Analysis of Pitch Type Prediction using Individualized Models

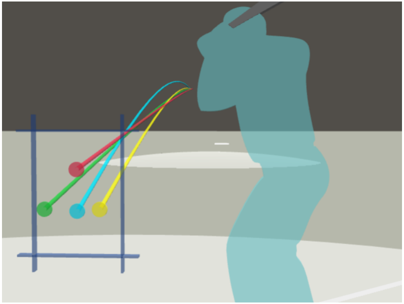
STA4241 Final Project

Zachary Allen

December 5, 2023

Pitch Types

- Pitchers throw different pitch types to disrupt the timing, vision, and swing of hitters
- Primarily determined by grip on the ball and arm/hand motion
- Results in pitches with varying velocities, spin, and horizontal/vertical movement



Justin Verlander 2023 Avg. Pitches to LHH
(via Baseball Savant)

Pitch Type	Velocity (mph)	Spin (rpm)
Four-seam Fastball	94.3	2419
Slider	86.7	2513
Curveball	78.0	2704
Changeup	85.0	1805

Pitch Type	Δx (ft.)	Δz (ft.)
Four-seam Fastball	-0.688	1.550
Slider	0.367	0.305
Curveball	0.595	-1.128
Changeup	-1.277	0.736

Knowledge of Pitch Types

If a batter had perfect knowledge of what pitch type would be thrown in advance, this would greatly increase their chance of making contact with the ball.

2017 Houston Astros

- Caught illegally stealing catcher-to-pitcher signals that indicate what pitch type would be thrown in real time by using cameras.
- Would bang on trash cans in dugout to indicate breaking pitches (curveball, slider, sinker, etc.)
- 93% Success Rate predicting off-speed pitches

Project Objective

The objective of this project is to create a binary classifier that predicts whether an upcoming pitch will be a fastball or not and is:

- Better than random guessing and baseline model
- Individualized - models are fit to data associated with specific pitchers
- Utilizing publicly available pitch data
- Based on techniques and models learned in STA4241

Feature Selection

For every pitch there are 85 features describing that pitch.

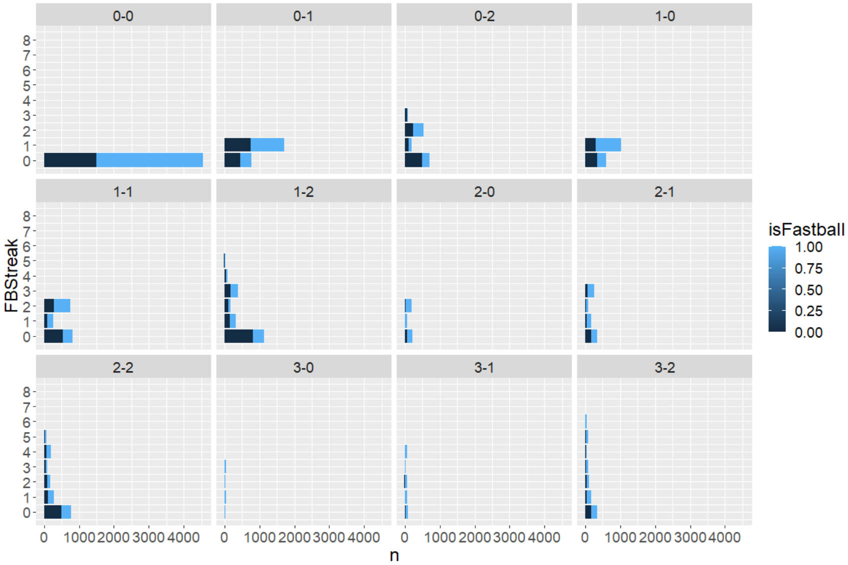
Features Used

- Count (12 factors)
- Runners on Each Base (2 factors each)
- Number of outs (3 factors)
- Hitter Stance (2 factors)
- Current Inning
- Year (factor)

New Features

- Pitches since last breaking ball (per batter)
- Most Recent Pitch Type (per batter, 3 factors)
- Times batter has faced pitcher (per game)

FBStreak



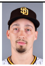
Justin Verlander Career FB/OS Breakdown for each (FBStreak, Count) Pair

Model Testing

1. Naive Bayes
2. Multiple Logistic Regression
3. Boosting

Naive Bayes

Predicting based on which class probability is greater leads to suboptimal test errors

	Name	Total Pitches	Test Error	Baseline Error
	Justin Verlander	20987	0.3510	0.4450
	Zach Eflin	12901	0.3783	0.3911
	Clayton Kershaw	19255	0.3555	0.4459
	Blake Snell	17420	0.4265	0.4771
	Shohei Ohtani	7613	0.3688	0.3757
	Average	**15635**	**0.3750**	**0.4426**

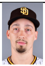
Naive Bayes - Varied Threshold

However, considering only predictions where the greater class probability is above a certain threshold leads to more favorable results.

Threshold	Accuracy	Num. Test Labels	% of Total Test Labels
0.50	0.6490400	3177	100.000000
0.60	0.6768169	2463	77.525968
0.75	0.7468220	944	29.713566
0.80	0.7629630	540	16.997167
0.90	0.8095238	126	3.966006
0.95	0.8644068	59	1.857098

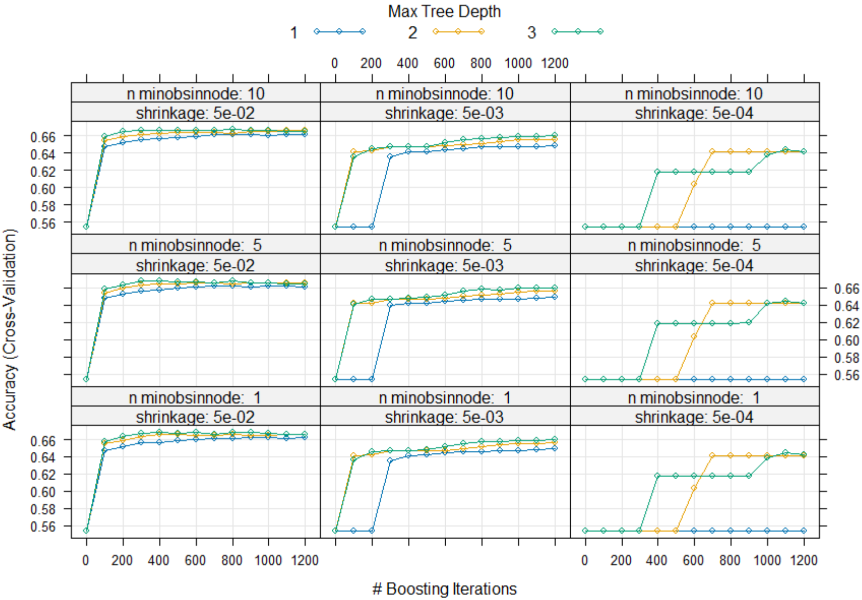
Table: Justin Verlander Naive Bayes, Accuracy by Threshold

Multiple Logistic Regression

	Name	Total Pitches	Test Error Rate	AUC
	Justin Verlander	20987	0.3516	0.6920
	Zach Eflin	12901	0.3839	0.6420
	Clayton Kershaw	19255	0.3488	0.6918
	Blake Snell	17420	0.4177	0.6179
	Shohei Ohtani	7613	0.362	0.6415
	Average	**15635**	**0.3728**	**0.65704**

- Most significant predictors can vary greatly between pitchers

Boosting



Max Tree Depth
1
2
3
n minobsinnode: 10
shrinkage: 5e-02
shrinkage: 5e-03
shrinkage: 5e-04
n minobsinnode: 5
n minobsinnode: 1
Accuracy (Cross-Validation)
Boosting Iterations
0.56
0.58
0.60
0.62
0.64
0.66
0
200
400
600
800
1000
1200

Boosting Results

Pitcher	Trees	Depth	Shrinkage	Min Obs In Node	Test Error
Justin Verlander	1200	3	0.05	1	0.3528
Zach Eflin	1200	3	0.05	1	0.3612
Clayton Kershaw	900	3	0.05	1	0.3292
Blake Snell	800	3	0.05	1	0.3972
Shohei Ohtani	300	3	0.05	5	0.3569
				Average	**0.3595**

Comparing All Results

Test Error Rates for All Models and Pitchers

Pitcher	Baseline	Naive Bayes	MLR	Boosting
Justin Verlander	0.4450	0.3510	0.3516	0.3528
Zach Eflin	0.3911	0.3783	0.3839	0.3612
Clayton Kershaw	0.4459	0.3555	0.3488	0.3292
Blake Snell	0.4771	0.4265	0.4177	0.3972
Shohei Ohtani	0.3757	0.3688	0.3620	0.3569
Average	**0.4426**	**0.3750**	**0.3728**	**0.3595**

Conclusion and Limitations

- Overall average test errors result in 6-8% improvements over baseline
  - Largest per-pitcher improvement was 12% (Kershaw, Boosting)
  - Varying thresholds leads to promising results
  - Teams could relay info to batter at just the right time, when most confident
- Missing potentially helpful predictors that cannot be measured
  - Difference between where pitcher intended to throw a pitch versus where it ended up
  - Pitching coach tendencies
  - Hitter weaknesses
- For Boosting, it is very likely better parameters exist that would further reduce test error rate.

References

Astros Cheating Analysis by Jake Mailhot:
1: Most Important Bangs of Astros' Scheme
2: How Much Did the Astros Really Benefit from Sign-Stealing?

BaseballR Library:
`https://billpetti.github.io/baseballr/index.html`

StatCast Column Descriptions:
`https://baseballsavant.mlb.com/csv-docs`

Project GitHub Repo:
`https://github.com/ZackAllen1/pitch-prediction`