Tony Owens

SML 320

March 8, 2024

Bayesian Pitch Valuation Project Progress Report

For my project, I intend to create a Bayesian pitch valuation model. In a baseball game, the goal of the pitcher is to prevent the offense from scoring. There are three main ways the pitcher can accomplish this: strike out the batter, the batter hits the ball in the air, and it is caught by a fielder before it hits the ground, a fielder fields the ball and throws it to a base before the corresponding runner arrives at the base. The pitcher's approach will depend heavily on the "state" of the game. The state is determined by the number of outs, and runners on base before the pitch is thrown. For example, if allowing the batter to contact a pitch has a high likelihood of giving up runs, the pitcher is more likely to attempt to strike out the batter. Conversely, if the batter contacting the ball is largely inconsequential, the pitcher may prioritize getting the batter to make weak contact with the pitch.

There are trade-offs to consider for the pitcher in every situation. If the pitcher prioritizes throwing pitches where the batter swings and misses, they run the risk of throwing pitches out of the strike-zone, which may lead to walks. This risk is magnified if the pitcher prioritizes throwing pitches in the strike-zone where the batter does not swing. On the other hand, if the batter prioritizes generating weak contact, they run the risk of the batter making strong contact, which can have greater consequences.

The fundamental utility of the proposed model is creating pitch valuations which incorporate the context of the game situation. As demonstrated, the value of a pitch depends partially on the game situation. Incorporating this context into the valuation process allows for

greater insight and evaluation of strategic decisions. The Bayesian framework also allows for outputs as probability distributions instead of point estimates. This provides greater context into the potential upside and downside of a pitch, providing additional strategic insight.

There are existing public models used for pitch valuation, perhaps the most famous of which is the "Stuff Scores" model. These models utilize the qualities of a pitch and assign it a score representing its value. These models are typically situationally agnostic. This constraint limits the ability of a user to determine the value of a pitch in a given situation. Stuff models are still useful in determining a pitcher's overall ability, but my approach aims to provide greater granularity to pitcher evaluation. Similarly to many publicly available models, this model will treat the difference in run expectancy before and after the pitch as the evaluation metric.

This model makes use of two data sources: Statcast data from 2018-2023 (excluding 2020), and Fangraphs data containing evaluations from the existing pitch valuation models. The Statcast dataset contains around 4.2 million rows, where each row represents a pitch. The columns contain information describing the pitch (velocity, movement, location), information describing the game state (the pitcher, the batter, the count, etc.) and information describing the outcome of the pitch. The data was acquired with the BaseballR package.

When examining the Statcast data, we find that both contact rate and swing rate vary based on where in the zone the pitch is thrown. This demonstrates that the location is an important factor in determining both the likelihood of a batter to swing, and the likelihood of the batter to make contact should they swing.

The Fangraphs data contains evaluations of pitches for different pitchers for the 2022 and 2023 seasons. I intend to treat 2022 and 2023 as testing data for the model, and for pitchers who appear in the 2022 and 2023 data, I will compare their cumulative pitch value from my model to

their value from the publicly available models. When examining the data, it is apparent that Stuff+ and Location+ exhibit noticeable variance based on the year of observation. In baseball analytics, a statistic ending with a + symbol indicates that it is normalized with 100 as average. These observations warrant further investigation to determine the root cause of the disparities. One hypothesis I have is that the decrease in the metrics is in part influenced by the implementation of the pitch clock in 2023. With less time to rest between pitches, pitchers cannot "dial in" as well and it impacts their performance.

For the conjugate priors, I opted for a Beta Binomial model for determining swing rate. I began with a fairly uninformative prior Beta(5,5) assuming batters swing at 50% of pitches they see. After sampling the Statcast data for 50 pitches, I observed 21 swings. This outputs a Beta(26,34) where I believe the swing rate is around 43%. After concern that 50 pitches is an insufficient sample, I sampled 500 different pitches and found 224 swings. Treating the Beta(26,34) as the prior, I arrive at a Beta(252,308) which predicts swing rate at around 45% lying between 40% and 50%.

For the next steps with the model, I first intend to develop a clear framework for predicting whether a batter will swing at a pitch, and whether they will contact the pitch should they swing. The model should output these results as probability distributions, which will provide further insight into the uncertainty behind the evaluations. From there, I will shift my focus to determining the value of a pitch based on the expected change in run expectancy. Finally, I will work to integrate these two steps into a model which outputs a single pitch valuation distribution. The ability to break the model down into many steps makes me optimistic about my ability to achieve the goals expressed.