MLB Pitch Prediction

Elevator Pitch

- Create a model and eventually dashboard that will help the offensive team (the hitting team) in pitch anticipation during an at-bat.

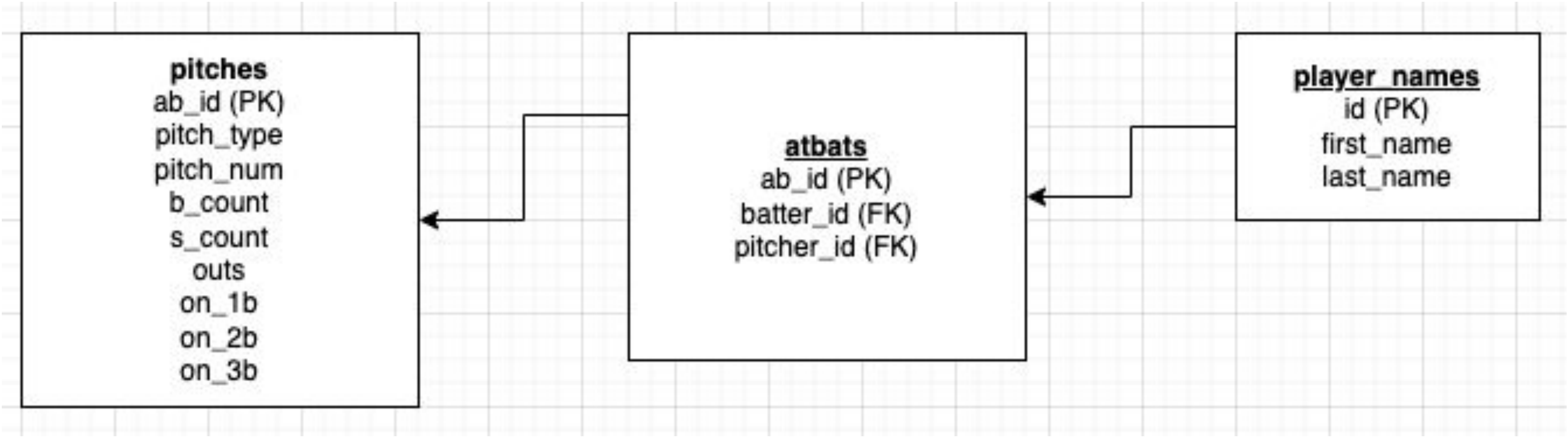
- Help predict pitch type based on batter, pitcher, runners on base, count during the at-bat, pitch sequence, etc.

- This dashboard would help teams predict what will be thrown therefore hopefully increasing the chances the batter will be successful.

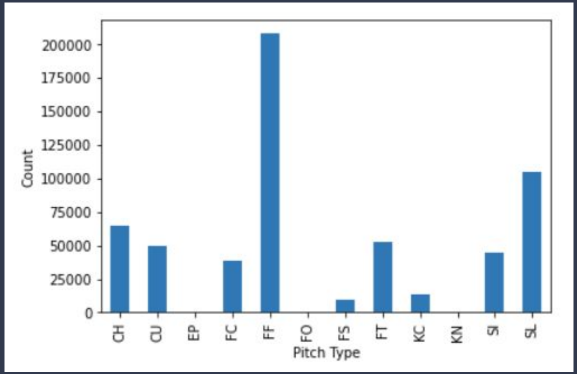
Data Summary

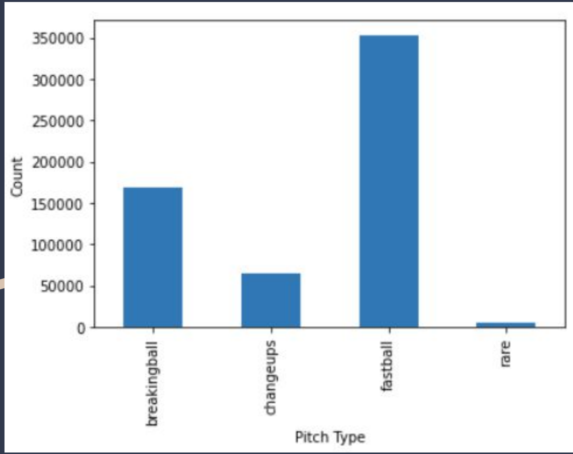
- Every pitch thrown during the Major League Baseball (MLB) 2015-2018 seasons
- 8 total datasets, 1.06GB
- Merged 3 datasets together
  - Pitch data
  - At-Bat data
  - Player data
- Merged dataset
  - ~2,150,000 rows
  - 54 columns
  - Too big for GitHub - revised down to 14 columns

Data Architecture



Data Cleaning



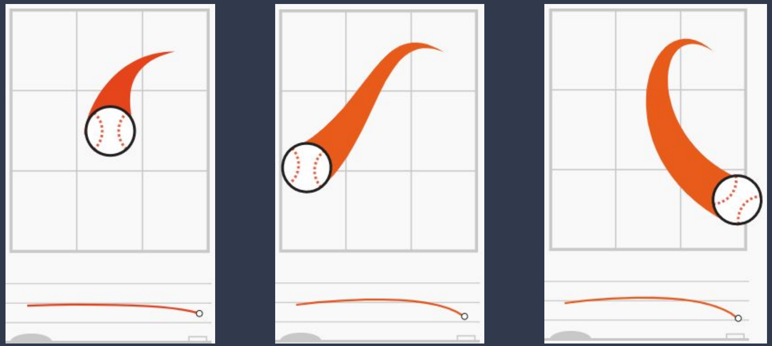


- Pitch count often not changing after a pitch is thrown - will be difficult to include this important feature in our model if it is unreliable

- Many other columns have questionable data but we will likely use filler values such as mean or mode

- Solution: Use parts of the dataset that don’t contain this issue (2019)

Generalizing Pitch Type

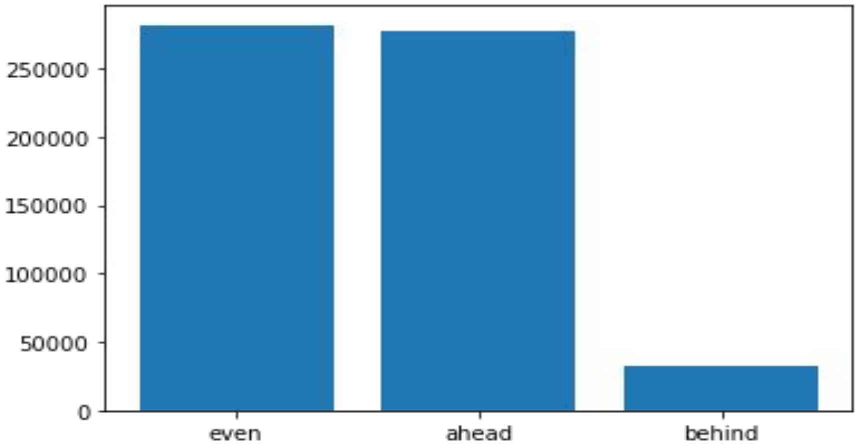


- Fastballs
  - Four seam fastball
  - Two Seam fastball
  - Sinker
  - Splitter
  - Cutter
- Breaking Balls
  - Curve ball
  - Slider
  - Screwball
  - Knuckle curve
  - Knuckle ball
- Changeups
  - Changeup
- Rare Pitches
  - Eephus
  - Pitchout
  - Unidentified
  - Intentional ball

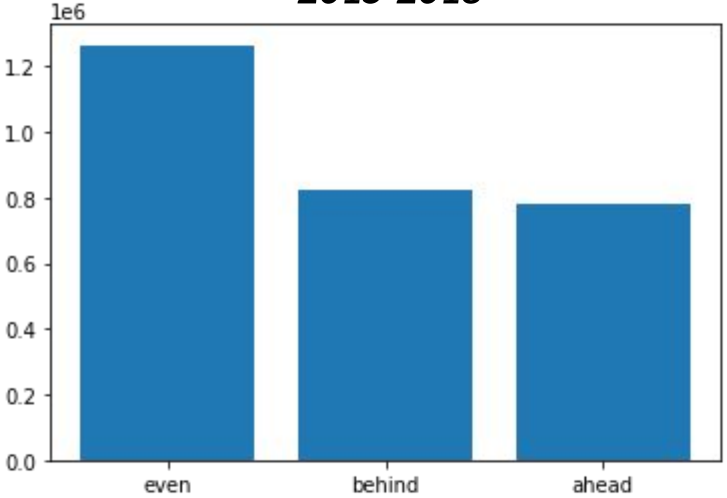
Cleaner data

- Many of the features in the 2019 dataset did not pass sanity check. Larger 2015-2018 ***did*** make sense

2019

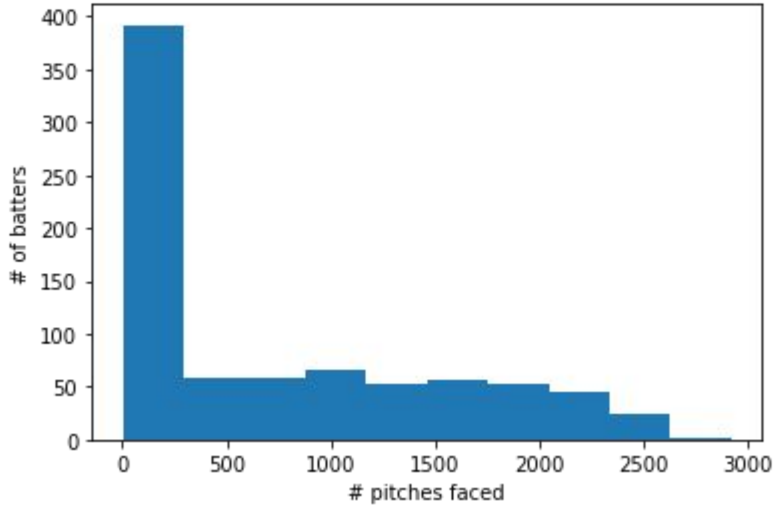


2015-2018



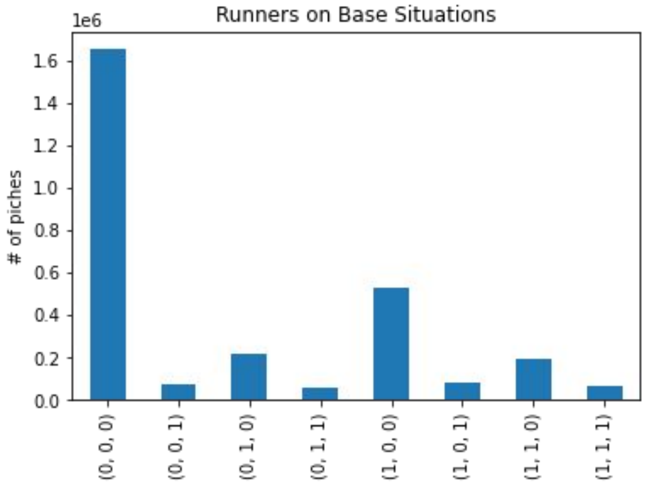
Feature Engineering

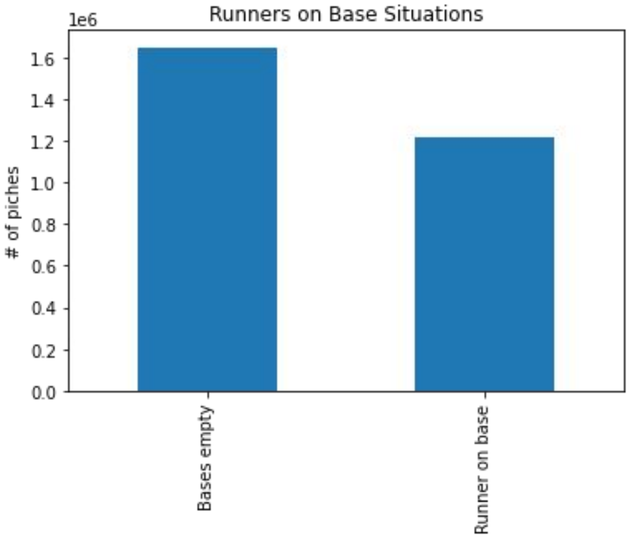
- Total the number of pitches a batter has faced
  - Eliminate batters below the median threshold of pitches faced
- Create a generalized strike/ball count feature
  - Individual counts are too specific and don't generate enough data (bad data)



Feature Engineering

- Balance Runners on Base Situations
  - We classified them into two categories, which satisfy relatively balanced and has distinct characteristics

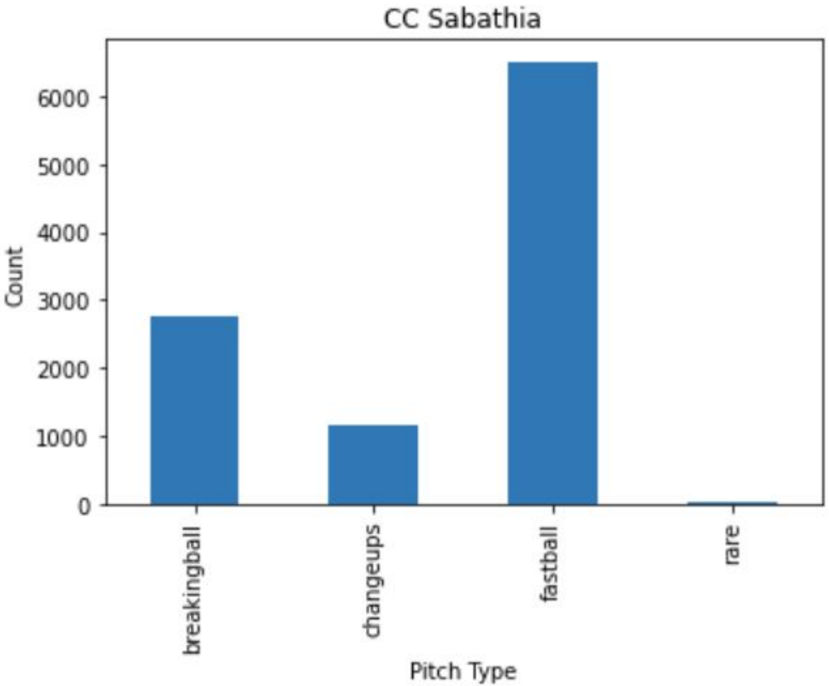


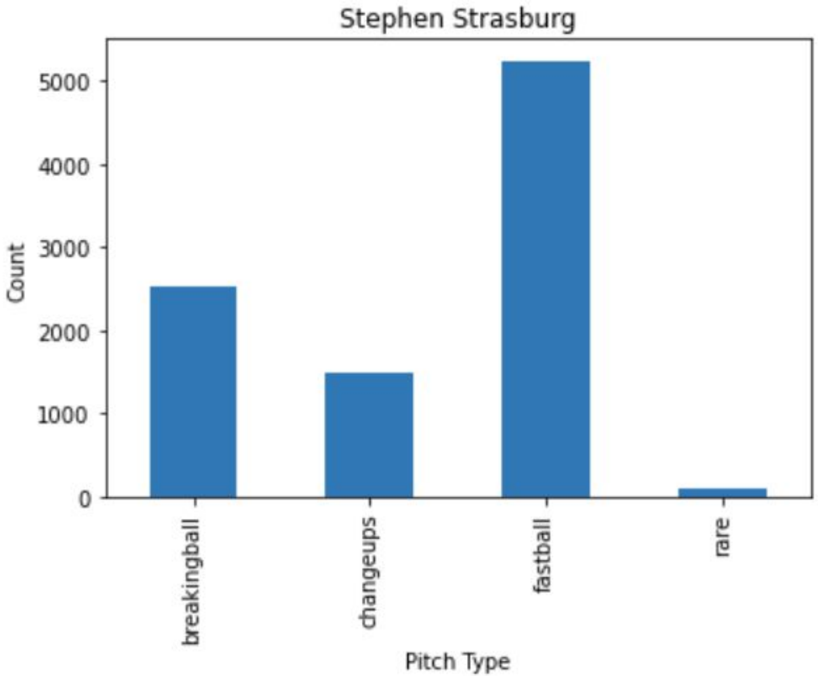


Building Basic Pitcher Tendencies

- Group by pitcher ID, situation, and pitch
- Group by pitcher ID and situation
- Using the cleaned pitch type feature
- Aggregating and finding frequency of balls thrown
- Pivot

Pitcher Tendencies





Building Basic Batter Knowledge

- Group by batter ID, situation, and pitch
- Group by batter ID and situation
- Aggregating and finding frequency of balls thrown ***to batter***
- Pivot

Summary: Data Cleaning and Feature Engineering

- 54 columns → 14 input columns
- Summarized columns
    - Pitch type
        - Fastball, breaking ball, change-up, rare
    - Runners on base
        - Empty, 1st base, RISP
    - At-bat count
        - Even, behind, ahead
    - Pitcher Situational Tendencies
        - 4 pitch type columns
        - Likelihood pitches types will be thrown (unknown pitch)
    - Batter Situational Experiences
        - 4 pitch type columns
        - Situational history of pitches seen

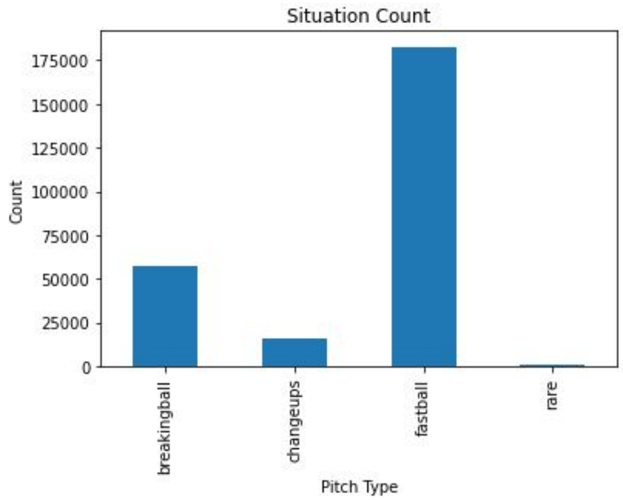
Baseline Model Probabilities

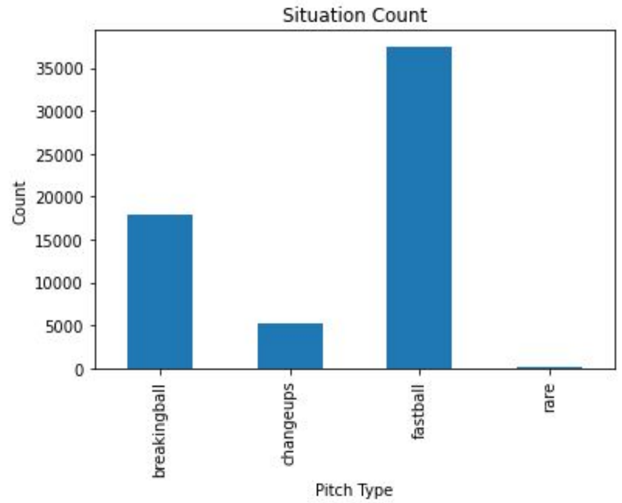
Situation: 1st Pitch, Even Count, 0 Outs

- Fastball - 71.13%
- Breaking Ball - 22.31%
- Changeup - 6.11%
- Rare - .43%

Situation: Breaking Ball, Ahead, 1 Out

- Fastball - 61.69%
- Breaking Ball - 29.53%
- Changeup - 8.57%
- Rare - .19%





Modeling

- Baseline model
  - Summary probabilities
  - Guess majority class

```
Enter the previous pitch: 1st_pitch
Enter the batter count:even
Enter the current outs: 1

prev_pitch_type     1st_pitch
batter_count             even
outs                      1.0
pitch_type           fastball
pitch_likelihood     0.656082
```

- Choices of models
  - SVM
  - Decision Trees
    - Gradient boosted classifier

Modeling

- Baseline model
  - Predict fastball
  - Most thrown pitch (61.9%)
    - Common strategy for hitters

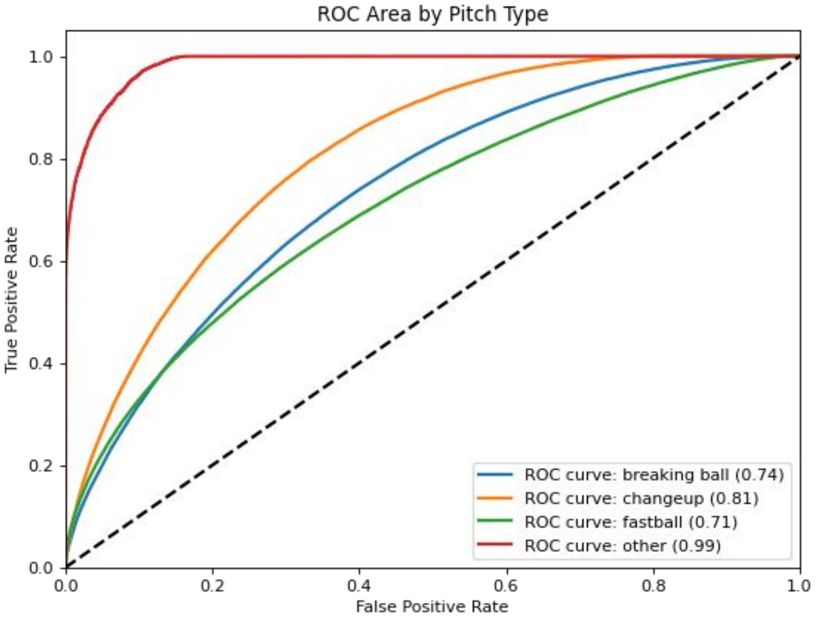
- Choices of models
  - SVM
  - Decision Trees
    - Random Forest
    - Gradient boosted classifier

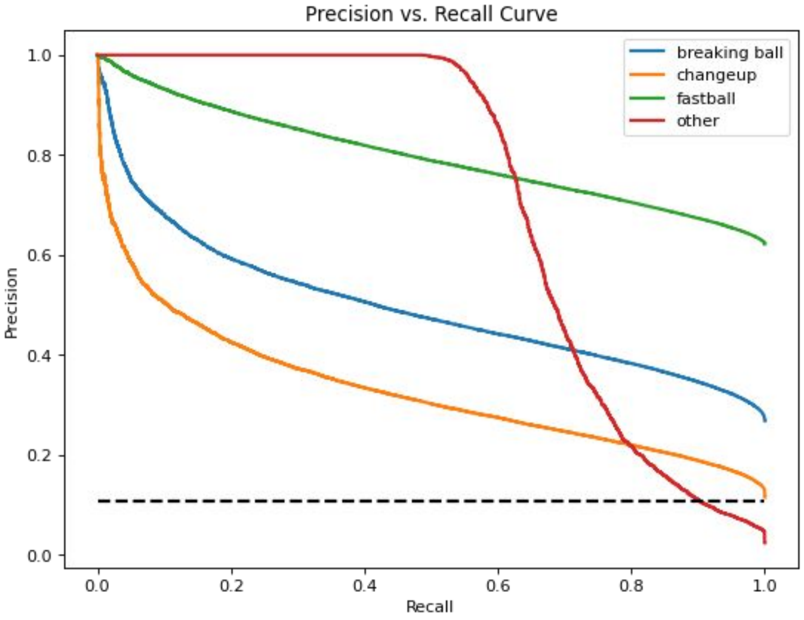
Results

	Baseline	SVM	XGBoost (Player Tendencies)	XGBoost (Player ID)
Accuracy	61.9%	63.7%	65%	65%
Precision	NA	89.8%	62%	62%
Recall	NA	68.61%	65%	65%

Results

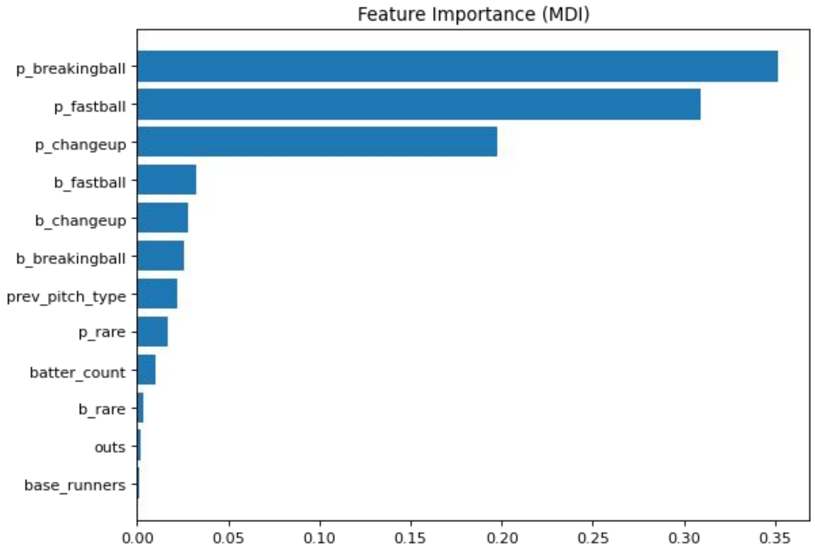
XGBoost (Player Tendencies)





Variable Importance

XGBoost (Player Tendencies)



Predicted

True

	breaking ball	changeup	fastball	other
breaking ball	28,782	1,039	84,740	47
changeup	3,981	3,825	37,897	18
fastball	18,943	2,491	244,801	11
other	188	21	1,227	1,929

Results

XGBoost (Player ID)

	precision	recall	f1-score	support
0	0.58	0.12	0.20	251833
1	0.49	0.01	0.02	96343
2	0.64	0.97	0.77	590459
3	0.85	0.56	0.68	7526
accuracy			$\boxed{0.64}$	946161
macro avg	0.64	0.41	0.42	946161
weighted avg	0.61	0.64	0.54	946161

Predicted

True

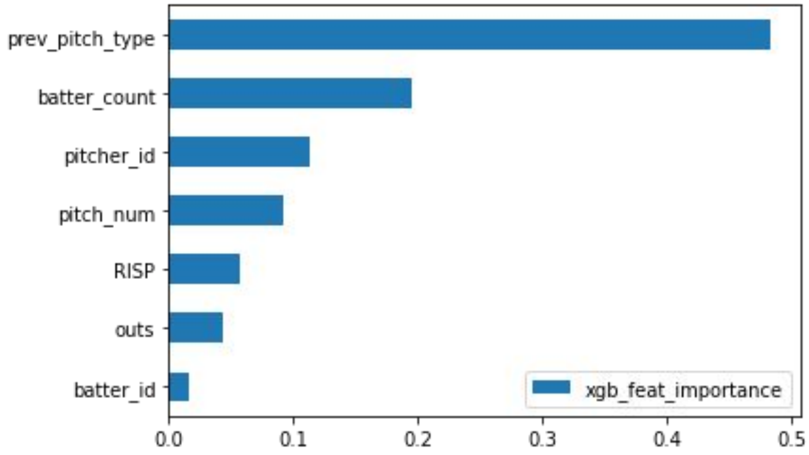
	breaking ball	change up	fastball	other
breaking ball	30226	176	221007	200
change up	3147	872	92729	70
fastball	18496	653	570704	462
other	193	4	3051	4171

Random Forest (Player ID)

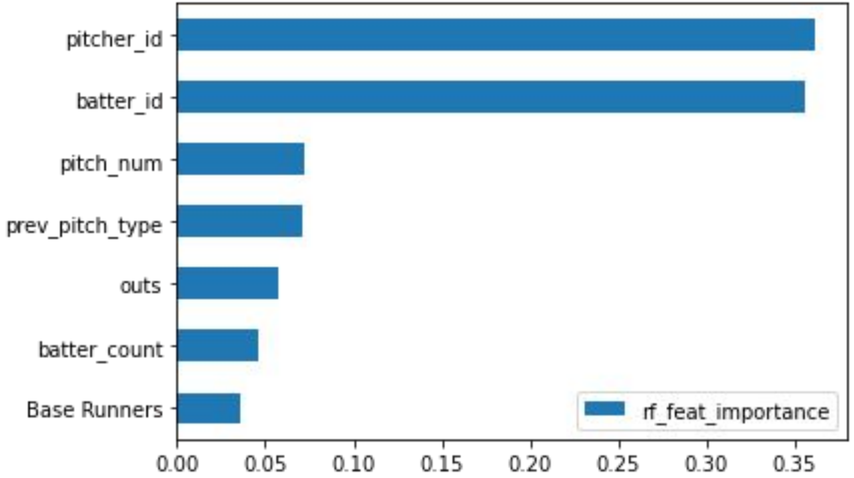
	precision	recall	f1-score	support
0	0.27	0.06	0.09	250763
1	0.12	0.00	0.00	96426
2	0.63	0.94	0.75	591404
3	0.01	0.00	0.01	7568
accuracy			$\boxed{0.60}$	946161
macro avg	0.25	0.25	0.21	946161
weighted avg	0.47	0.60	0.49	946161

Variable Importance

XGBoost (Player ID)



Random Forest (Player ID)



Model Valuation

- A run is estimated around $120,000
- 162 games a year
- Every two hits averages a single run
- 146 pitches a game

146 X 61.9% = 90

146 X 65% = 95

5 more pitches a game = 95-90

(Hits Yielded From Additional Information) X .5 X $120,000 X 162

Hit Increase Per Game	.5	1	2	3	4	5
Yearly Monetary value	$4,860,000	$9,720,000	$19,440,000	$29,160,000	$38,880,000	$48,600,000

Dashboard Demo

Clayton Kershaw
Pitcher
(477132)

Buster Posey
Catcher
(457763)

Conclusions / Limitations

- Outperformed baseline by ~3%
- Generalized features
  - Pitch type
  - Runners on base
- Removed batters/pitchers with limited data

- **Next Steps**
  - Add features
    - Score
    - Inning
  - Connect dashboard to database

Citations

https://community.fangraphs.com/what-is-a-run-worth/