Predicting Strikeout Percentage Among MLB Relief Pitchers

By: Luke VanHouten
Mentored by: Alex Bank

DRP Outline

- Motivation
  - I am very interested in baseball statistics
  - I want to work with prediction
  - There is more complexity and nuance with pitching statistics
  - I would like to work with a complex dataset
    - Pitching provides this

Baseball Context

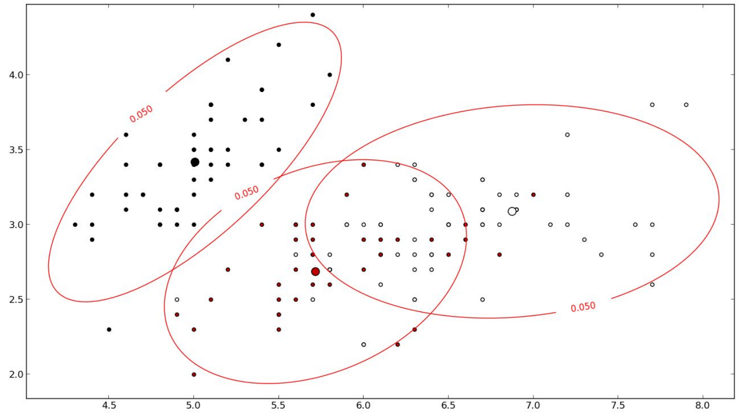
- Pitcher
  - The pitcher’s role is to throw the baseball to home plate, hopefully without the batter being able to hit the ball
  - They aim for the strike zone, an imaginary box by the batter’s torso
- Strikes and strikeouts
  - If a batter does not hit a ball that goes through this zone, it is a strike
  - 3 strikes and you’re out
  - Strikeout percentage is a stat that lets us standardize strikeouts to not take into account how much a player has played

What I Learned - Reading

- “Predicting Major League Baseball Strikeout Rates from Differences in Velocity and Movement Among Player Pitch Types” - Eric Martin
    - MIT Sloan Sports Analytics Conference 2019
    - Want to predict strikeout percentage based on pitch data among different pitch types for starting pitchers
        - Pitch data includes pitch speed and movement, to be aggregated
        - Cluster pitches based on pitch data rather than name of pitch
            - Some players may throw different pitches that still look the same to the batter

What I Learned - Reading

- Clustering uses Gaussian mixture models
    - Identifying probabilistic regions to separate pitches into different categories
    - As opposed to k-means clustering, where these clusters are set beforehand
        - Pitches now have a lot of variability within each pitch type that can be captured with a distribution
- Max velocity, strike percentage, and vertical movement IQR best predictors of strikeout percentage among these clusters



Scope of My Project

- Use techniques found in the paper to predict strikeout percentage at the pitcher level, as opposed to per pitch type
- Instead, have clustering be just an extension of the prediction model
- Focus on relief pitchers
    - Future question brought up in original paper
    - Include stats for handedness matchups and game leverage
        - More important for relief pitchers as managers choose when they enter the game

Data Sources - Statcast

- Statcast pitch data - 2015-2021
    - Ignoring 2020 season due to effects of COVID
        - Allows for 2021 to be easily split to test data
    - 4.4 million pitches for these 6 years
    - Scrape from Baseball Savant to PostgreSQL server using R
- Learned some SQL to more efficiently the group pitchers through queries
    - Still have all pitch data available

public.statcast: 4,479,611 rows total, limited to 1,000 » Next ⬍ Show all ▼ Sorting (3) ▼ Columns (98/98) ▼ Filter

pitch_t...	game_date	release_speed	release_pos_x	release_pos_z	player_name	batter	pitcher	events	description	spin_dir	spin_rate_deprecated	break_angle_deprecat...	break_length_depreca...	zone	des
SL	2015-04-06	82.6	-1.27	5.54	Zimmerman, Ryan	475,582	112,526		ball	(NULL)	(NULL)	(NULL)	(NULL)	14	Rya
SL	2015-04-06	81.7	-1.41	5.68	Zimmerman, Ryan	475,582	112,526		called_strike	(NULL)	(NULL)	(NULL)	(NULL)	7	Rya
SL	2015-04-06	83	-1.03	5.7	Desmond, Ian	435,622	112,526		foul	(NULL)	(NULL)	(NULL)	(NULL)	14	Ian
SL	2015-04-06	84.1	-1.22	5.56	Taylor, Michael A.	572,191	112,526		ball	(NULL)	(NULL)	(NULL)	(NULL)	14	Mic
SL	2015-04-06	84.1	-1.12	5.58	Desmond, Ian	435,622	112,526		called_strike	(NULL)	(NULL)	(NULL)	(NULL)	9	Ian
SL	2015-04-06	83.3	-1.34	5.54	Taylor, Michael A.	572,191	112,526		foul_tip	(NULL)	(NULL)	(NULL)	(NULL)	14	Mic
SL	2015-04-06	84.5	-1.47	5.46	Zimmerman, Ryan	475,582	112,526		foul	(NULL)	(NULL)	(NULL)	(NULL)	8	Rya
SL	2015-04-06	83.4	-1.26	5.57	Desmond, Ian	435,622	112,526	field_out	hit_into_play	(NULL)	(NULL)	(NULL)	(NULL)	14	Ian
SL	2015-04-06	86.4	-0.86	5.71	Scherzer, Max	453,286	112,526		called_strike	(NULL)	(NULL)	(NULL)	(NULL)	9	Ma
SL	2015-04-06	82.3	-1.33	5.72	Escobar, Yunel	488,862	112,526		called_strike	(NULL)	(NULL)	(NULL)	(NULL)	3	Yu
SL	2015-04-12	83.3	-0.97	5.68	Markakis, Nick	455,976	112,526	field_out	hit_into_play	(NULL)	(NULL)	(NULL)	(NULL)	8	Nic
SL	2015-04-12	82.1	-1.22	5.64	Gomes, Jonny	430,404	112,526		foul	(NULL)	(NULL)	(NULL)	(NULL)	14	Jor
SL	2015-04-12	85.4	-1	5.8	Young Jr., Eric	458,913	112,526	field_out	hit_into_play	(NULL)	(NULL)	(NULL)	(NULL)	2	Eri
SL	2015-04-17	83.6	-1.2	5.8	Realmuto, J.T.	592,663	112,526		called_strike	(NULL)	(NULL)	(NULL)	(NULL)	6	J.T
SL	2015-04-17	84.7	-1.03	5.79	Stanton, Giancarlo	519,317	112,526		foul	(NULL)	(NULL)	(NULL)	(NULL)	14	Gia
SL	2015-04-17	84.4	-1.3	5.64	Stanton, Giancarlo	519,317	112,526		foul	(NULL)	(NULL)	(NULL)	(NULL)	9	Gia
SL	2015-04-17	83.7	-1.41	5.54	Stanton, Giancarlo	519,317	112,526		foul	(NULL)	(NULL)	(NULL)	(NULL)	8	Gia
SL	2015-04-17	85.1	-1.04	5.49	Realmuto, J.T.	592,663	112,526		blocked_ball	(NULL)	(NULL)	(NULL)	(NULL)	14	J.T
SL	2015-04-17	83.9	-1.33	5.66	Hechavarria, Adeiny	588,751	112,526	field_out	hit_into_play	(NULL)	(NULL)	(NULL)	(NULL)	8	Ade
SL	2015-04-17	81.7	-1.29	5.73	Prado, Martín	445,988	112,526		called_strike	(NULL)	(NULL)	(NULL)	(NULL)	5	Ma
SL	2015-04-17	83.4	-1.28	5.57	Morse, Michael	434,604	112,526	field_out	hit_into_play	(NULL)	(NULL)	(NULL)	(NULL)	9	Mic
SL	2015-04-17	85.4	-1.32	5.64	Realmuto, J.T.	592,663	112,526		foul	(NULL)	(NULL)	(NULL)	(NULL)	6	J.T
SL	2015-04-23	81.7	-1.19	5.63	Gomes, Jonny	430,404	112,526		swinging_strike	(NULL)	(NULL)	(NULL)	(NULL)	9	Jor
SL	2015-04-23	82.3	-1.22	5.53	Gomes, Jonny	430,404	112,526		ball	(NULL)	(NULL)	(NULL)	(NULL)	9	Jor
SL	2015-04-23	82.2	-1.28	5.9	Freeman, Freddie	518,692	112,526		ball	(NULL)	(NULL)	(NULL)	(NULL)	13	Fre
SL	2015-04-23	83.7	-1.16	5.75	Markakis, Nick	455,976	112,526		called_strike	(NULL)	(NULL)	(NULL)	(NULL)	4	Nic
SL	2015-04-23	82.9	-0.9	5.6	Gomes, Jonny	430,404	112,526		ball	(NULL)	(NULL)	(NULL)	(NULL)	14	Jor
SL	2015-04-23	82.1	-1.26	5.56	Gomes, Jonny	430,404	112,526		foul	(NULL)	(NULL)	(NULL)	(NULL)	6	Jor
SL	2015-04-29	84.3	-1.21	5.65	Stanton, Giancarlo	519,317	112,526		called_strike	(NULL)	(NULL)	(NULL)	(NULL)	9	Gia
SL	2015-04-29	82.5	-1.28	5.79	Morse, Michael	434,604	112,526	field_out	hit_into_play	(NULL)	(NULL)	(NULL)	(NULL)	14	Mic
SL	2015-04-29	83.3	-1.29	5.65	Hechavarria, Adeiny	588,751	112,526	double	hit_into_play	(NULL)	(NULL)	(NULL)	(NULL)	5	Ade
SL	2015-04-29	82	-1.32	5.57	Realmuto, J.T.	592,663	112,526	field_out	hit_into_play	(NULL)	(NULL)	(NULL)	(NULL)	14	J.T
SL	2015-04-29	82.1	-1.29	5.61	Ozuna, Marcell	542,303	112,526	strikeout	swinging_strike	(NULL)	(NULL)	(NULL)	(NULL)	14	Ma
SL	2015-04-29	83.3	-1.04	5.51	Prado, Martín	445,988	112,526		ball	(NULL)	(NULL)	(NULL)	(NULL)	14	Ma
SL	2015-04-29	84.2	-1.24	5.49	Stanton, Giancarlo	519,317	112,526		ball	(NULL)	(NULL)	(NULL)	(NULL)	14	Gia
SL	2015-04-29	81.8	-1.23	5.64	Ozuna, Marcell	542,303	112,526		swinging_strike	(NULL)	(NULL)	(NULL)	(NULL)	14	Ma
SL	2015-04-29	83.3	-1.19	5.56	Ozuna, Marcell	542,303	112,526		blocked_ball	(NULL)	(NULL)	(NULL)	(NULL)	14	Ma

Data Sources - Other

- FanGraphs
    - Leverage Index
    - Access using baseballr API
- Chadwick Bureau
    - Player ID’s to join above datasets
    - 16 CSV files
- Merge all of these to create one dataframe, aggregated per pitcher per year

Data Sources - Statistics to Add

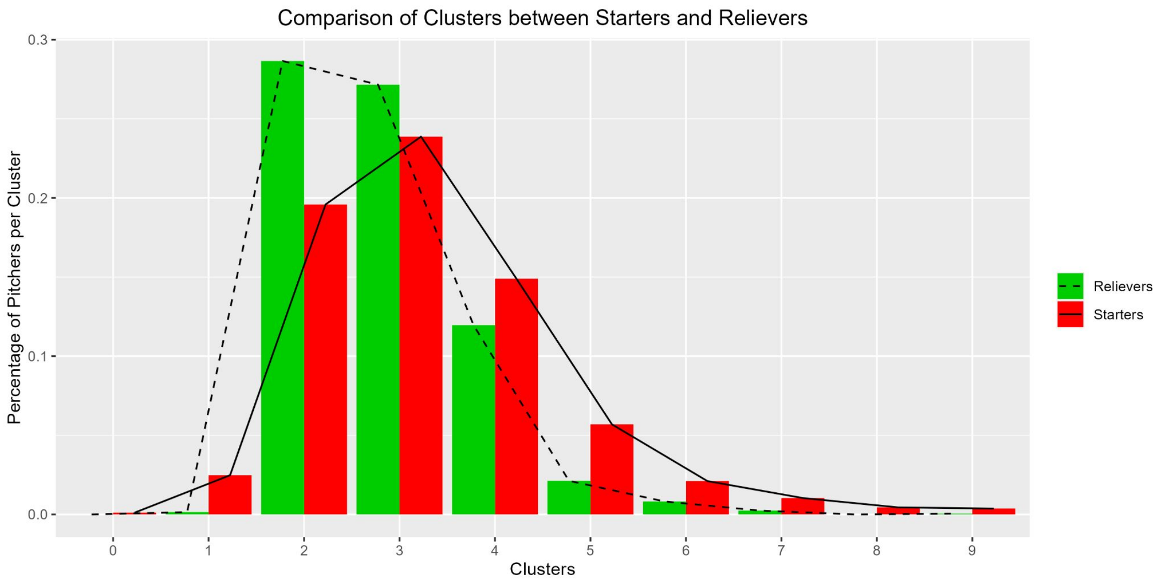
- Strikeout Percentage
- Strike Percentage
  - Percentage of pitches that are not balls
- First Batter Platoon Advantage
  - The percentage of batters faced that are the first for that pitcher's appearance and have the same handedness as the pitcher
    - Pitchers have more favorable matchups when this the case

Data Sources - Statistics to Add

- Pitches, Games, Pitches per Game
    - Relief pitcher eligible if <1000 pitches, >3 games, <45 pitches per game
- Leverage Indexes
    - The amount the win probability can change per event in the game
    - pLI
        - The average leverage index for all events during an appearance
    - gmLI
        - The average leverage index when the reliever enters the game

Data Sources - Statistics to Add

- Number of Clusters
    - Implemented the Gaussian mixture models for all pitches per player per year for number of clusters
    - Only clustered by pitch speed and movement to improve computation speed
    - Relievers on average had one less pitch cluster than starting pitchers



Model - Setup

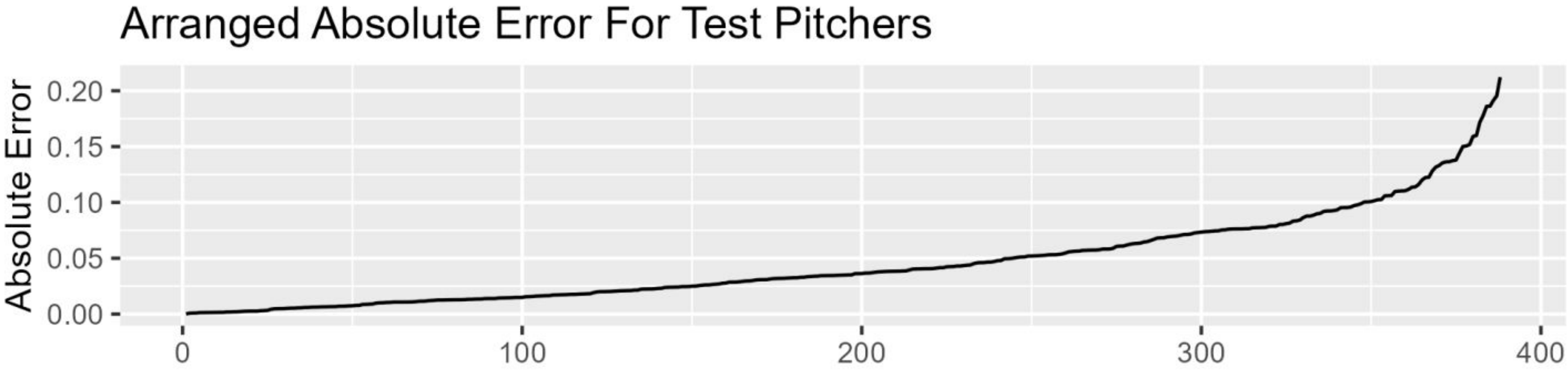
- XGBoost used for predictive machine learning model
  - Simple and powerful
- All data split into training and testing data
  - Training data 2015-2019
    - 1670 pitchers
  - Testing data 2021
    - 388 pitchers
  - ~20% train test split

Model - Setup

- Hyperparameter tuning - targeting MAE
  - Example hyperparameter: ETA
    - Controls the learning rate of the model
- Want to find the the combination of these hyperparameters that give the lowest MAE
  - Used a grid search for this

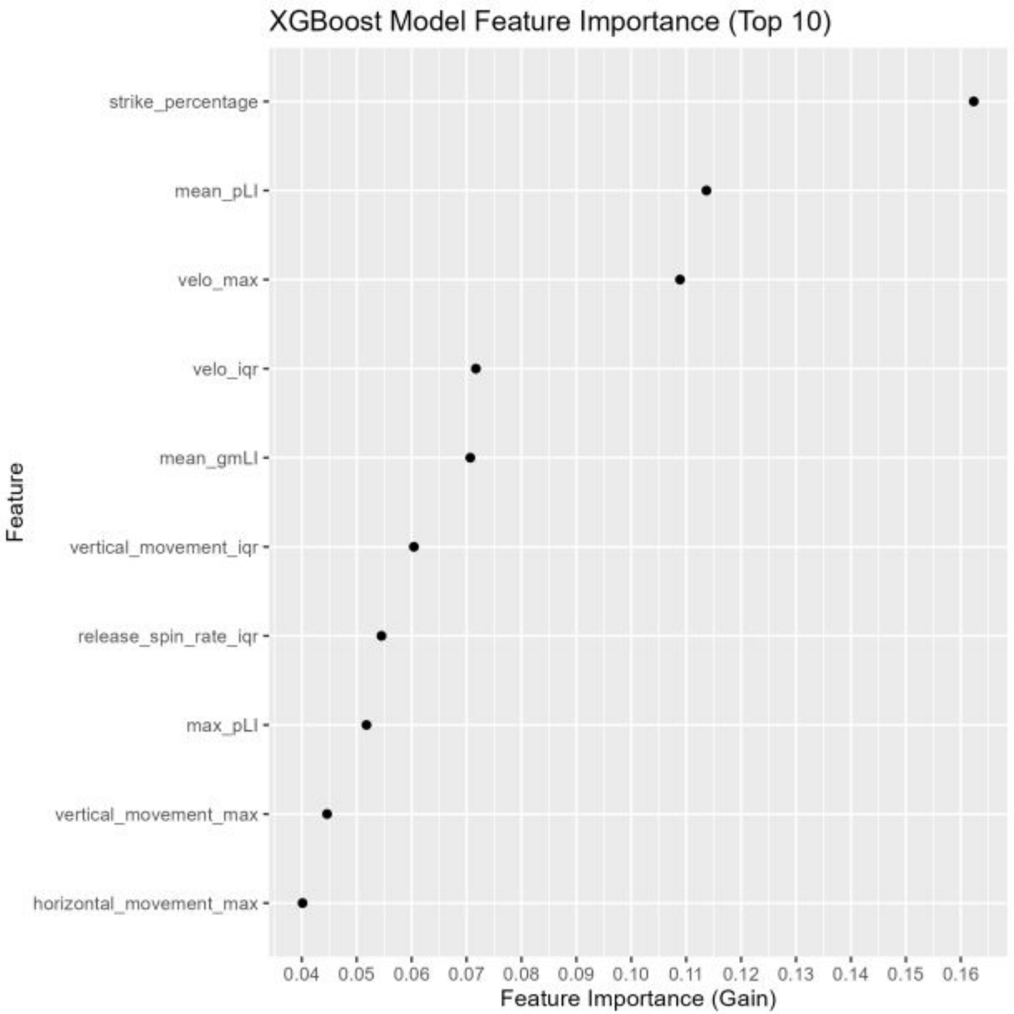
Model - Results

- Final MAE ~0.046, or 4.6%
    - MAE in original paper 0.0294 (2.94%), so off by a factor of ~1.5
        - Less data and different method likely accounts for more error



Model - Results

- Feature Importance
  1. Strike percentage
  2. Average leverage index for all game events
  3. Max velocity
  4. Velocity IQR
  5. Average leverage index at the beginning of the inning



Future

- Dealing with outliers in pitch clustering
- Doing the prediction within each cluster
    - Similar to the original paper
- Pitch sequencing
    - Different permutations of sequences of the pitch clusters

Thank you!