# Machine learning driven baseball pitch analysis

Presented by:
Noah Bloss
Chayton Bouwmeester
Julius Foverskov
Frederik Hansen
Simon Johannsen

UNIVERSITY OF COPENHAGEN

# Outline

- Data & Goals

- Clustering of pitch types

- Classification of pitch types

- Classification of outcomes

- Prediction and evaluation

- Conclusion

- The future

# Goals for the project

- Clustering of pitch types

- Classifying pitch types

- Predicting the type of pitch

- Predicting the outcome of a pitch

- Create a rating for each individual pitch

- Be able to evaluate pitchers based on their pitches

# Baseball

- ## What is a Pitch?
    - Different ways of throwing the ball
    - Fastball, curveball etc.
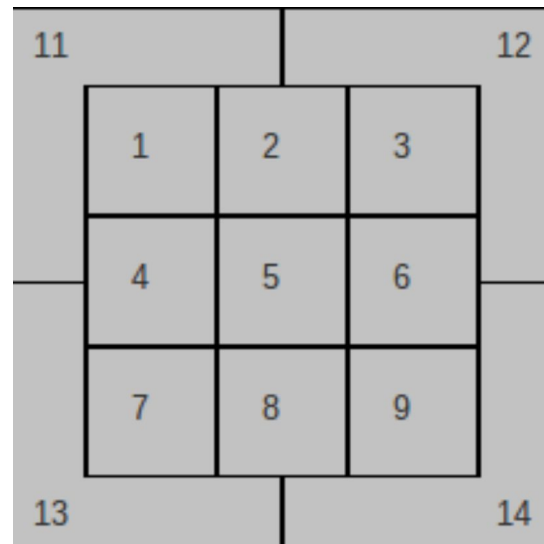




Credits: BeyondTheBox

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| FF  | FS  | SI  | KC  | CH  | FC  | SL  | ST  | CU  | PO  | CS  | FA  | SV  | KN  | NaN | EP  | SC  |
| 0   | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  | 13  | 14  | 15  | 16  |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
```
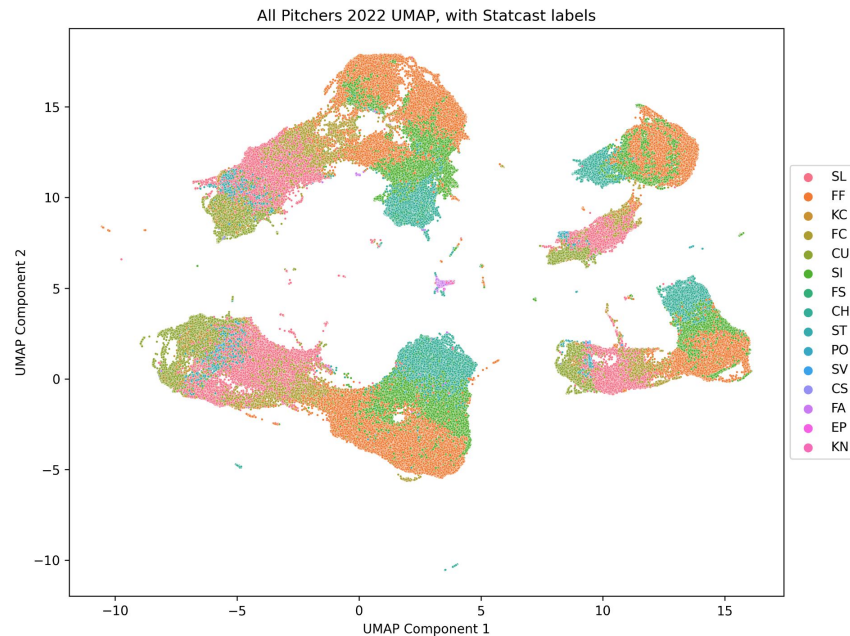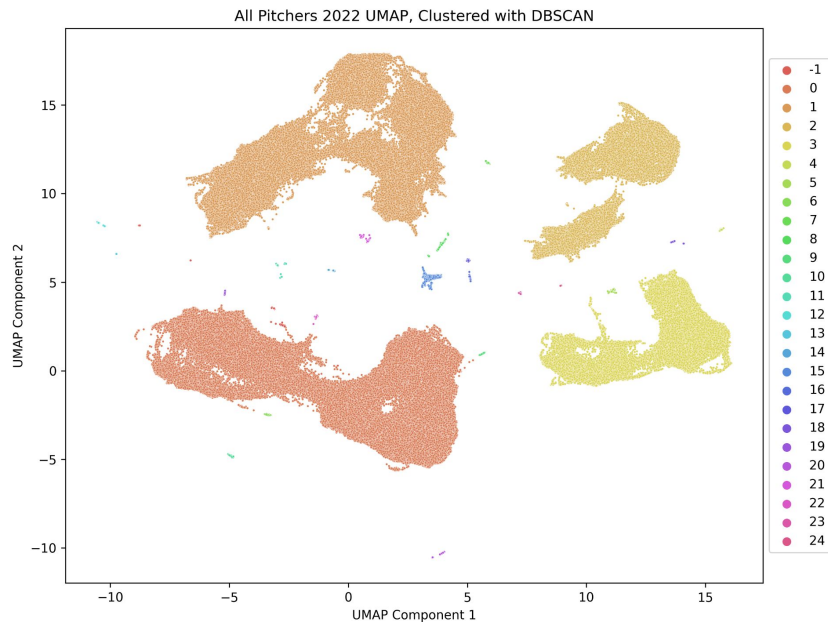
# The Data

- MLB pitch-data from Statcast

  - 2017-2022

  - 3,987,305 pitches in total

- 93 features to pick from

  - Lots of 'useless' information

- What is important for the pitch?

  - Everything about the ball from the point of release

- Converting string values to integers

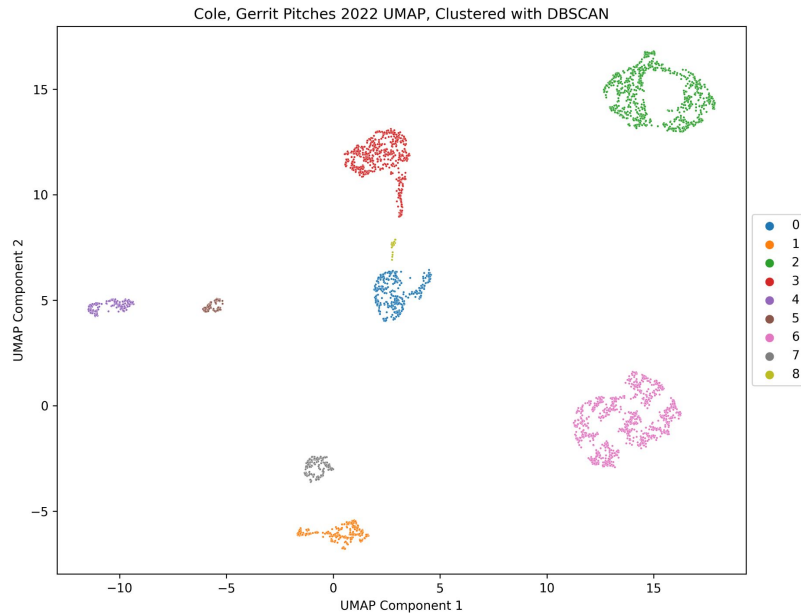| | pitch_type | release_speed | release_pos_x | release_pos_z | zone | plate_x | plate_z | vx0 | vy0 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 94.7 | -1.81 | 5.49 | 12.0 | 0.12 | 4.16 | 6.768865 | -137.755566 |
| 1 | 0 | 94.7 | -1.77 | 5.55 | 11.0 | -0.16 | 3.55 | 5.838900 | -137.772022 |
| 2 | 1 | 84.0 | -1.58 | 5.95 | 1.0 | -0.55 | 3.23 | 3.701940 | -122.258047 |
| 3 | 1 | 85.7 | -1.56 | 5.83 | 14.0 | 0.84 | 1.33 | 7.050352 | -124.700065 |
| 4 | 2 | 94.2 | -1.97 | 5.74 | 11.0 | -1.22 | 3.00 | 4.856891 | -137.187852 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 651619 | 0 | 96.1 | -3.28 | 5.13 | 5.0 | -0.19 | 2.41 | 9.899700 | -139.626080 |
| 651620 | 8 | 79.7 | -3.36 | 5.48 | 7.0 | -0.41 | 1.85 | 4.691537 | -116.064292 |
| 651621 | 0 | 96.7 | -3.21 | 5.33 | 9.0 | 0.48 | 2.07 | 11.547336 | -140.289508 |
| 651622 | 0 | 96.1 | -3.38 | 5.33 | 11.0 | -1.17 | 2.75 | 7.600636 | -139.767747 |
| 651623 | 0 | 95.3 | -3.28 | 5.47 | 11.0 | -1.16 | 3.06 | 7.498831 | -138.723511 |

# Clustering of different pitch types

- Pitch types are highly individualised
- See some clusters, but what are they?



All Pitchers 2022 UMAP, Clustered with DBSCAN



All Pitchers 2022 UMAP, with Statcast labels

# Umap clustering for 1 pitcher (Gerrit Cole)

- Finding new pitch type clusters compared to statcast labels using DBSCAN



Cole, Gerrit Pitches 2022 UMAP, Clustered with DBSCAN



Cole, Gerrit Pitches 2022 UMAP with Statcast Labels
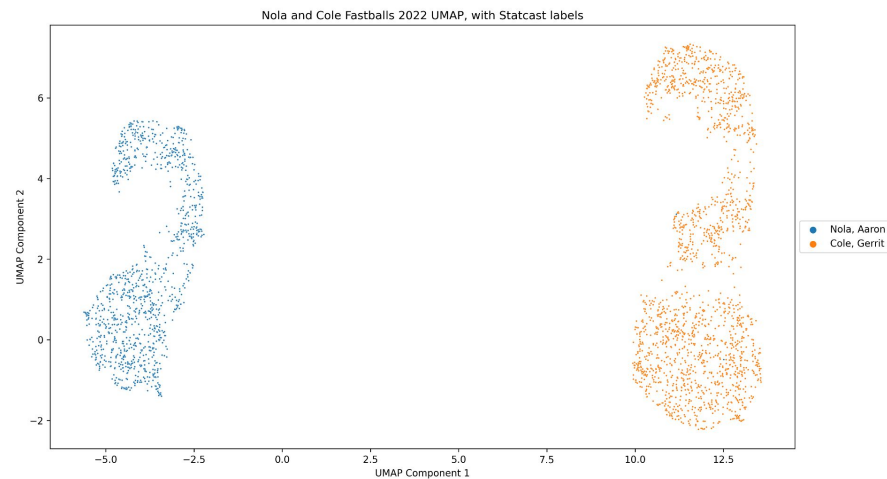
# Umap clustering for 1 pitcher (Aaron Nolan)

- Seems to work for all players
- Can't distinguish between some pitch types



Nola, Aaron Pitches 2022 UMAP, Clustered with DBSCAN

Nola, Aaron Pitches 2022 UMAP with Statcast Labels
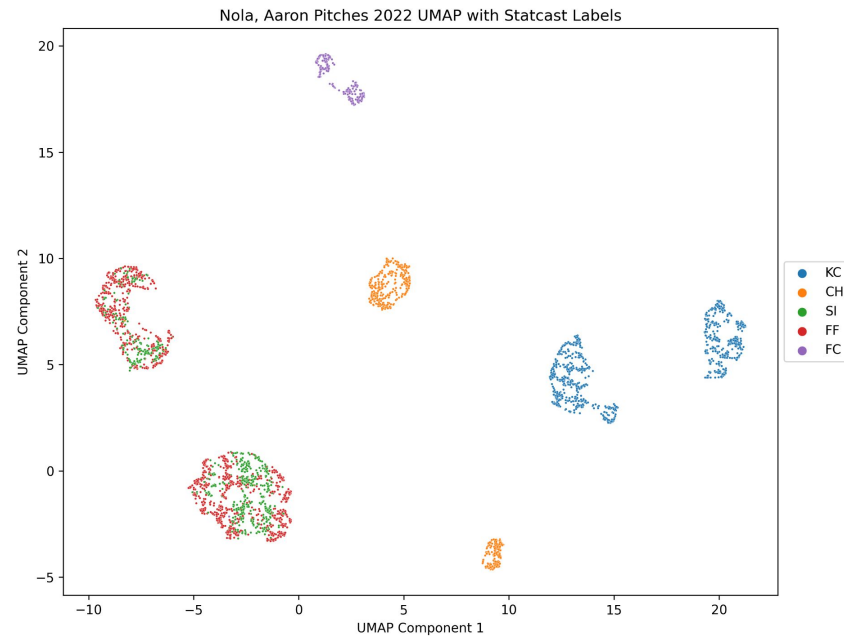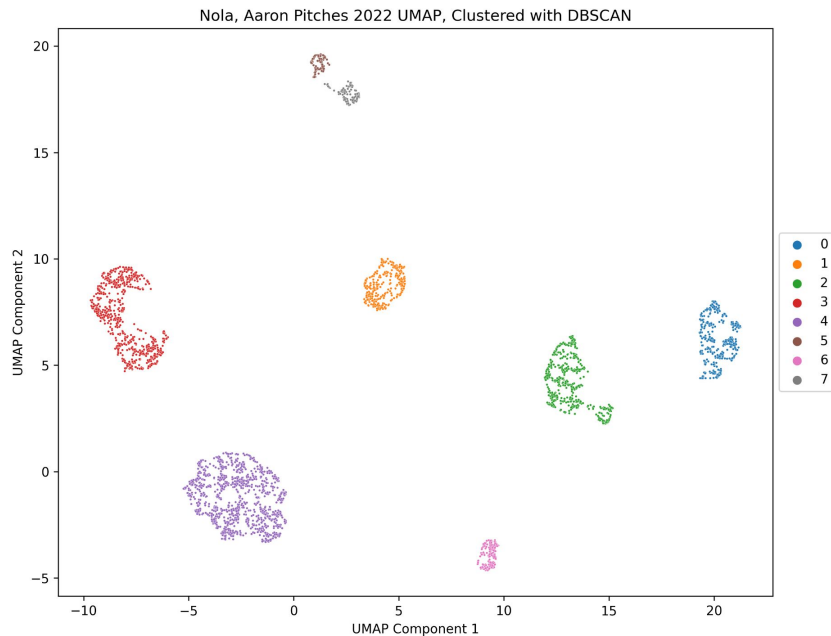
# Umap cluster for 2 pitchers

- Pitchers have distinct throwing styles even for similar pitches



Nola and Cole Fastballs 2022 UMAP, Clustered with DBSCAN



Nola and Cole Fastballs 2022 UMAP, with Statcast labels
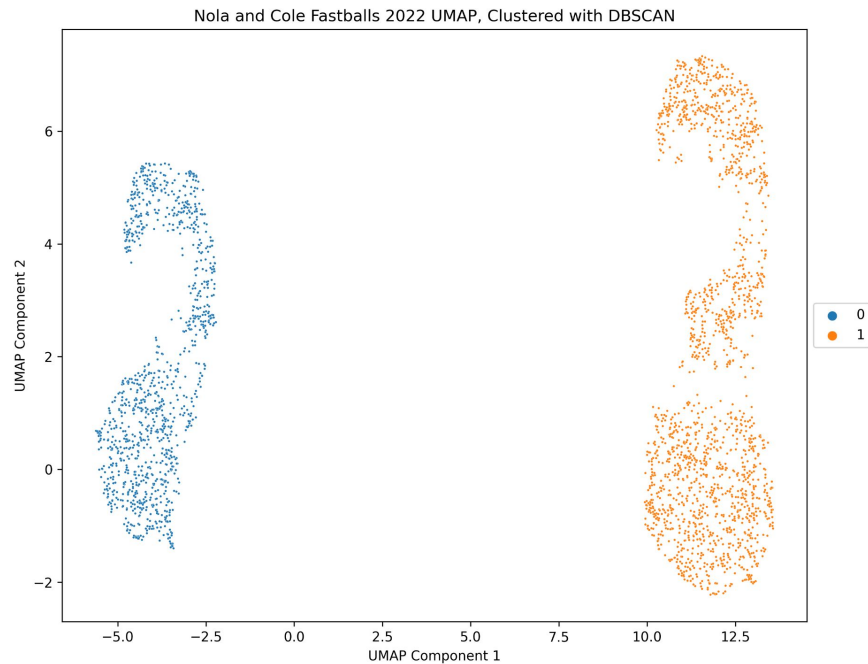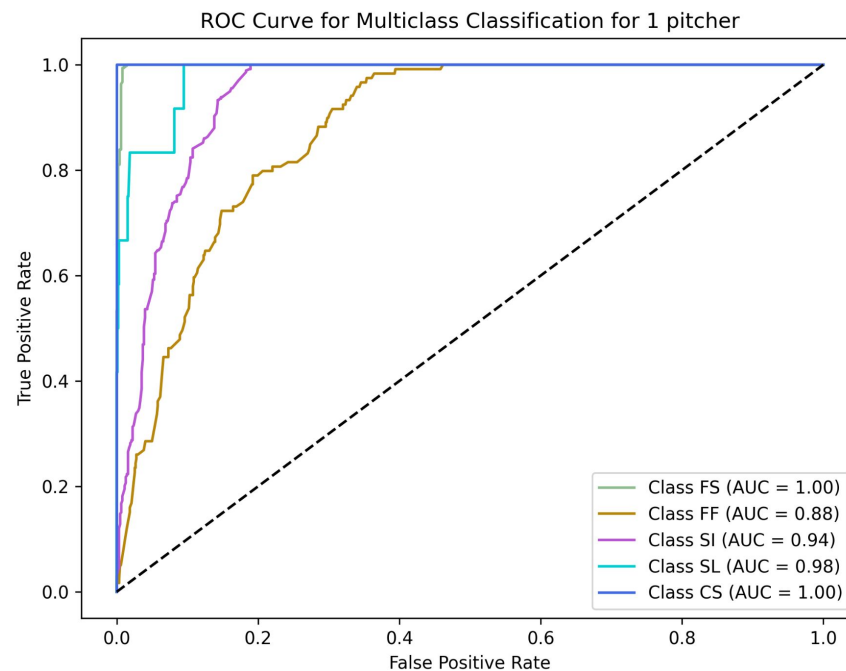
# Umap cluster for 10 pitchers

- Pitchers have distinct throwing styles even for similar pitches
- Some players do throw similarly though!
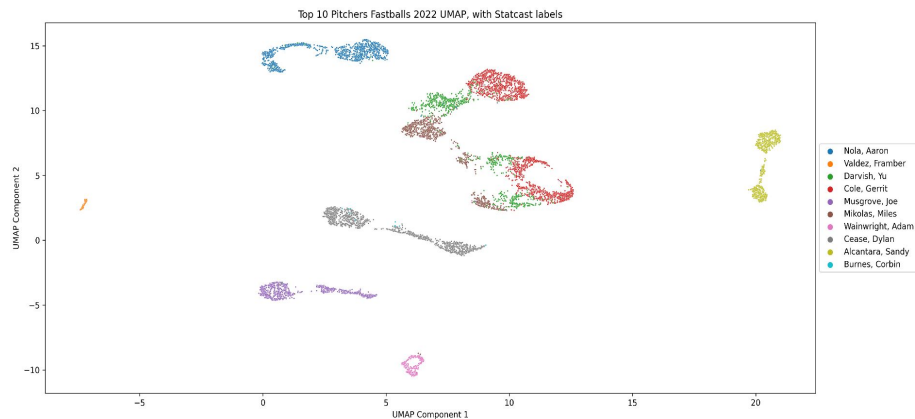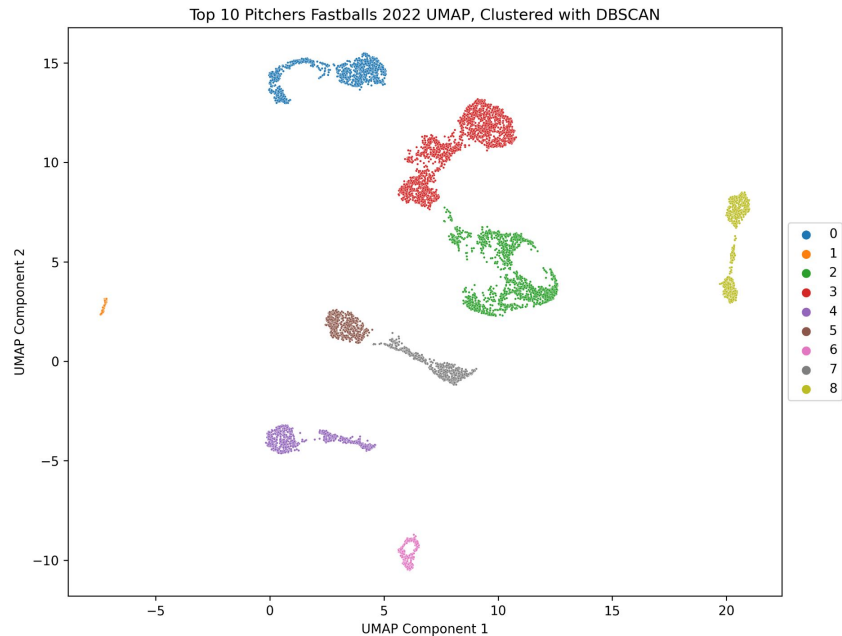
# Classification of pitch types (1 pitcher)

- Easy to classify highly represented pitches.
  - Similar pitches causes trouble
- Multiclass classification (LGBM)

| pitch_type_counts | |
| --- | --- |
| FS | 2482 |
| FF | 1767 |
| SI | 551 |
| SL | 89 |
| CS | 14 |



ROC Curve for Multiclass Classification for 1 pitcher

Class FS (AUC = 1.00)
Class FF (AUC = 0.88)
Class SI (AUC = 0.94)
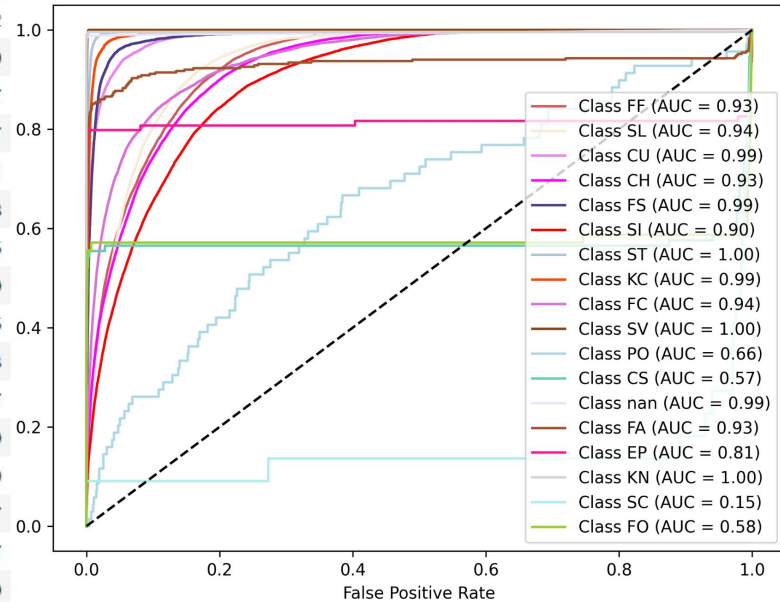Class SL (AUC = 0.98)
Class CS (AUC = 1.00)

# Classification of pitch types (All pitchers)

- Harder to classify with more than one pitcher
- Feature importance changes
  - General pitch features becomes more dominant
- Neural network was tried with much less success

| pitch_type_counts | |
| --- | --- |
| FF | 1137650 |
| SL | 573632 |
| CU | 541340 |
| CH | 343087 |
| FS | 256707 |
| SI | 199991 |
| ST | 84998 |
| KC | 52015 |
| FC | 40609 |
| SV | 20945 |
| PO | 8873 |
| CS | 3677 |
| nan | 1649 |
| FA | 600 |
| EP | 497 |
| KN | 367 |
| SC | 319 |
| FO | 77 |



ROC Curve for Multiclass Classification for All Pitchers (2017-2021)

Class FF (AUC = 0.93)
Class SL (AUC = 0.94)
Class CU (AUC = 0.99)
Class CH (AUC = 0.93)
Class FS (AUC = 0.99)
Class SI (AUC = 0.90)
Class ST (AUC = 1.00)
Class KC (AUC = 0.99)
Class FC (AUC = 0.94)
Class SV (AUC = 1.00)
Class PO (AUC = 0.66)
Class CS (AUC = 0.57)
Class nan (AUC = 0.99)
Class FA (AUC = 0.93)
Class EP (AUC = 0.81)
Class KN (AUC = 1.00)
Class SC (AUC = 0.15)
Class FO (AUC = 0.58)

# Overview of Models

- Binary Classification
  - "Handmade" good vs bad outcome

- Multiclass Classification
  - Pick the most important categories of each outcome

- Regression
  - Assign (domain knowledge) weights to each outcome

# Binary Classifier

- Able to predict 'positive' or 'negative outcomes quite well

- Works slightly better when looking at just fastballs, but slightly worse with only curveballs

- Quite a simple model, misses a lot of nuance
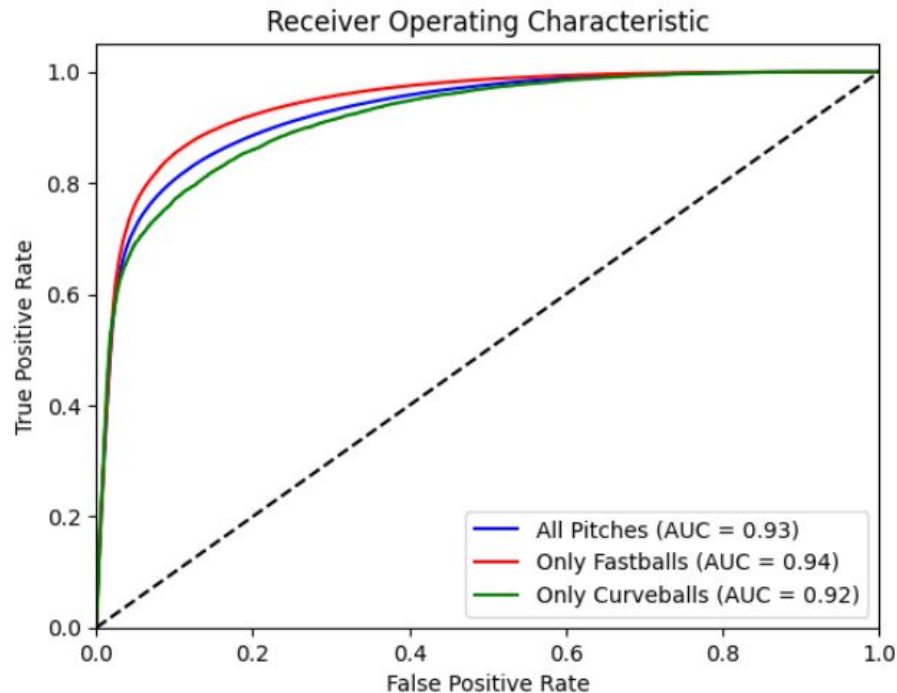
# Predicting the Outcome of a Pitch

- ● Classified three possible outcomes
  - ○ Hit into play = 1
  - ○ Strike = 2
  - ○ Ball = 3
- ● XGboost
  - ○ Initial score: 0.659
  - ○ Optimized score: 0.862



Feature Importance for Pitch Outcome Classification



ROC Curve for Pitch Outcome
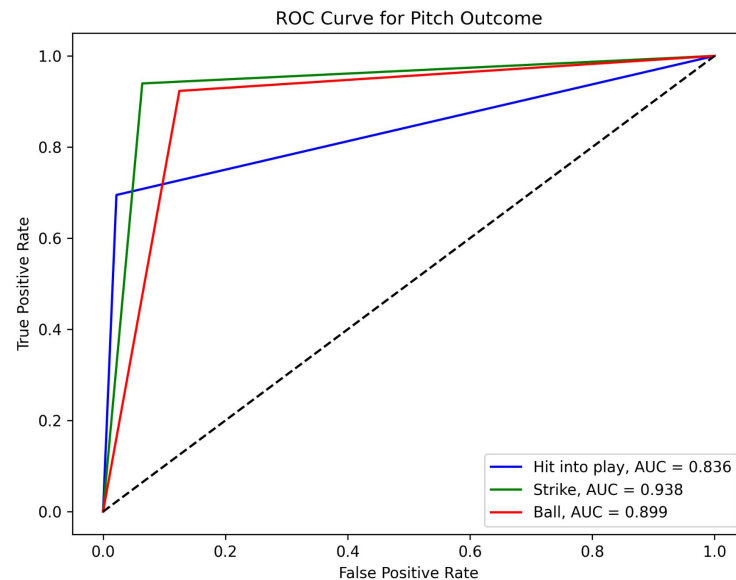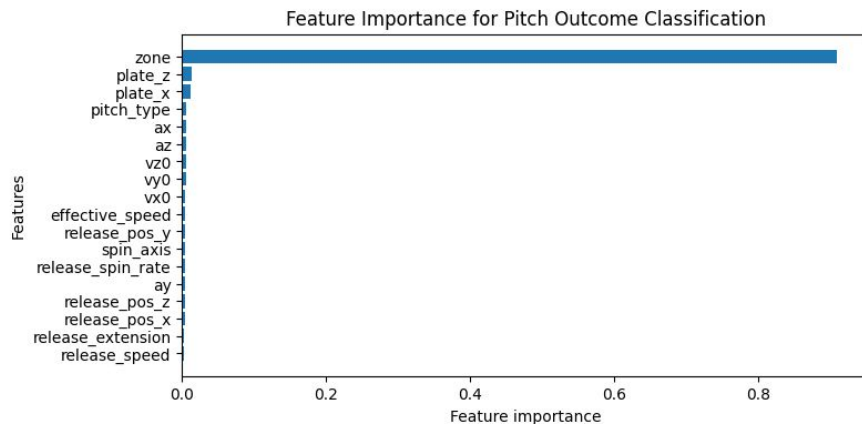
# Rating a pitch

- We created a scale to rate each pitch -> [-20,2]

- Train a LightGBM regression model separately for each pitch type on seasons 2017-2021

- Apply model to pitches based on the pitch type for 2022 pitches

| Outcome | Score |
|---|---|
| Ball | -1.5 |
| Foul | 1 |
| Swinging Strike | 1.5 |
| Called Strike | 2 |
| Hit Into Play | 0 |
| Home Run | -20 |
| Hit By Pitch | -5 |

# LightGBM Regressor Predictions (2022 Season)

n_pitches > 1500

| player_name | |
|---|---|
| ★ Rodón, Carlos | 0.177377 |
| Gausman, Kevin | 0.160355 |
| Webb, Logan | 0.147598 |
| Wheeler, Zack | 0.135382 |
| Strider, Spencer | 0.130871 |
| Rasmussen, Drew | 0.129163 |
| Snell, Blake | 0.124835 |
| Nola, Aaron | 0.121506 |
| Woodruff, Brandon | 0.118187 |
| ★ Alcantara, Sandy | 0.117598 |
| Urías, Julio | 0.117013 |
| Eovaldi, Nathan | 0.110395 |
| ★ Kershaw, Clayton | 0.109649 |
| ★ McClanahan, Shane | 0.109468 |
| Ashcraft, Graham | 0.107940 |
| Darvish, Yu | 0.103417 |
| Junis, Jakob | 0.102155 |
| ★ Verlander, Justin | 0.097099 |
| Cobb, Alex | 0.095898 |
| Gallen, Zac | 0.093069 |

- Starting Pitchers vs Relief Pitchers

- Mean score across all pitches thrown in the 2022 season

- Applying the regression model to pitches in the 2022 season

- Star denotes a player selected as an All-Star for the 2022 season

n_pitches > 250

| player_name | |
|---|---|
| ★ Clase, Emmanuel | 0.296691 |
| Williams, Devin | 0.209196 |
| Graterol, Brusdar | 0.205992 |
| Kimbrel, Craig | 0.200092 |
| ★ Rodón, Carlos | 0.177377 |
| Morejon, Adrian | 0.177167 |
| deGrom, Jacob | 0.170946 |
| Gausman, Kevin | 0.160355 |
| Fairbanks, Pete | 0.156196 |
| Scott, Tanner | 0.155824 |
| Akin, Keegan | 0.155582 |
| Brasier, Ryan | 0.149683 |
| Megill, Trevor | 0.147786 |
| Webb, Logan | 0.147598 |
| ★ Díaz, Edwin | 0.146494 |
| Nardi, Andrew | 0.143753 |
| Minter, A.J. | 0.137327 |
| Sewald, Paul | 0.137157 |
| Bickford, Phil | 0.135504 |
| Wheeler, Zack | 0.135382 |

# Evaluation of Regression Predictions

- Generally predicts outcomes fairly well

- MAE: 1.007

- Home run weight affects the error measurement

- Error is higher for rarer pitch types



Histogram of Absolute Error for All Pitches
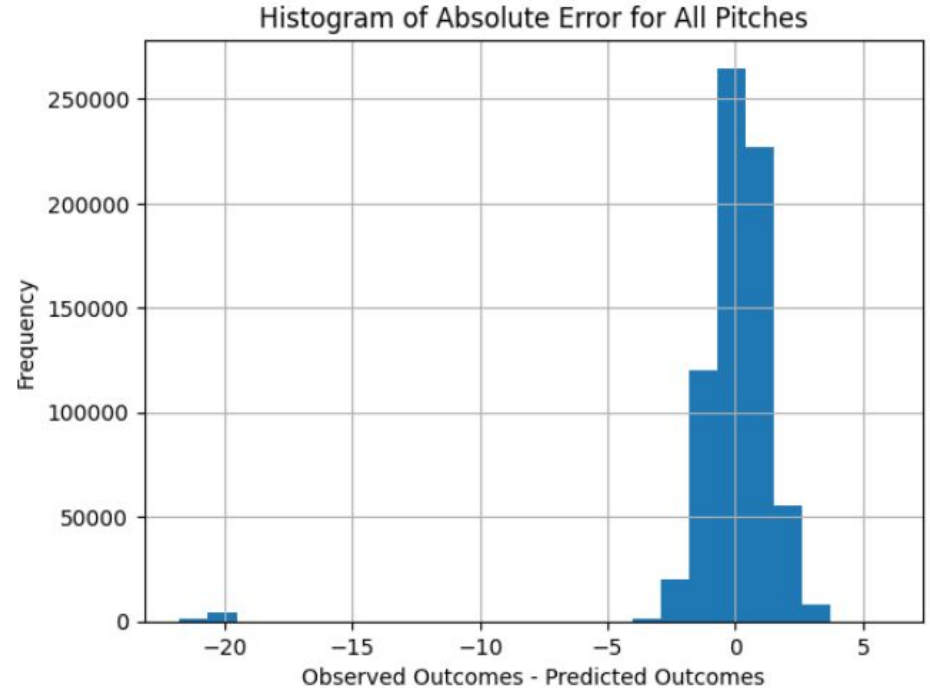
# How Do We Compare to Traditional Metrics?

- Training regressor on 2017-2020, and predicting on 2021

- Find intersection of ERA dataset and our model
  - Rank players
  - Calculate correlation

| 2021 ERA vs 2022 ERA | Our model on 2021 pitchers vs 2022 ERA |
|---|---|
| 0.444 | 0.586 |

# Conclusion

- Classify pitch types, to a good degree

- Classify pitch outcomes

  - Simplified version of rating pitchers

- Pitch types are highly individualised

- This sort of scoring framework has potential!

- Domain knowledge is important

# Future ideas

- Predict injuries based on change in pitch score

- Find best pitchers for specific environments eg. low air pressure

- Are early pitches predictive of later pitches?

- Context is probably important! Rating outcomes based on encounters between batter and pitcher instead of each individual pitch.

- Develop a more objective scoring scale

Thank you for your attention

# Appendix

Things we tried that didn't work and all other things we have done.

# Classification of pitches with LGBM (NN did not give any useable results for this)

Features is found by using permutation_importance (sklearn)

- For 1 pitcher features used is:
  - ['release_speed', 'fielder_7', 'fielder_6', 'fielder_5']
- For all pitchers features used is:
  - ['release_speed', 'release_pos_x', 'release_pos_z', 'batter', 'pitcher', 'description', 'spin_dir']

Random search was used to find hyperparameters

- For 1 pitcher hyperparameters used is:
  - lr=0.237, max_depth=28 and num_leaves=54
- For all pitchers hyperparameters used is:
  - lr=0.0217, max_depth=39 and num_leaves=57

# Classification of outcomes with XGboost(NN took forever to train and produced worse results than XGboost)

- Created an original model with 19 features
- Initial parameters: booster = 'gbtree', use_label_encoder=False,
- n_estimators=15, max_depth = 9, max_leaves = 5,
- max_delta_step = 10
- Used Bayesian Optimization on n_estimators, max_depth, max_leaves
- Final parameters: booster = 'gbtree', use_label_encoder=False,
- n_estimators=5, max_depth = 20, max_leaves = 11,
- max_delta_step = 10

# Pitch Scoring Regressor Notes

- Only used naive hyperparameter tuning, could probably be improved through more rigorous methods
- Runtime/memory was starting to become an issue
- Scale was developed in collaboration with a baseball expert (Charles)

**Training Variables Used**: 'release_speed', 'zone', 'p_throws', 'pfx_x', 'pfx_z', 'stand', 'plate_x', 'plate_z', 'vx0', 'vy0', 'vz0', 'ax', 'ay', 'az', 'effective_speed', 'release_spin_rate', 'release_extension', 'release_pos_x', 'release_pos_z', 'release_pos_y', 'spin_axis'

**Other Relevant Variables**: 'description', 'events', 'pitch_type'

**Hyperparameters**: 'learning_rate': 0.01, 'max_depth': 50, 'n_estimators': 1000, 'num_leaves': 52

# Regressor Feature Importances

- LightGBM In-Built feature importances for each individual pitch-type model
- Found that the final coordinates of the pitch were quite often the most important variables

**CH**
```
Feature: plate_x, Importance: 6787
Feature: plate_z, Importance: 5954
Feature: vx0, Importance: 3267
Feature: release_pos_z, Importance: 2973
Feature: ay, Importance: 2803
Feature: release_spin_rate, Importance: 2745
Feature: az, Importance: 2689
Feature: vz0, Importance: 2598
Feature: release_pos_y, Importance: 2487
Feature: pfx_z, Importance: 2466
Feature: pfx_x, Importance: 2441
Feature: zone, Importance: 2186
Feature: spin_axis, Importance: 2109
Feature: release_pos_x, Importance: 1890
Feature: effective_speed, Importance: 1770
Feature: ax, Importance: 1723
Feature: vy0, Importance: 1351
Feature: release_speed, Importance: 1331
Feature: stand, Importance: 992
Feature: release_extension, Importance: 435
Feature: p_throws, Importance: 3
```

**CS**
```
Feature: plate_x, Importance: 1691
Feature: vx0, Importance: 1627
Feature: plate_z, Importance: 1493
Feature: release_spin_rate, Importance: 1175
Feature: ax, Importance: 1156
Feature: vz0, Importance: 1152
Feature: pfx_x, Importance: 1100
Feature: ay, Importance: 1061
Feature: release_speed, Importance: 1035
Feature: release_pos_z, Importance: 906
Feature: effective_speed, Importance: 894
Feature: release_pos_x, Importance: 775
Feature: release_pos_y, Importance: 757
Feature: spin_axis, Importance: 671
Feature: az, Importance: 615
Feature: zone, Importance: 556
Feature: pfx_z, Importance: 525
Feature: vy0, Importance: 446
Feature: release_extension, Importance: 141
Feature: stand, Importance: 121
Feature: p_throws, Importance: 0
```

**CU**
```
Feature: plate_x, Importance: 7011
Feature: plate_z, Importance: 5398
Feature: release_pos_z, Importance: 3133
Feature: vx0, Importance: 2959
Feature: vz0, Importance: 2928
Feature: ax, Importance: 2716
Feature: release_spin_rate, Importance: 267₿
Feature: ay, Importance: 2590
Feature: release_pos_y, Importance: 2515
Feature: release_pos_x, Importance: 2435
Feature: zone, Importance: 2134
Feature: spin_axis, Importance: 2042
Feature: effective_speed, Importance: 2036
Feature: pfx_z, Importance: 1985
Feature: pfx_x, Importance: 1969
Feature: az, Importance: 1778
Feature: release_speed, Importance: 1570
Feature: stand, Importance: 1411
Feature: vy0, Importance: 1162
Feature: release_extension, Importance: 544
Feature: p_throws, Importance: 6
```

**EP**
```
Feature: vx0, Importance: 1742
Feature: plate_x, Importance: 1619
Feature: ax, Importance: 1591
Feature: plate_z, Importance: 1302
Feature: release_pos_x, Importance: 1268
Feature: ay, Importance: 1253
Feature: release_pos_z, Importance: 1235
Feature: vz0, Importance: 1122
Feature: release_pos_y, Importance: 1084
Feature: pfx_x, Importance: 1032
Feature: release_spin_rate, Importance: 976
Feature: zone, Importance: 846
Feature: spin_axis, Importance: 769
Feature: az, Importance: 719
Feature: pfx_z, Importance: 600
Feature: effective_speed, Importance: 533
Feature: release_speed, Importance: 520
Feature: vy0, Importance: 492
Feature: stand, Importance: 281
Feature: release_extension, Importance: 238
Feature: p_throws, Importance: 0
```

**FC**
```
Feature: plate_x, Importance: 6226
Feature: plate_z, Importance: 5970
Feature: vx0, Importance: 2945
Feature: ay, Importance: 2943
Feature: vz0, Importance: 2929
Feature: release_pos_z, Importance: 2908
Feature: release_pos_y, Importance: 2831
Feature: release_spin_rate, Importance: 2799
Feature: release_pos_x, Importance: 2621
Feature: pfx_z, Importance: 2569
Feature: spin_axis, Importance: 2168
Feature: pfx_x, Importance: 2074
Feature: zone, Importance: 1932
Feature: az, Importance: 1862
Feature: ax, Importance: 1838
Feature: effective_speed, Importance: 1703
Feature: release_speed, Importance: 1439
Feature: vy0, Importance: 1337
Feature: stand, Importance: 1112
Feature: p_throws, Importance: 497
Feature: release_extension, Importance: 297
```

**FF**
```
Feature: plate_x, Importance: 8083
Feature: plate_z, Importance: 7691
Feature: vx0, Importance: 3022
Feature: az, Importance: 2522
Feature: release_spin_rate, Importance: 2519
Feature: release_pos_z, Importance: 2459
Feature: stand, Importance: 2258
Feature: vz0, Importance: 2251
Feature: zone, Importance: 2249
Feature: release_pos_x, Importance: 2116
Feature: pfx_z, Importance: 2028
Feature: release_pos_y, Importance: 1939
Feature: pfx_x, Importance: 1904
Feature: ax, Importance: 1880
Feature: ay, Importance: 1830
Feature: effective_speed, Importance: 1586
Feature: spin_axis, Importance: 1584
Feature: release_speed, Importance: 1323
Feature: vy0, Importance: 1157
Feature: release_extension, Importance: 478
Feature: p_throws, Importance: 121
```

**FO**
```
Feature: release_pos_x, Importance: 1313
Feature: release_spin_rate, Importance: 1200
Feature: vz0, Importance: 1059
Feature: ay, Importance: 902
Feature: plate_z, Importance: 702
Feature: vx0, Importance: 685
Feature: release_pos_z, Importance: 642
Feature: zone, Importance: 624
Feature: plate_x, Importance: 589
Feature: ax, Importance: 521
Feature: release_pos_y, Importance: 479
Feature: pfx_x, Importance: 421
Feature: pfx_z, Importance: 417
Feature: vy0, Importance: 409
Feature: effective_speed, Importance: 398
Feature: az, Importance: 349
Feature: release_speed, Importance: 313
Feature: spin_axis, Importance: 252
Feature: stand, Importance: 103
Feature: release_extension, Importance: 44
Feature: p_throws, Importance: 0
```

# Regressor Feature Importances (Cont.)

**FS**
Feature: plate_x, Importance: 5302
Feature: plate_z, Importance: 5116
Feature: release_pos_y, Importance: 3375
Feature: ay, Importance: 3257
Feature: vz0, Importance: 3118
Feature: release_spin_rate, Importance: 3116
Feature: release_pos_z, Importance: 3027
Feature: vx0, Importance: 2793
Feature: pfx_x, Importance: 2566
Feature: release_pos_x, Importance: 2420
Feature: spin_axis, Importance: 2281
Feature: az, Importance: 2278
Feature: pfx_z, Importance: 2209
Feature: ax, Importance: 2129
Feature: zone, Importance: 2016
Feature: effective_speed, Importance: 1902
Feature: vy0, Importance: 1628
Feature: release_speed, Importance: 1402
Feature: stand, Importance: 593
Feature: release_extension, Importance: 471
Feature: p_throws, Importance: 1

**KC**
Feature: plate_x, Importance: 5751
Feature: plate_z, Importance: 4813
Feature: vz0, Importance: 3308
Feature: release_spin_rate, Importance: 3271
Feature: release_pos_z, Importance: 3157
Feature: release_pos_y, Importance: 3100
Feature: release_pos_x, Importance: 2985
Feature: ay, Importance: 2944
Feature: vx0, Importance: 2525
Feature: ax, Importance: 2394
Feature: effective_speed, Importance: 2369
Feature: spin_axis, Importance: 2130
Feature: pfx_z, Importance: 2028
Feature: pfx_x, Importance: 1958
Feature: az, Importance: 1956
Feature: release_speed, Importance: 1831
Feature: zone, Importance: 1759
Feature: vy0, Importance: 1569
Feature: stand, Importance: 726
Feature: release_extension, Importance: 421
Feature: p_throws, Importance: 5

**KN**
Feature: vx0, Importance: 1195
Feature: plate_z, Importance: 901
Feature: release_spin_rate, Importance: 850
Feature: zone, Importance: 771
Feature: ay, Importance: 743
Feature: plate_x, Importance: 689
Feature: pfx_z, Importance: 674
Feature: spin_axis, Importance: 668
Feature: vz0, Importance: 650
Feature: az, Importance: 617
Feature: ax, Importance: 584
Feature: pfx_x, Importance: 472
Feature: release_pos_y, Importance: 413
Feature: release_pos_z, Importance: 406
Feature: release_pos_x, Importance: 383
Feature: release_speed, Importance: 379
Feature: effective_speed, Importance: 269
Feature: vy0, Importance: 174
Feature: release_extension, Importance: 122
Feature: stand, Importance: 55
Feature: p_throws, Importance: 0

**SC**
Feature: vy0, Importance: 316
Feature: vz0, Importance: 218
Feature: pfx_x, Importance: 187
Feature: ay, Importance: 176
Feature: release_pos_y, Importance: 174
Feature: zone, Importance: 148
Feature: release_pos_x, Importance: 146
Feature: pfx_z, Importance: 127
Feature: release_speed, Importance: 122
Feature: vx0, Importance: 112
Feature: ax, Importance: 73
Feature: release_pos_z, Importance: 73
Feature: release_spin_rate, Importance: 53
Feature: plate_z, Importance: 27
Feature: plate_x, Importance: 21
Feature: effective_speed, Importance: 14
Feature: az, Importance: 9
Feature: spin_axis, Importance: 4
Feature: stand, Importance: 0
Feature: release_extension, Importance: 0
Feature: p_throws, Importance: 0

**SI**
Feature: plate_x, Importance: 7883
Feature: plate_z, Importance: 6812
Feature: vx0, Importance: 3324
Feature: ay, Importance: 2587
Feature: pfx_z, Importance: 2573
Feature: release_spin_rate, Importance: 2526
Feature: ax, Importance: 2341
Feature: stand, Importance: 2262
Feature: vz0, Importance: 2252
Feature: release_pos_z, Importance: 2251
Feature: zone, Importance: 2187
Feature: pfx_x, Importance: 2120
Feature: az, Importance: 2114
Feature: release_pos_y, Importance: 2087
Feature: spin_axis, Importance: 1844
Feature: release_pos_x, Importance: 1812
Feature: effective_speed, Importance: 1525
Feature: release_speed, Importance: 1126
Feature: vy0, Importance: 968
Feature: release_extension, Importance: 405
Feature: p_throws, Importance: 1

**SL**
Feature: plate_x, Importance: 7056
Feature: plate_z, Importance: 6395
Feature: release_pos_z, Importance: 2912
Feature: vz0, Importance: 2858
Feature: vx0, Importance: 2650
Feature: release_spin_rate, Importance: 2596
Feature: ay, Importance: 2565
Feature: pfx_z, Importance: 2540
Feature: release_pos_x, Importance: 2429
Feature: ax, Importance: 2379
Feature: release_pos_y, Importance: 2317
Feature: spin_axis, Importance: 2061
Feature: pfx_x, Importance: 2020
Feature: az, Importance: 1768
Feature: zone, Importance: 1712
Feature: effective_speed, Importance: 1660
Feature: stand, Importance: 1556
Feature: release_speed, Importance: 1416
Feature: vy0, Importance: 1208
Feature: p_throws, Importance: 555
Feature: release_extension, Importance: 347

**ST**
Feature: plate_x, Importance: 5187
Feature: plate_z, Importance: 4017
Feature: ay, Importance: 3467
Feature: vz0, Importance: 3420
Feature: release_spin_rate, Importance: 3042
Feature: release_pos_z, Importance: 2952
Feature: vx0, Importance: 2902
Feature: spin_axis, Importance: 2748
Feature: release_pos_x, Importance: 2691
Feature: ax, Importance: 2649
Feature: effective_speed, Importance: 2409
Feature: release_pos_y, Importance: 2407
Feature: az, Importance: 2402
Feature: pfx_z, Importance: 2383
Feature: pfx_x, Importance: 2137
Feature: release_speed, Importance: 1827
Feature: zone, Importance: 1740
Feature: vy0, Importance: 1688
Feature: stand, Importance: 576
Feature: release_extension, Importance: 356
Feature: p_throws, Importance: 0

**SV**
Feature: plate_x, Importance: 4702
Feature: release_spin_rate, Importance: 4468
Feature: plate_z, Importance: 4206
Feature: vz0, Importance: 3453
Feature: ay, Importance: 3433
Feature: vx0, Importance: 3290
Feature: release_pos_x, Importance: 3137
Feature: release_pos_z, Importance: 3043
Feature: release_pos_y, Importance: 3020
Feature: ax, Importance: 2487
Feature: spin_axis, Importance: 2388
Feature: az, Importance: 2051
Feature: pfx_z, Importance: 2045
Feature: pfx_x, Importance: 1988
Feature: effective_speed, Importance: 1879
Feature: vy0, Importance: 1862
Feature: release_speed, Importance: 1375
Feature: zone, Importance: 1130
Feature: stand, Importance: 606
Feature: release_extension, Importance: 437
Feature: p_throws, Importance: 0

# How is a pitch type classified?

- Defining the most common pitch type (Fastballs)
  - "a baseball pitch thrown at full speed and often appearing to rise slightly as it nears the plate" - Meriam-webster dictionary
  - "a type of pitch that results from a pitcher throwing the baseball as hard and as fast as possible" - Wonderopolis

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| FF  | FS  | SI  | KC  | CH  | FC  | SL  | ST  | CU  | PO  | CS  | FA  | SV  | KN  | NaN | EP  | SC  |
|  0  |  1  |  2  |  3  |  4  |  5  |  6  |  7  |  8  |  9  | 10  | 11  | 12  | 13  | 14  | 15  | 16  |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
```

# Confusion Matrix - LGBM Multiclassification Model

- Early "Toy Model", we later
  chose to focus on less categories
  which gave better accuracy.
  Eventually we used a regression model
  that captured the same features



|  | precision | recall | f1-score | support |
|---|---|---|---|---|
| ball | 0.73 | 0.92 | 0.81 | 353711 |
| blocked_ball | 0.29 | 0.04 | 0.07 | 24818 |
| called_strike | 0.51 | 0.72 | 0.60 | 175168 |
| foul | 0.39 | 0.29 | 0.33 | 184801 |
| hit_into_play | 0.43 | 0.41 | 0.42 | 181571 |
| other | 0.50 | 0.13 | 0.20 | 23166 |
| swinging_strike | 0.44 | 0.17 | 0.24 | 106998 |
| accuracy |  |  | 0.57 | 1050233 |
| macro avg | 0.47 | 0.38 | 0.38 | 1050233 |
| weighted avg | 0.54 | 0.57 | 0.54 | 1050233 |
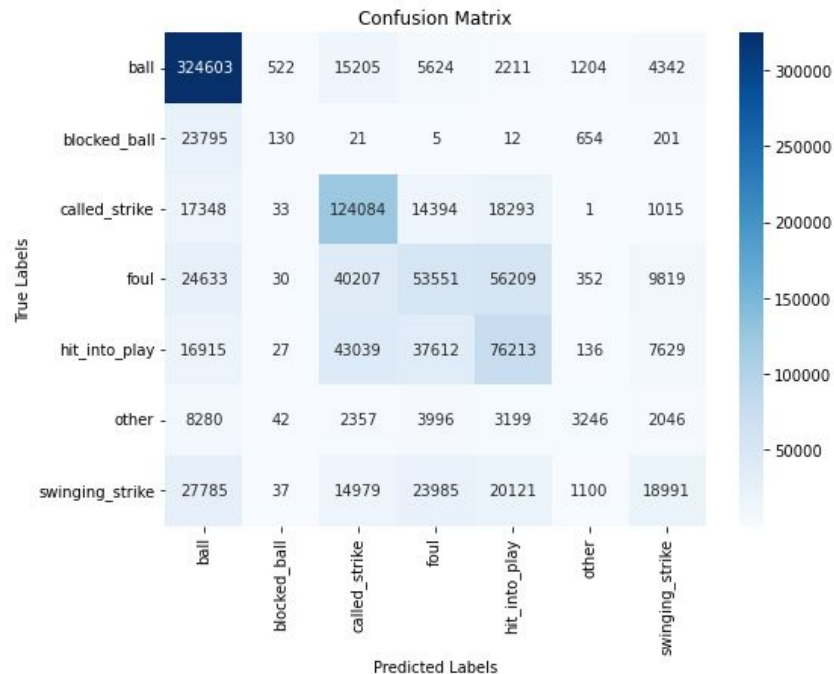
# Location specific pitches

Our idea was to see if pitches, initially just fastballs would be different in different locations. We didn't see much difference for the pitcher having the most pitches in "Denver, Colorado", so we abandoned this idea for now.



Senzatela, Antonio Fastballs 2022 UMAP with Location Labels