



# *To Ball or Not to Ball*

-Dylan Perlstein



# *Table of contents*

**01**

## ***Model***

- Predictors
- Assumption

**02**

## ***Predictions***

- Validity
- Accuracy

**03**

## ***Applications***

- Baseball betting





**01**

# *Modelling*



# Model

## Logistic Regression

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

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

- 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
    - $\text{mean(MLBwOBA)} + \text{rWoba} / (50 + \text{PA})$
    - 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 pitchers 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

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



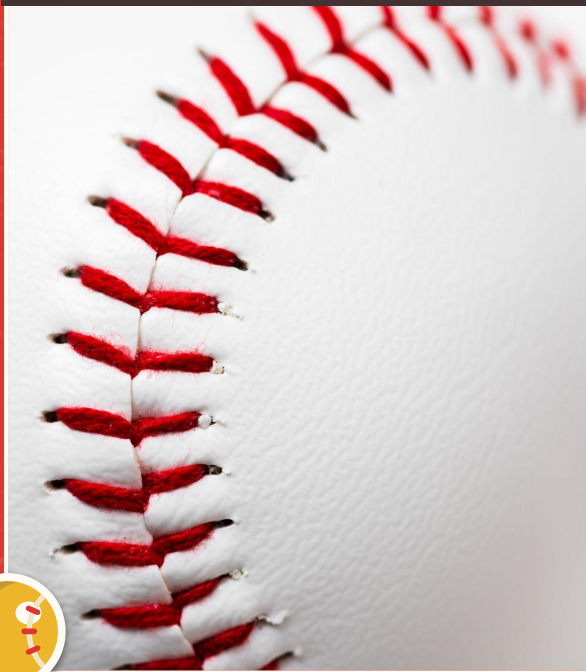


03



# *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

The screenshot shows a live baseball game between the NY Yankees and the BAL Orioles. The score is 0-1 in the 3rd inning. The current batter is Gunnar Henderson, and the pitcher is Clarke Schmidt. The game is in the 3rd inning, 1st out. Three betting options are shown: Strike/Foul (+105), Ball/Hit by Pitch (+125), and In Play (+400).

	1	2	3	4	5	6	7	8	9	T
NY YANKEES	0	0	0	-	-	-	-	-	-	0
BAL ORIOLES	1	0	0	-	-	-	-	-	-	1

LIVE SGP ▼ 3rd 1 Out

Gunnar Henderson - 2nd Plate Appearance - 3rd Inning - 2nd Pitch (vs. Clarke Schmidt)

Strike/Foul +105	Ball/Hit by Pitch +125	In Play +400
---------------------	---------------------------	-----------------

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

The screenshot shows a live baseball game between the Washington Nationals (WAS) and the Miami Marlins (MIA). The score is WAS 2, MIA 0. The game is in the 3rd inning, 2 outs, with a runner in scoring position (RISP). The batter is Vidal Brujan, and the pitcher is Jake Irvin. The odds for the current play are:

Outcome	Odds
Strike/Foul	+110
Ball/Hit by Pitch	+125
In Play	+380

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?***



**CREDITS:** This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**