Predicting MLB Pitches

Metis Final Project

Patrick Bovard

Agenda

- Motivation and Goals
- Data Collection and Modeling
- Results and Takeaways
- Future Improvements

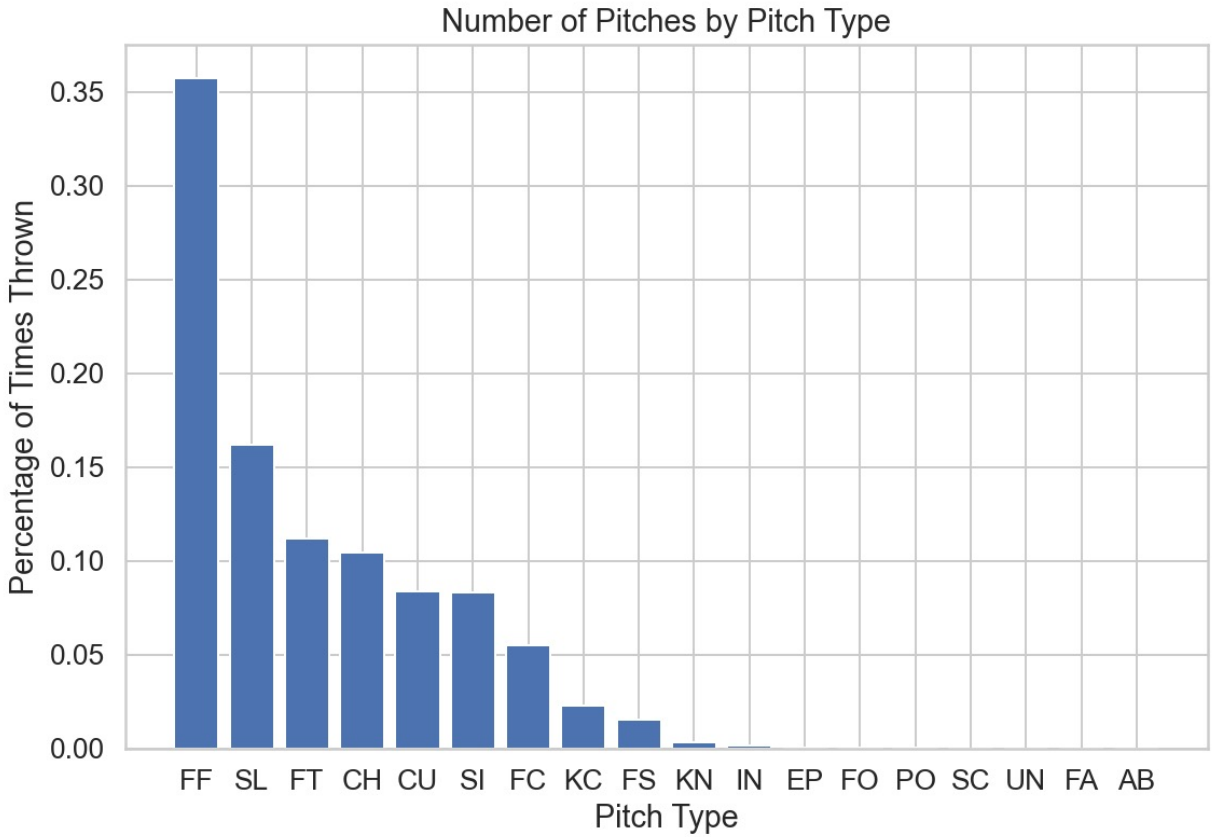
It takes 0.425 seconds for a 95 mile per hour pitch to reach home plate.

- Within that time batter needs to:
  - Identify pitch type
  - Identify location
  - Decide whether or not to swing
  - Swing
  - Hit the ball
- How is this possible?
  - Guessing
  - Sign stealing
  - Studying / Preparation

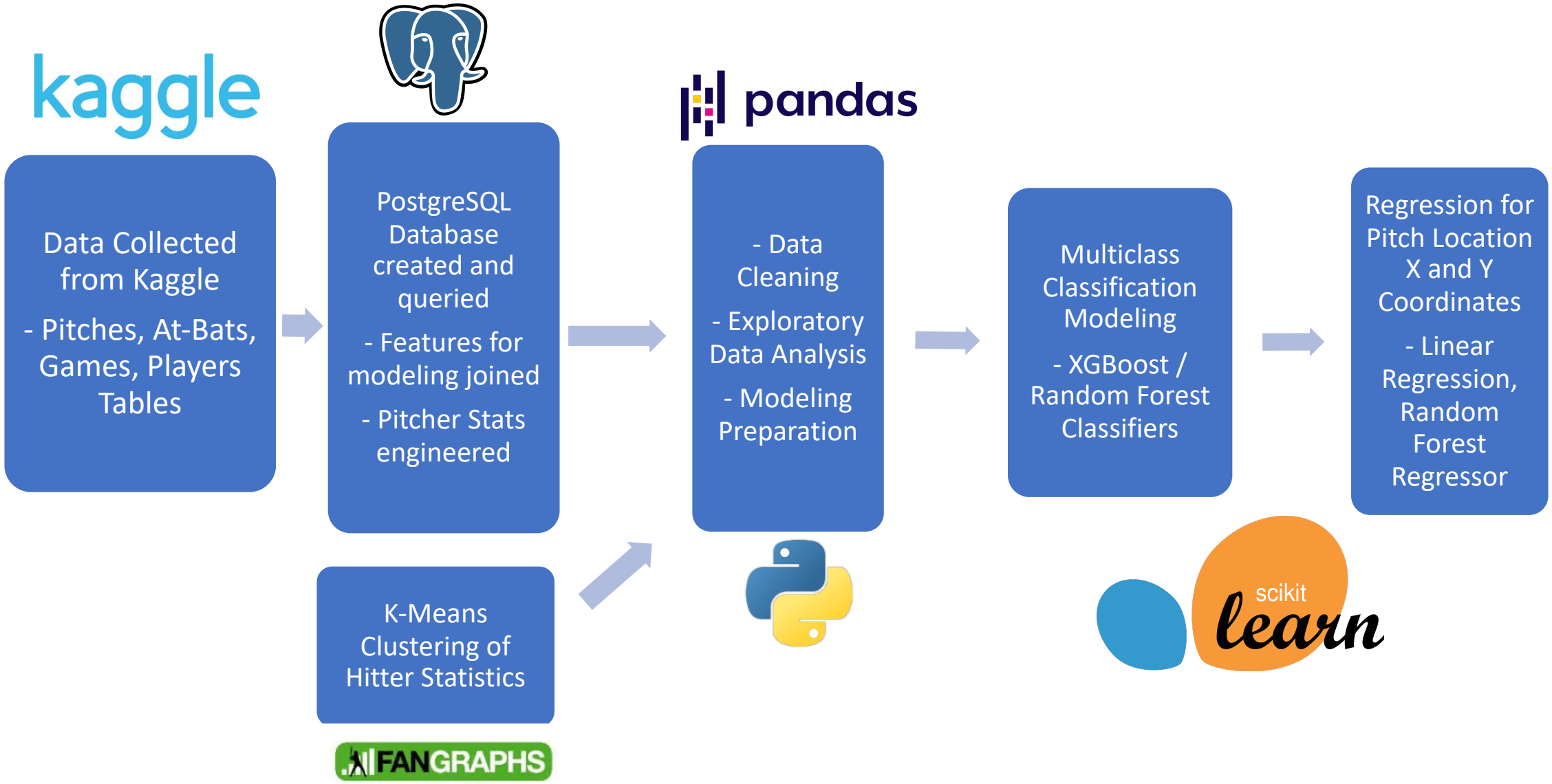
Can Machine Learning help?

Data from every MLB Pitch over 2015-2019 was utilized in this project

- ~3.5 million total pitches
- ~18 pitch types
- Kaggle Dataset from Paul Schale
- Supplementary Data:
  - Hitting Statistics from FanGraphs
- Project Goal: Develop two sequential predictive models:
  - Pitch Type – Classification
  - Pitch Location – Regression
- KEY: Features must be known by the batter before the pitch!



General Project Timeline and Tools



The final models included features from three “buckets”:

Game Situation:

Outs, Balls, Strikes, Runners on Base, Score

Batter Info:

Hitter “Cluster”, Hitting Side

Pitcher Stats:

Total pitches thrown, last pitch information, running pitch type breakdown

Individual final models were trained and validated for 50 pitchers.

- Pitch Type: XGBoost Classifier

METRIC	VALUE

Individual final models were trained and validated for 50 pitchers.

- Pitch Type: XGBoost Classifier

METRIC	VALUE
Avg. Pitch Type Prediction Accuracy	31%

Individual final models were trained and validated for 50 pitchers.

- Pitch Type: XGBoost Classifier

METRIC	VALUE
Avg. Pitch Type Prediction Accuracy	31%
Avg. Non-Primary Pitch Prediction Precision	23%
Avg. Non-Primary Pitch Prediction Recall	23%

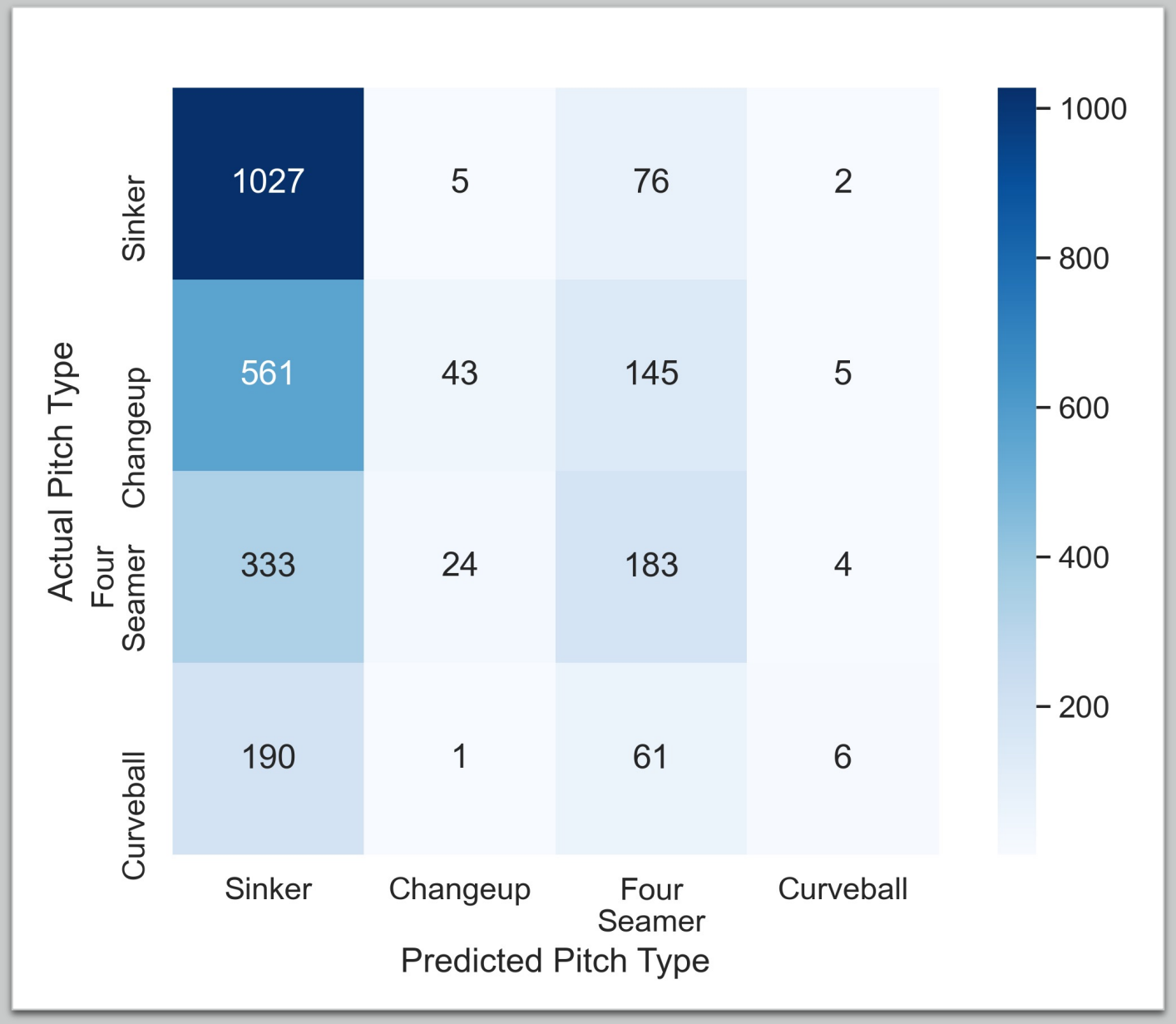
Kyle Hendricks' pitch arsenal was predicted with the highest precision and recall for non-primary pitches

Accuracy: 47% | Precision: 46%
Recall: 34%

Pitch Arsenal (w/ pitch rate):

- Sinker: 47%
- Changeup: 27%
- Four Seam Fastball: 13%
- Curveball: 8%
- Cutter: 5%

Establishes hitters with a sinker/changeup mix, doesn't rely on four seam velocity



The model predicted Carlos Carrasco's pitches well when considering the second most likely pitch

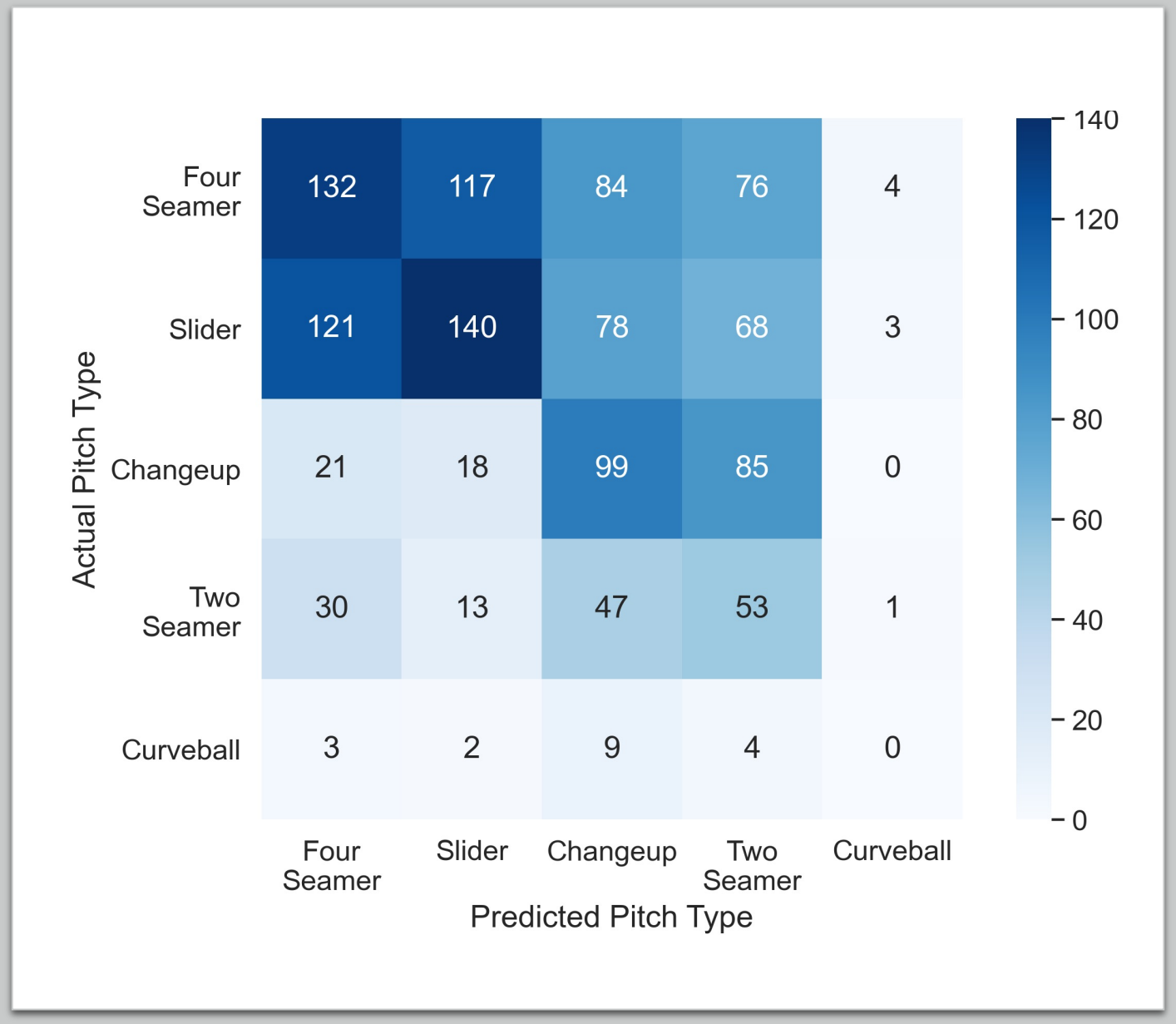
Accuracy: 35% | Precision: 28%

Recall: 29%

Pitch Arsenal (w/ pitch rate):

- Four Seam Fastball: 36%
- Slider: 21%
- Changeup: 18%
- Two Seamer: 14%
- Curveball: 11%

Second most likely pitch type matched the actual pitch type 47% of the time



The model struggled with pitchers who drastically changed pitch types, like Chris Sale

Accuracy: 23%

Precision: 32%

Recall: 28%

- Sinker Usage fell drastically in 2017
- Replaced with Four seam fastball

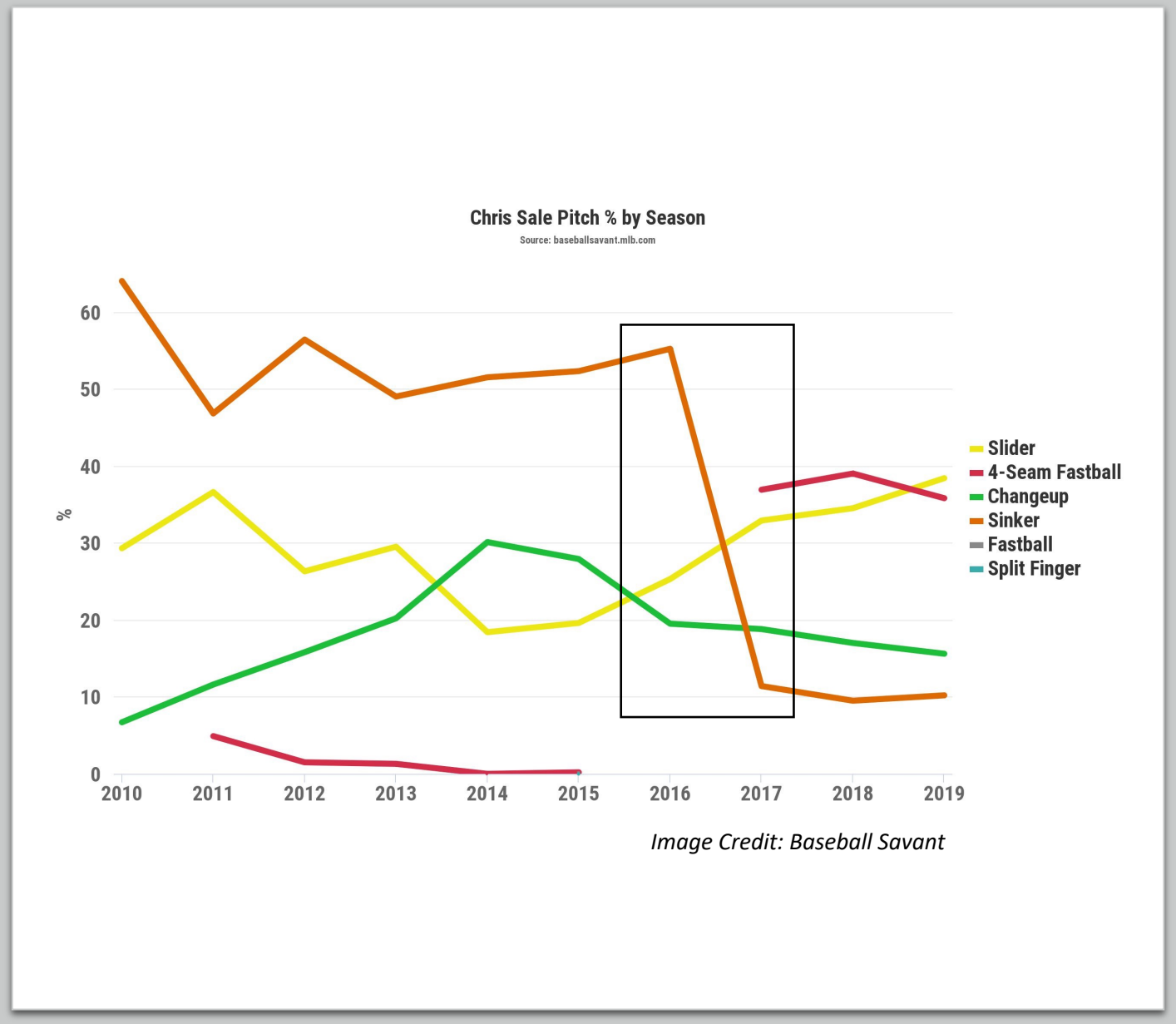


Image Credit: Baseball Savant

Individual final models were trained and validated for 50 pitchers.

- Pitch Location: Linear Regression
  - X Coordinate First
  - Y Coordinate Second

METRIC	VALUE

Individual final models were trained and validated for 50 pitchers.

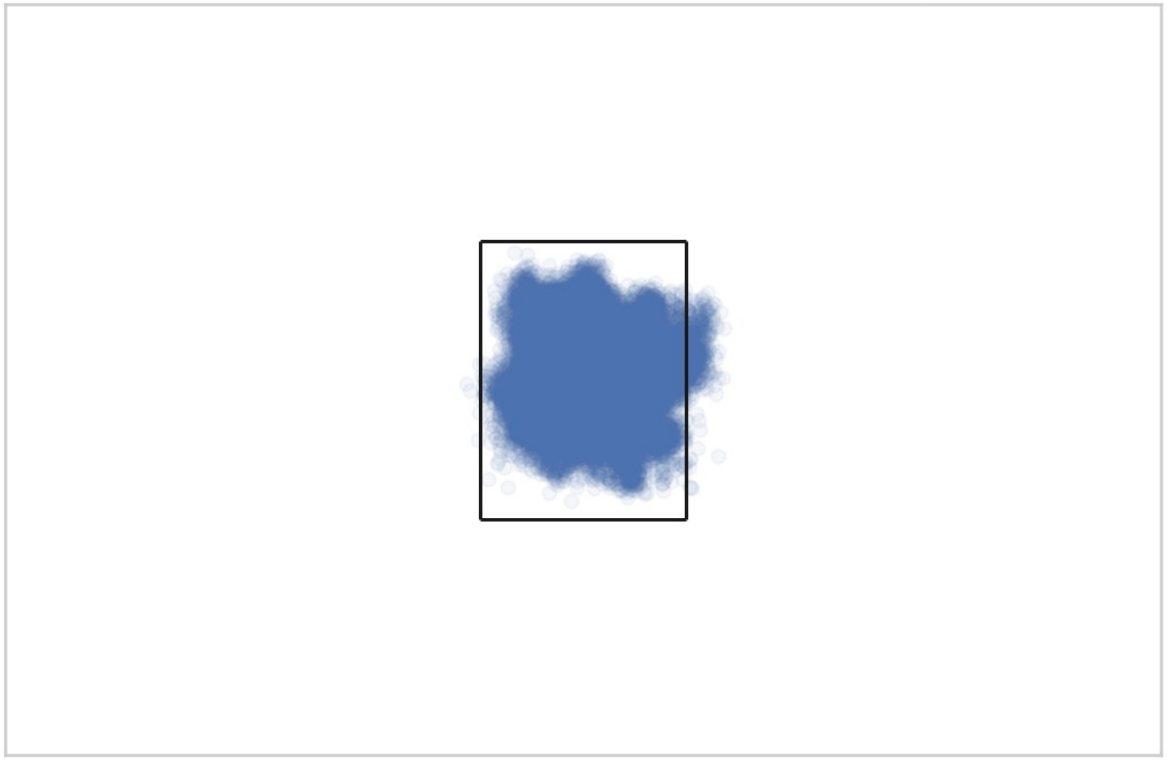
- Pitch Location: Linear Regression
  - X Coordinate First
  - Y Coordinate Second

METRIC	VALUE
Avg. X Coordinate Mean Average Error	7.76 in.
Avg. Y Coordinate Mean Average Error	8.25 in.

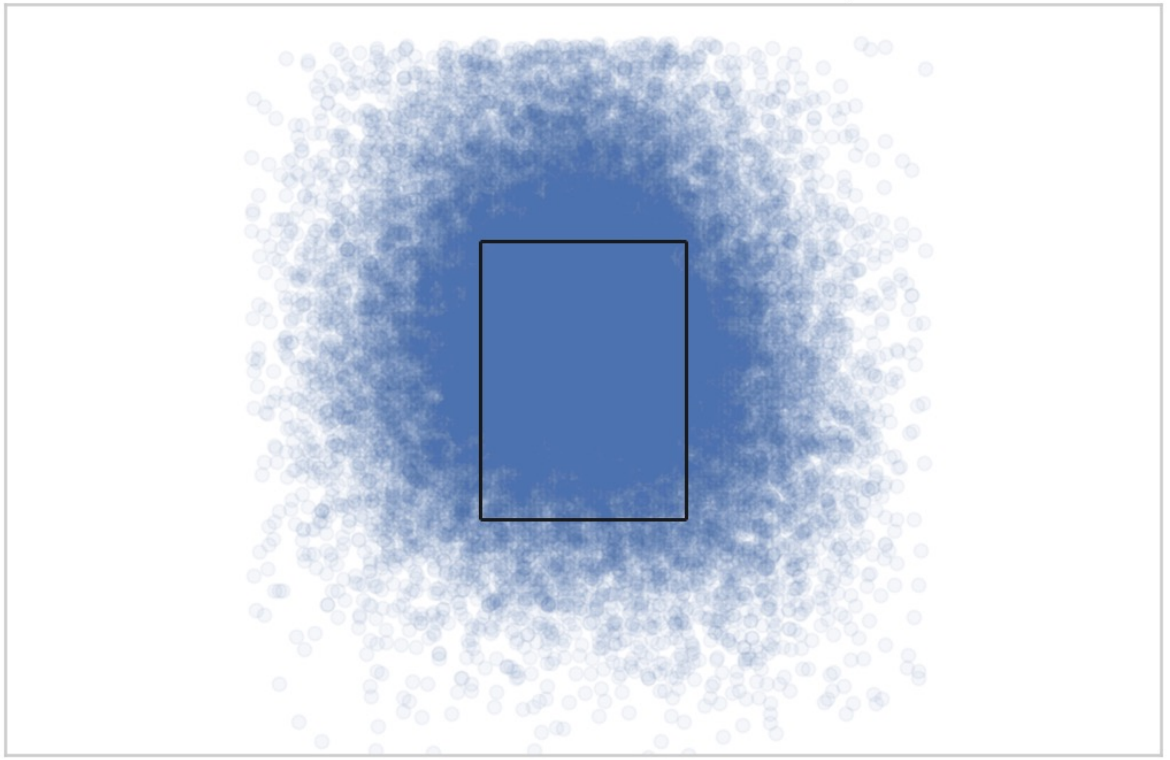
Predicted pitch locations tended to predict with more concentration within the strike zone, replicating intent of the pitch

Four Seam Fastball: Middle of the zone, upper portion

Model Predicted Pitch Locations - 2019 Test Data, FF



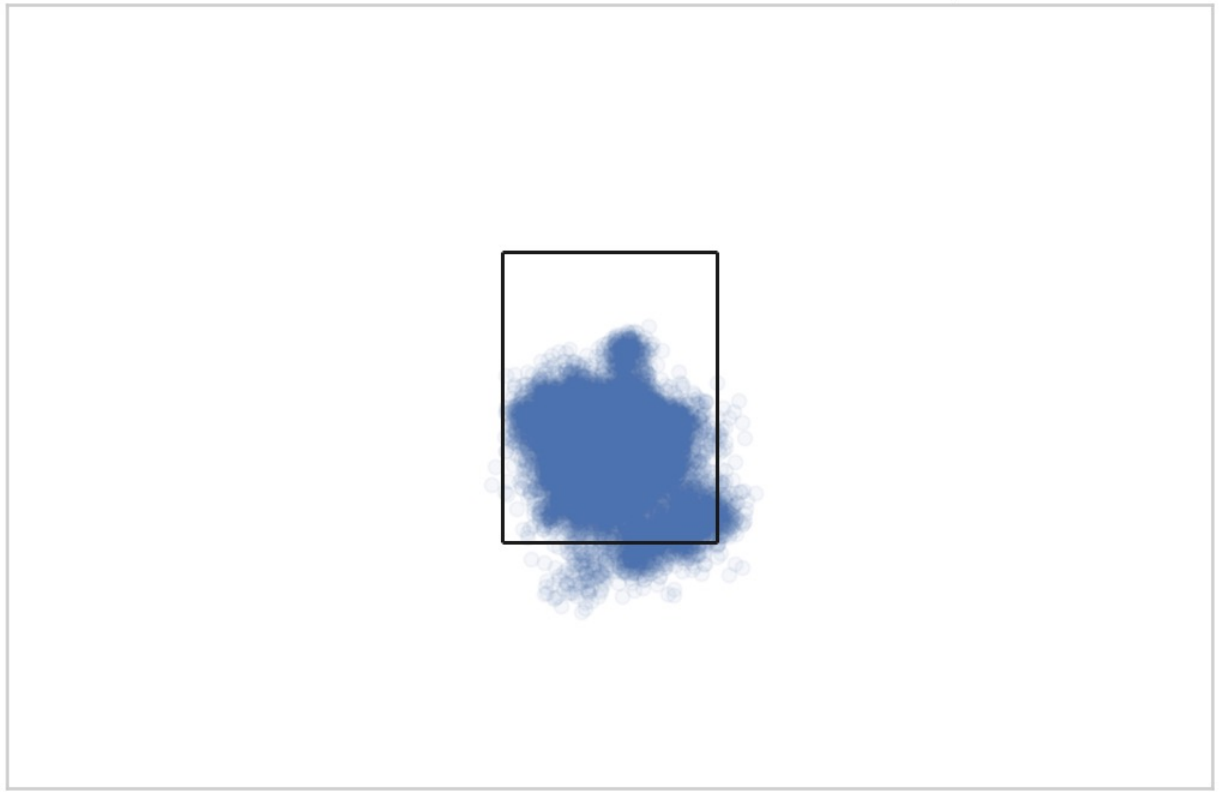
Actual Pitch Locations - 2019 Test Data, FF



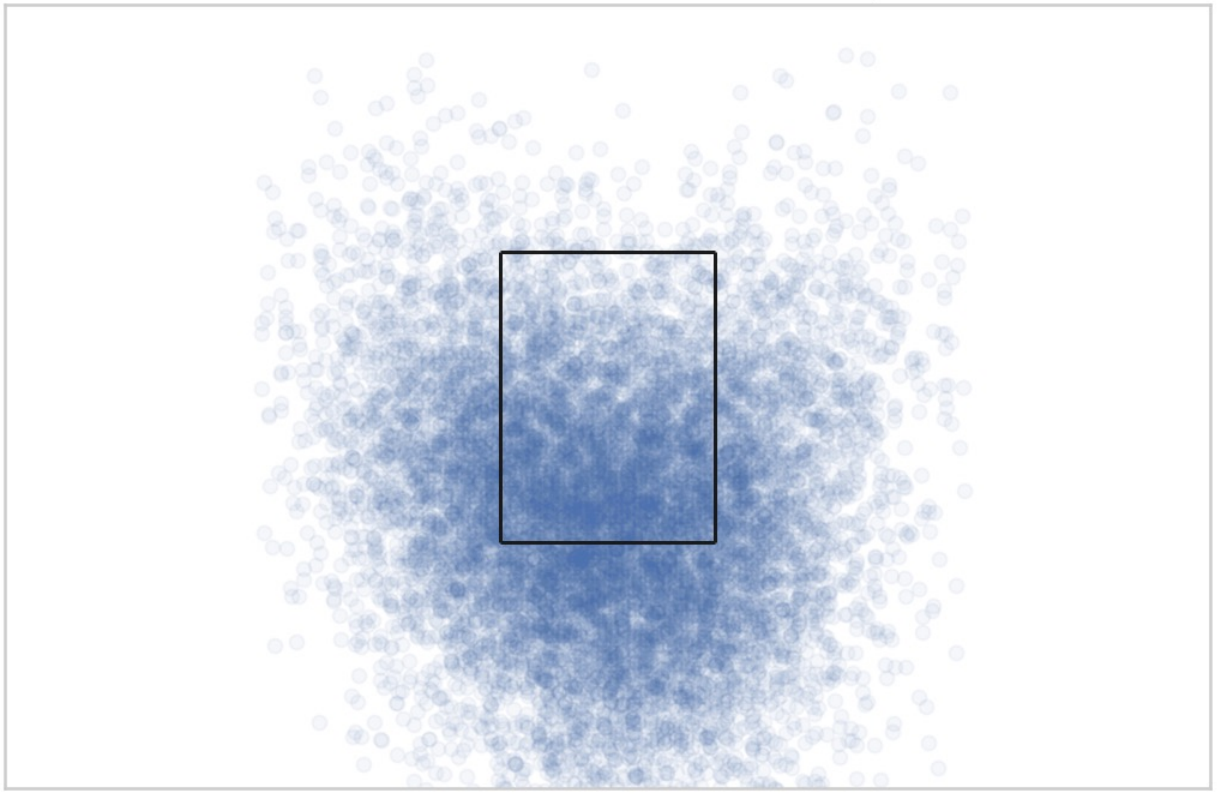
Predicted pitch locations tended to predict with more concentration within the strike zone, replicating intent of the pitch

Changeup: Lower in the zone, trending towards the right-side edge

Model Predicted Pitch Locations - 2019 Test Data, CH



Actual Pitch Locations - 2019 Test Data, CH



Future Improvements

- Web App Development
- Individual hyptertuning of pitcher models
  - Continued feature engineering to improve prediction

Thank you!

Feel free to connect on LinkedIn or check out my project on GitHub!

LinkedIn:
https://www.linkedin.com/in/patrick-bovard/

GitHub Repository:
https://github.com/pbovard63/Predicting_MLB_Pitches

APPENDIX

50 Pitchers included in final model:

- Max Scherzer, Justin Verlander', 'Chris Archer', 'Jose Quintana', 'Chris Sale', 'Rick Porcello', 'Jon Lester', 'Corey Kluber', 'Gio Gonzalez', 'Julio Teheran', 'Jake Arrieta', 'Zack Greinke', 'Cole Hamels', 'Trevor Bauer', 'Gerrit Cole', 'Jacob deGrom', 'Dallas Keuchel', 'Jake Odorizzi', 'James Shields', 'Kyle Gibson', 'Marco Estrada', 'J.A. Happ', 'Kevin Gausman', 'Tanner Roark', 'Mike Fiers', 'Ian Kennedy', 'Mike Leake', 'Kyle Hendricks', 'David Price', 'Carlos Martinez', 'Carlos Carrasco', 'Andrew Cashner', 'Jeff Samardzija', 'Madison Bumgarner', 'Jason Hammel', 'Masahiro Tanaka', 'CC Sabathia', 'Robbie Ray', 'Wade Miley', 'Clayton Kershaw', 'Danny Duffy', 'Bartolo Colon', 'Patrick Corbin', 'Sonny Gray', 'Chase Anderson', 'Johnny Cueto', 'Francisco Liriano', 'Hector Santiago', 'Jordan Zimmermann', 'Felix Hernandez'

Final Model Features:

Classification:

- 'Cluster','inning', 'top', 'on_1b', 'on_2b', 'on_3b', 'b_count', 's_count', 'outs', 'stand_R', 'pitcher_run_diff','last_pitch_speed', 'last_pitch_px', 'last_pitch_pz','pitch_num','cumulative_pitches', 'cumulative_ff_rate', 'cumulative_sl_rate', 'cumulative_ft_rate', 'cumulative_ch_rate', 'cumulative_cu_rate', 'cumulative_si_rate', 'cumulative_fc_rate', 'cumulative_kc_rate', 'cumulative_fs_rate', 'cumulative_kn_rate', 'cumulative_ep_rate', 'cumulative_fo_rate', 'cumulative_sc_rate', 'Last_Pitch_Type_Num', 'last_5_ff', 'last_5_sl', 'last_5_ft', 'last_5_ch', 'last_5_cu', 'last_5_si', 'last_5_fc', 'last_5_kc', 'last_5_fs', 'last_5_kn', 'last_5_ep', 'last_5_fo'

Regression – Px:

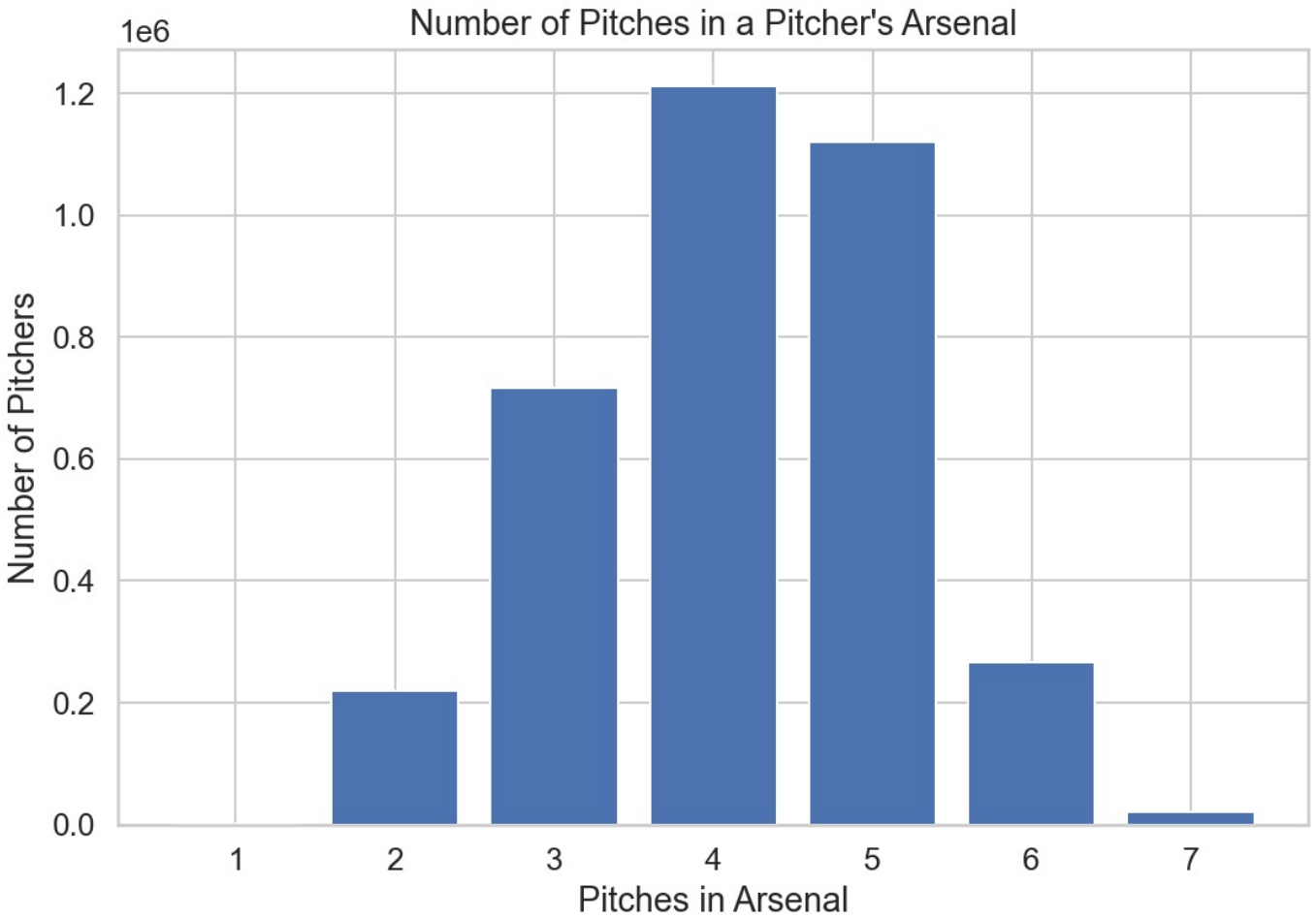
- 'Cluster','inning', 'top', 'on_1b', 'on_2b', 'on_3b', 'b_count', 's_count', 'outs', 'stand_R', 'pitcher_run_diff','last_pitch_speed', 'last_pitch_px', 'last_pitch_pz','pitch_num','cumulative_pitches', 'Last_Pitch_Type_Num', 'Pitch_Type_Num'

Regression – Pz:

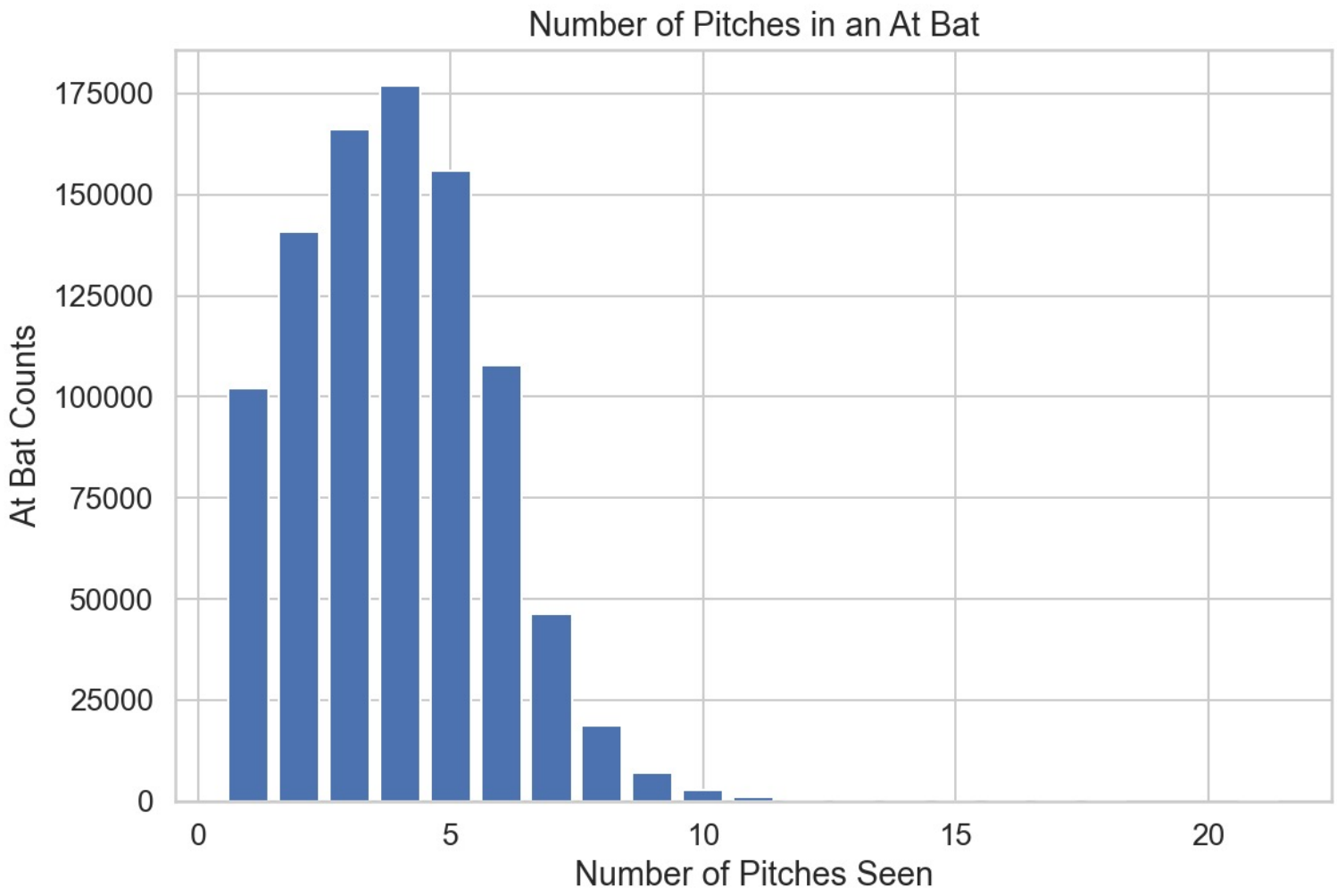
- 'Cluster','inning', 'top', 'on_1b', 'on_2b', 'on_3b', 'b_count', 's_count', 'outs', 'stand_R', 'pitcher_run_diff','last_pitch_speed', 'last_pitch_px', 'last_pitch_pz','pitch_num','cumulative_pitches', 'Last_Pitch_Type_Num', 'Pitch_Type_Num', 'px'

Different pitchers throw different numbers of pitches, complicating predictions

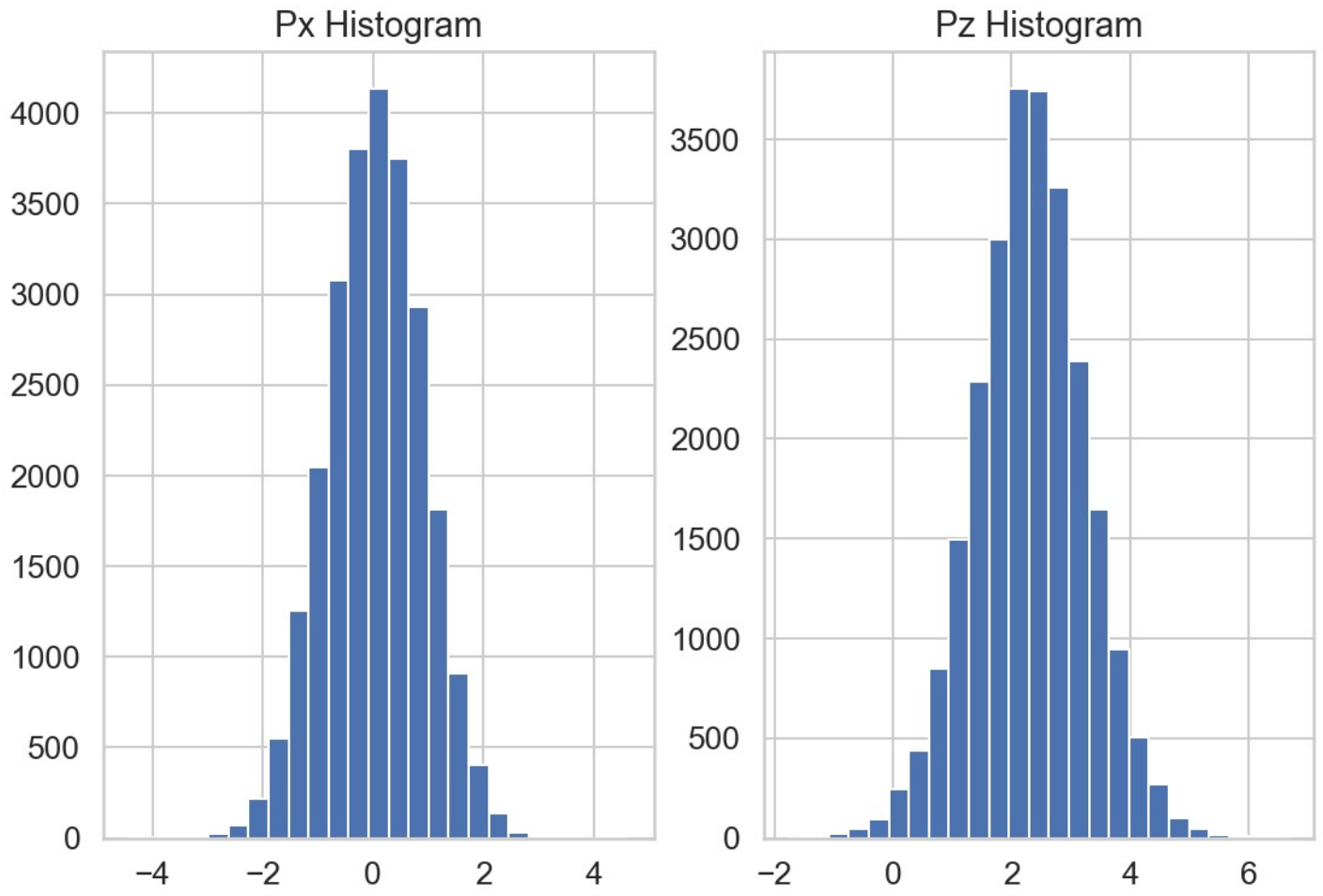
- Solution: Individual models for each different pitcher



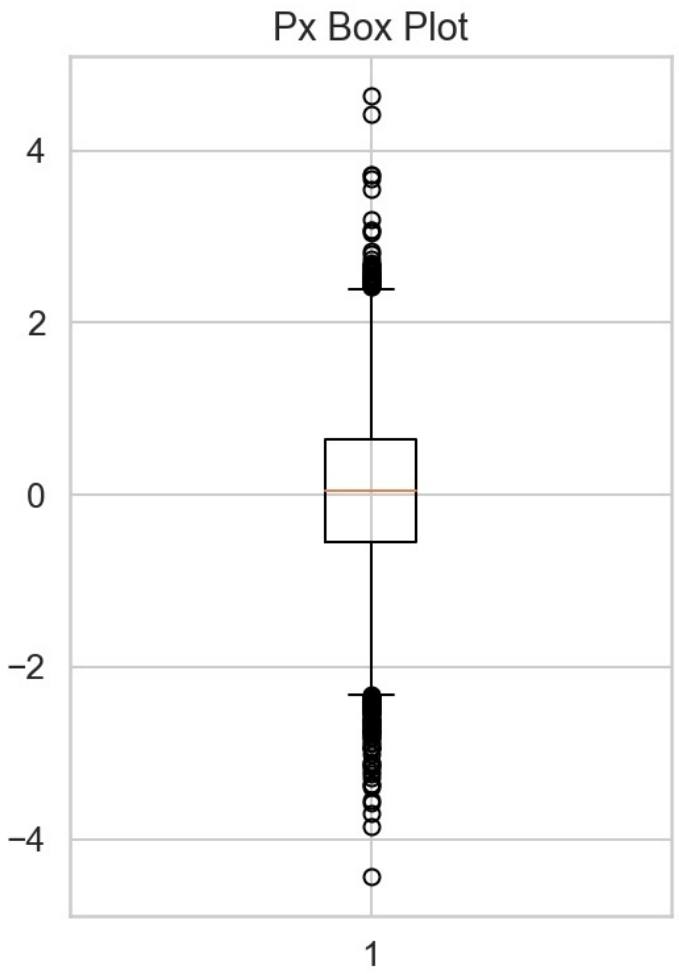
Pitches in an at bat have a definitive right tail.

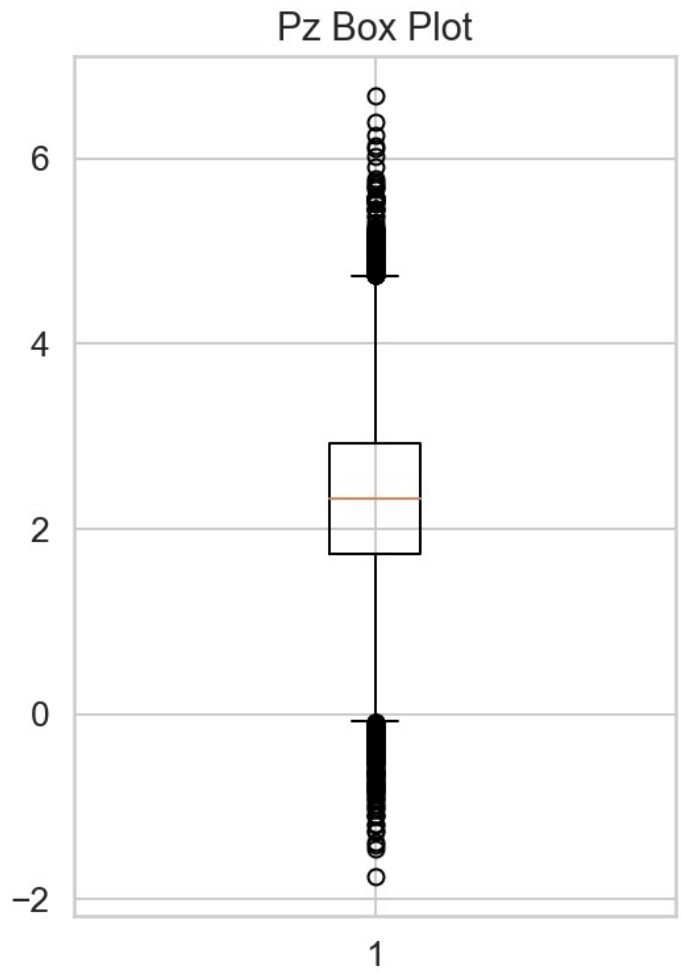


Pitch location shows a relatively normal distribution for the two coordinates.

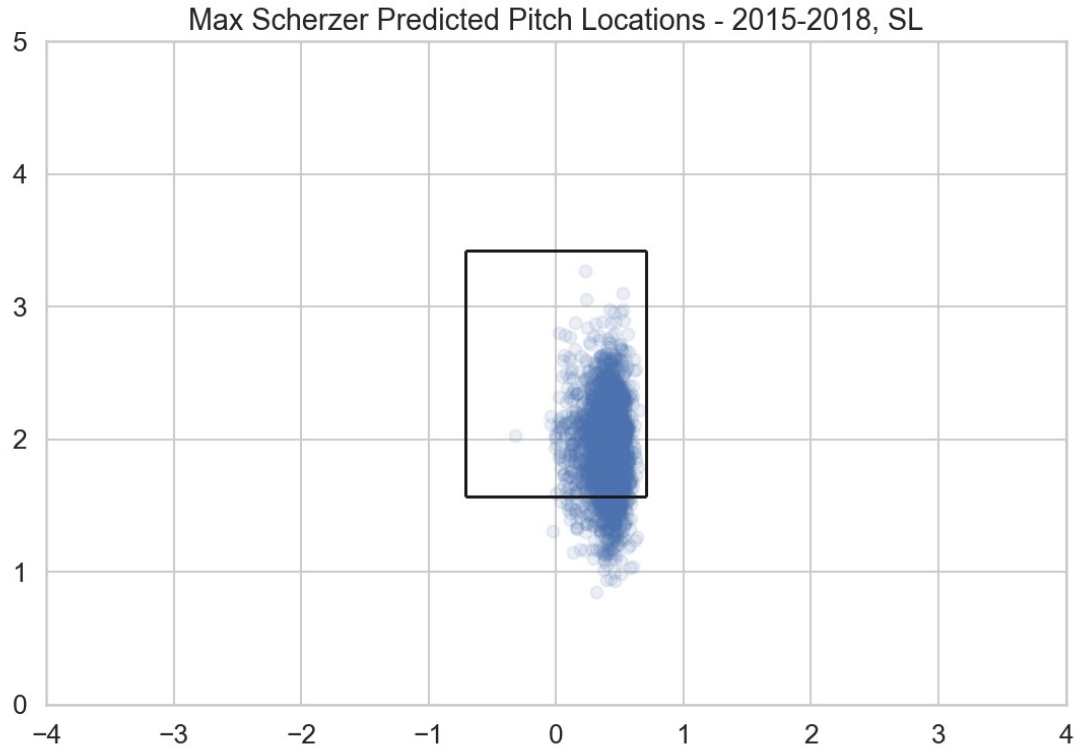


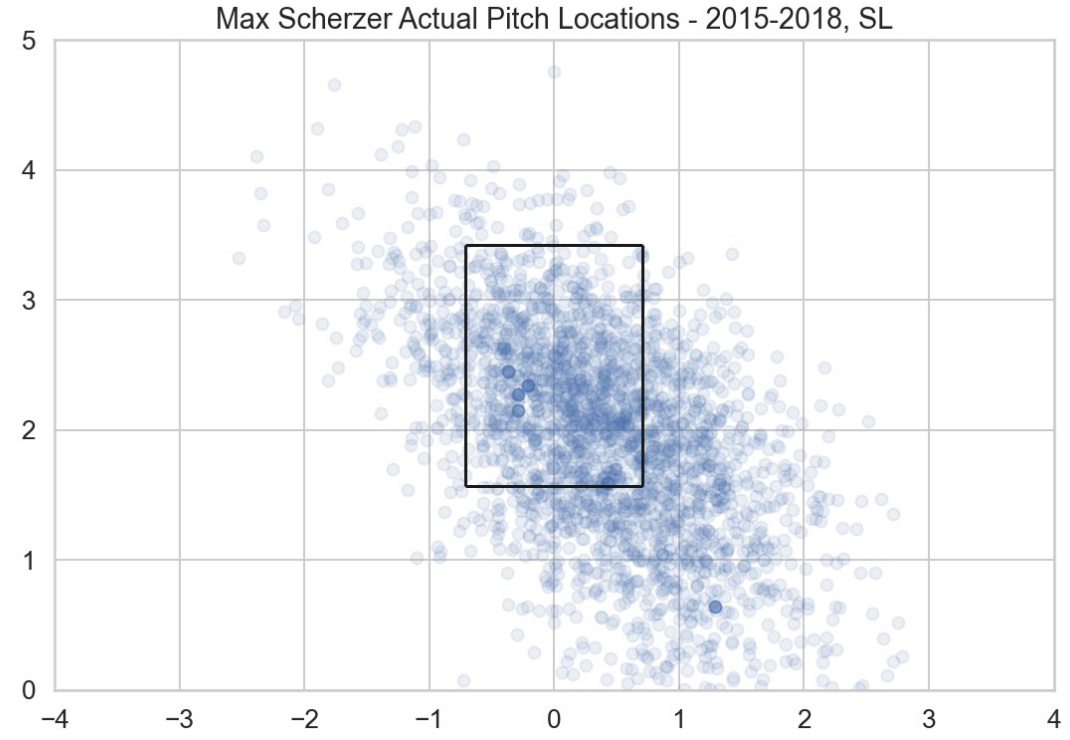
Both Px and Pz showed a high number of outlier points.





Pitch locations tend to predict in or near the strike zone, possibly showing how a pitcher is trying to throw a certain pitch type.





Pitch locations tend to predict in or near the strike zone, possibly showing how a pitcher is trying to throw a certain pitch type.