**Algos:**

1. **K-means/K-prototypes**
2. **Gaussian Mixture Model**
3. **Hierarchical Clustering**

**Use historic matchups from batter vs. all pitchers from same cluster 🡺 predict PA RV**

**Features (47`)**

1. **Player:**
   1. Handedness
   2. Position(SP/RP)
   3. Release\_pos\_x
   4. Release\_pos\_z
2. **Pitch type percentage/speed/break(x/z)/spin\_rate/spin\_axis**
   1. **Fastball:** FF
   2. **Sinker:** FT, SI
   3. **Slider/Cutter:** SL, FC
   4. **Curve:** CB, KC
   5. **Off-speed:** CH, FS
   6. **Remove:** Slow FA, EP, KN
   7. Number of pitch type
3. **Contact profile**
   1. Avg exit velo
   2. Barrel + Solid percent
   3. **Launch Angle percent:** GB, LD, FB, PU
   4. GB/FB ratio
4. **Swing profile**
   1. Whiff rate
   2. O-Swing % (Chase Rate)
   3. Zone Contact %
5. **Production Stats**
   1. K/9, H/9, BB/9, HR/9

Questions:

1. Pitch success rate?
2. Replace missing pitch data with 0?
3. Neutralize handedness?
   1. Standardize categorical data? (L/R)
4. If matchup is less than 600 pa, regress to the mean?

**RE24+PA Outcome Matrix**

In this situation, the sample space is plate appearances, each of which can result in a single, double, triple, home run, walk, hit-by-pitch, error, sacrifice hit, sacrifice fly, fielder’s choice, catcher’s interference, or out. The batter will be called out if an out, sacrifice hit, or sacrifice fly occurs.

**True PA Outcome:**

1. Field out
2. Single
3. Strikeout
4. Home run
5. Fielders choice
6. Grounded into double play
7. Hit by pitch
8. Walk
9. Double
10. Sac bunt
11. Sac fly
12. Force out
13. Field error
14. Triple
15. Fielders choice out
16. Double play
17. Strike out double play
18. Sac bunt double play
19. Sac fly double play
20. Triple play

**Fake PA Outcome:**

1. Caught stealing 2b
2. Wild pitch
3. Caught stealing home
4. Pick off 1b
5. Catcher interference
6. Caught stealing 3b
7. Pickoff 2b
8. Game advisory
9. Pickoff caught stealing home
10. Passed ball
11. Pickoff caught stealing 3b
12. Stolen base 2b
13. Runner double play
14. Pick off 3b
15. Pick off caught stealing 2b
16. Other out

'field\_out',

Field error

'single',

'strikeout',

1. Strikeout double play

'home\_run',

'fielders\_choice',

1. Force out
2. Fielders choice out

'grounded\_into\_double\_play',

1. Double play
2. Sac bunt double play
3. Sac fly double play

'walk'

1. Hit by pitch

'double',

'sac\_bunt',

'sac\_fly',

'triple',

Table

Description automatically generated with medium confidence'triple\_play'

**Limitations**

1. Clustering

Statcast pitch types are shitty 🡺 have to use a less granular grouping

Statcast strike zone sizes are shitty

Pitch movement data are shitty

Have to use a less granular grouping because each pitcher’s arsenal is different

1. Prediction:

new pitchers and new batters

Lots of pitchers in new batters

Brand new pitcher: use past batter’s avg

Brand new batter: use league vs pitcher avg

Pitcher not in cluster: