Baseball pitch trajectory density estimation for predicting future pitcher outcomes

Scott Powers and Vicente Iglesias

Conference of Texas Statisticians 2024

RICE UNIVERSITY
Sport Analytics

The Problem

MLB teams spend A LOT of money on pitchers ...

PLAYER	POS	TEAM SIGNED WITH	AGE AT SIGNING	START	END	YRS	VALUE
Shohei Ohtani	SP	LAD	29	2024	2033	10	$700,000,000
Yoshinobu Yamamoto	SP	LAD	25	2024	2035	12	$325,000,000
Gerrit Cole	SP	NYY	29	2020	2028	9	$324,000,000
Stephen Strasburg	SP	WSH	31	2020	2026	7	$245,000,000
Jacob deGrom	SP	TEX	34	2023	2027	5	$185,000,000
Aaron Nola	SP	PHI	30	2024	2030	7	$172,000,000
Patrick Corbin	SP	WSH	29	2019	2024	6	$140,000,000

spotrac.com

... and they don't always know who the best ones are.

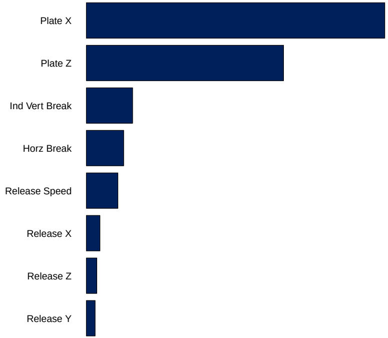
Standard of Practice

- Observe $X \in \mathbb{R}^9$ describing each pitch trajectory
- Observe $Y \in \mathbb{R}$ describing the run value of the pitch outcome
- Estimate $f(x) = \mathbb{E}[Y \mid X = x]$
  - This is a standard supervised learning problem
- Evaluate the pitcher using $\frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_i)$
  - This turns out to work better than using actual outcomes

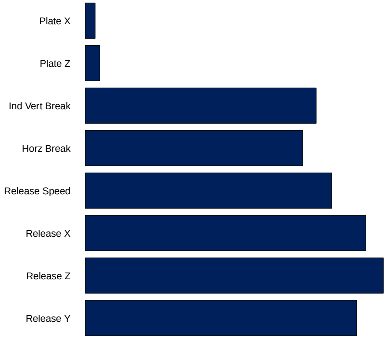
Let's call this the **Descriptive** model, e.g. Healey (2019)

The Conundrum

Variable Importance[1]



Variable Reliability[2]



[1] fractional contribution of each feature's splits to gradient boosting pitch model
[2] (between-pitcher variance) / (total variance); varies by pitch type (here: RHB FB)

Why Supervised Learning Isn't Enough

Supervised Learning

Pitch		Outcome
x_1	$\rightarrow$	y_1
x_2	$\rightarrow$	y_2
x_3	$\rightarrow$	y_3
	$\dots$	
x_n	$\rightarrow$	y_n
x^*	$\rightarrow$	$\hat{y}$

Our Problem

Pitcher	Pitch		Outcome
A	x_1	$\rightarrow$	y_1
A	x_2	$\rightarrow$	y_2
B	x_3	$\rightarrow$	y_3
		$\dots$	
C	x_n	$\rightarrow$	y_n
		$\searrow$	
A			$\hat{y}$

Why Supervised Learning Isn't Enough

Supervised Learning

Pitch		Outcome
x_1	$\rightarrow$	y_1
x_2	$\rightarrow$	y_2
x_3	$\rightarrow$	y_3
	$\dots$	
x_n	$\rightarrow$	y_n
x^*	$\rightarrow$	$\hat{y}$

Our Solution

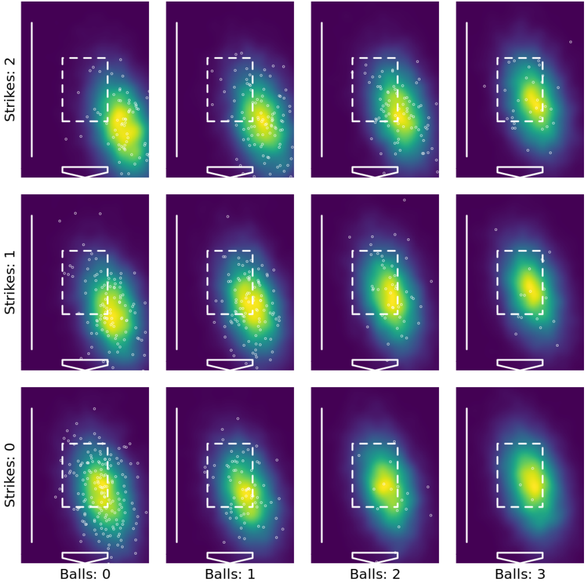
Pitcher	Pitch		Outcome
A	x_1	$\rightarrow$	y_1
A	x_2	$\rightarrow$	y_2
B	x_3	$\rightarrow$	y_3
		$\dots$	
C	x_n	$\rightarrow$	y_n
	$\downarrow$		
A	$\hat{p}(x)$	$\rightarrow$	$\int \hat{p}(x)\hat{f}(x)$

Our Solution

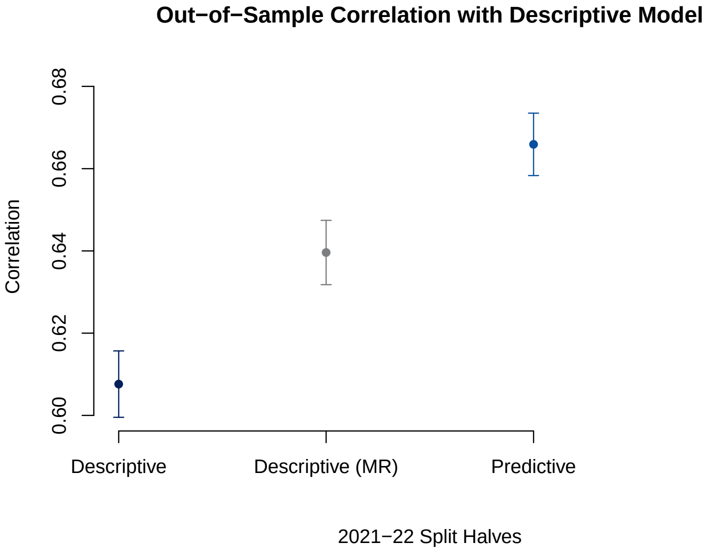
1. Estimate the probability distribution over pitch trajectories
   - Depends on pitcher, batter side, count, etc.
   - We use a Bayesian hierachical model to share information
   - Expensive to sample from posterior (81 parameters per pitcher)
   - We find MAP model fit using automatic differentiation
2. Fit a model to predict pitch outcome given its trajectory
   - We use gradient boosting, not the focus today
3. Integrate the model 2. w.r.t. the distribution 1.

Let's call this the **Predictive** model

Dylan Cease's Slider vs RHB in All Counts



Does It Work?


Out−of−Sample Correlation with Descriptive Model
Correlation
0.60
0.62
0.64
0.66
0.68
Descriptive
Descriptive (MR)
Predictive
2021−22 Split Halves

Next steps

- Better (simpler?) parameterization for distribution model
- Relax Gaussian assumption (unimodal with specific tails)

Thank You!

saberpowers.github.io

References

Healey G (2019) “A Bayesian method for computing intrinsic pitch values using kernel density and nonparametric regression estimates” *Journal of Quantitative Analysis in Sports* 15(1) 59-74