{M_t\}$, and batting orders can be evaluated in terms of expected points.

Literature review

Using Markov chains to model baseball games isn't a novel approach, as previous work has already been done and can easily be found via a quick google search. Examples of such works are the thesis of Daniel Joseph Ursin from the University of Wisconsin Milwaukee [1] as well as the A Markov Chain Approach to Baseball paper available on ResearchGate [2].

That being said, we do not wish to use these as either inspiration or reference given the rather straightforward nature of the modelling task at hand. The only potential issue which is obvious at first glance is the possibility that the chain is unable to attain its absorbing state (the 25^{th} one), which is mathematically expressed as

$$\exists\ t : \left(\prod_{t=1}^{T} M_t \right)_{1,25} = 0$$

and shall be dealt with by tweaking the faulty M_t, most likely by replacing it with the average of its team $\mathbf{E}[M]$ or by some other player's $M_{a \neq t} \approx M_t$. Such a problem could arise whenever we do not possess enough data tied to a player so to properly estimate his transition matrix M. All in all, we would like our work to be as original as can be.

Data

We will use the play-by-play data available on the fangraphs website. Said data was previously acquired for the seasons which occurred between the years 2011 and 2021 using a custom webscrapping tool which is available as a Python package [3]. The data was then cleaned and transformed using a custom R processing tool which is also available as a package [4]. (Note that both tools were coded between 2020 and 2021, and the data downloaded and cleaned in 2021 as part of a personal project).

Methodology

As for the methodology, we simply intend on testing diverse ways of computing M for players using different weights (decaying average) and addressing the issue where the absorbing state cannot be reached. Each method will then be benchmarked by computing the log-loss of the estimated probabilities, and an evaluation of the final result will be given, most likely by comparing it to the implied probabilities derived from the historical moneylines odds which we also scrapped.

References

[1] Daniel Joseph Ursin. *A Markov Model for Baseball with Applications*. University of Wisconsin Milwaukee, 2014. URL: https://dc.uwm.edu/cgi/viewcontent.cgi?article=1969&context=etd.

[2] Bruce Bukiet, Elliotte Harold, and José Palacios. "A Markov Chain Approach to Baseball". In: *Operations Research* 45 (Feb. 1997), pp. 14–23. DOI: 10.1287/opre.45.1.14.

[3] Francis Fregeau. *mlb webscrapper*. 2021. URL: https://github.com/c-Stats/mlb_webscrapper.

[4] Francis Fregeau. *mlb database*. 2021. URL: https://github.com/c-Stats/mlb_database.