To Ball or Not to Ball

-Dylan Perlstein

★★★★★★★

Table of contents

01 Modelling

Model

Logistic Regression

- Used a Logistic Regression to predict the logit(p(Ball)),
- Then got the log(odds ratio) to get the probabilities of a ball

$$\textit{Ball} \sim \beta_0 + \beta_1 \textit{HandMatch} + \beta_2 \textit{BatterQuality} + \beta_3 \textit{PitcherQuality} + \beta_i \textit{Count:Outs:Base Sate}$$

- As we can see, the count, outs, and base states are interacted, leading to 96 different Betas for the interactions

Parameters

- Count
  - Balls and Strikes
- Base State
  - Only looked at 4 base states: “Loaded”, “RISP”, “Men On”, “Empty
- Outs
  - 0, 1, 2
- Batter and Pitcher handedness
  - This is a binary category for if the batter and pitcher use the same hand or not
  - For Example: 1 if the Pitcher is R and the batter is R, 0 if Pitcher is R and Batter is L
- Batter and Pitcher Quality metrics
  - rWoba: a rolling metric that computes weighted on base average (basically it adds weights to each at bat outcome and takes the Sum/number of at bats)
  - Used Empirical Bayes to account for lack of sample size early in the season
    - mean(MLBwOBA) + rWoba / (50 + PA)
    - 50 = “Fake” data, PA is actual number of plate appearances

02

Predictions

Validity

Are we seeing things that make sense logically?

- Hand Match: The coefficient is negative. In baseball, batters, on average, are worse against the same hand, so getting a ball, a good batter event, should be less likely
- Beta for 3 balls 0 strikes 0 outs Empty: -17
  - Why does this make sense? In baseball, hitters swing about 10% of the time when it is 3-0, and pithers know this, so they “steal” a strike knowing they aren’t swinging
- As we can see, the model is lining up with general baseball knowledge: Good Sign!

Accuracy

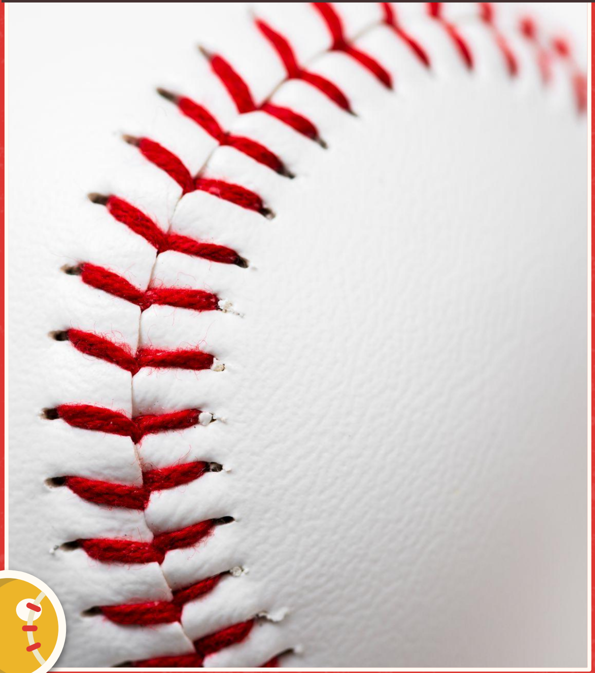
Log Loss

- When modelling, I tried many different models, based on different interactions and treating variables as numeric or categorical and needed to assess which was the best
- Actual: 1 if the pitch was a ball, 0 if not

$$log\ loss = \frac{1}{N}\sum_{i=1}^{N} actual * log(p(Ball)) + (1 - actual) * (1 - log(p(Ball))$$

03

Applications

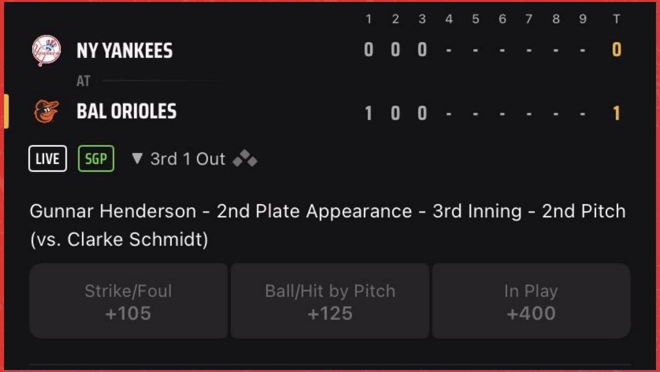


Sports Betting

- Disclaimer: I do not condone gambling and sportsbooks are hard to beat
- It's really hard to beat the book on game odds, but live odds are much harder for them to predict so there may be an edge
- Sportsbooks now offer betting lines on the outcome of each individual pitch!
- This is what my model is predicting, so if my model's odds of a ball are higher than the implied gambling odds, ARBITRAGE = $$$

Example 1

Gunnar Henderson vs Clarke Schmidt: 1 out, 0-1 with Empty base state

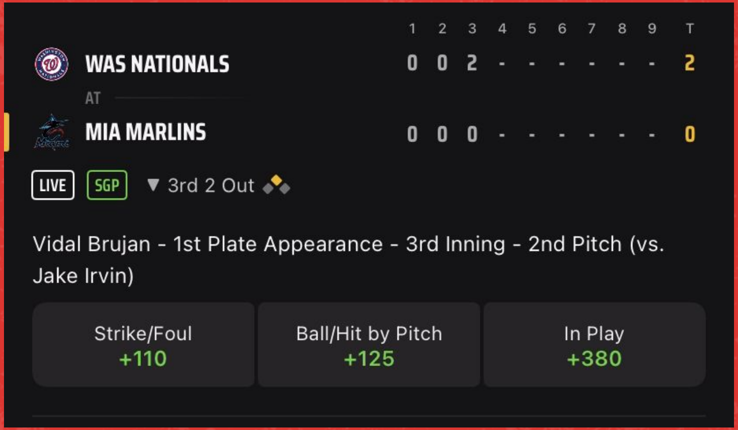


My model predicts a ball with probability .617, while the implied probability of +125 odds is .44, therefore there is value in this bet

What actually happened? It was a ball

Example 2

Vidal Brujan vs Jake Irvin: 2 outs, 0-1 with RISP



Here, again, we have +125 odds for a ball which equals an implied probability of.44. Taking into account the game states and the hitter and batter, my model gives a probability of ball of .31, which is far lower than the odds, therefore I would not bet.

What happened? A strike, which is not a ball.

Thanks!

Do you have any questions?