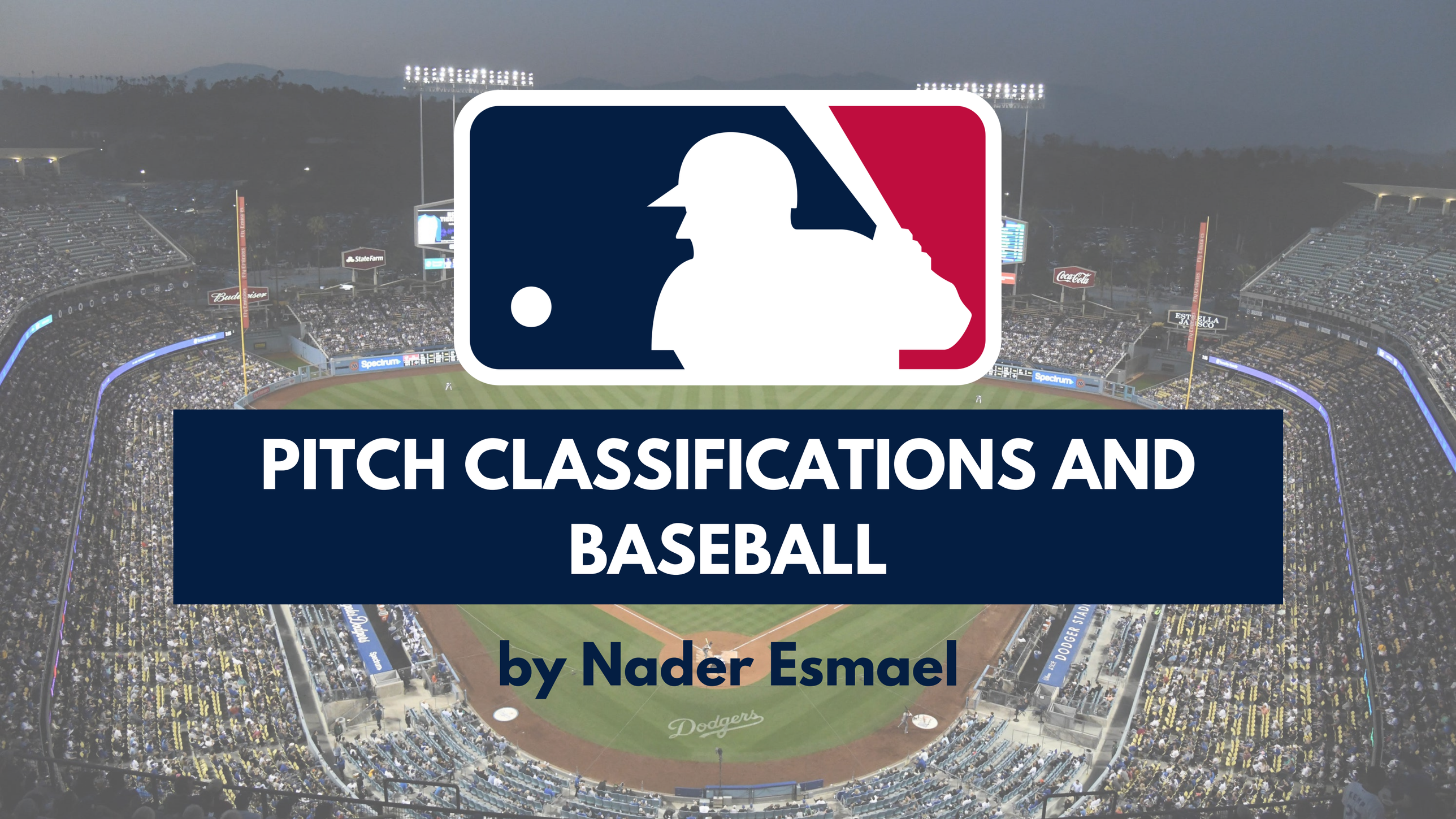
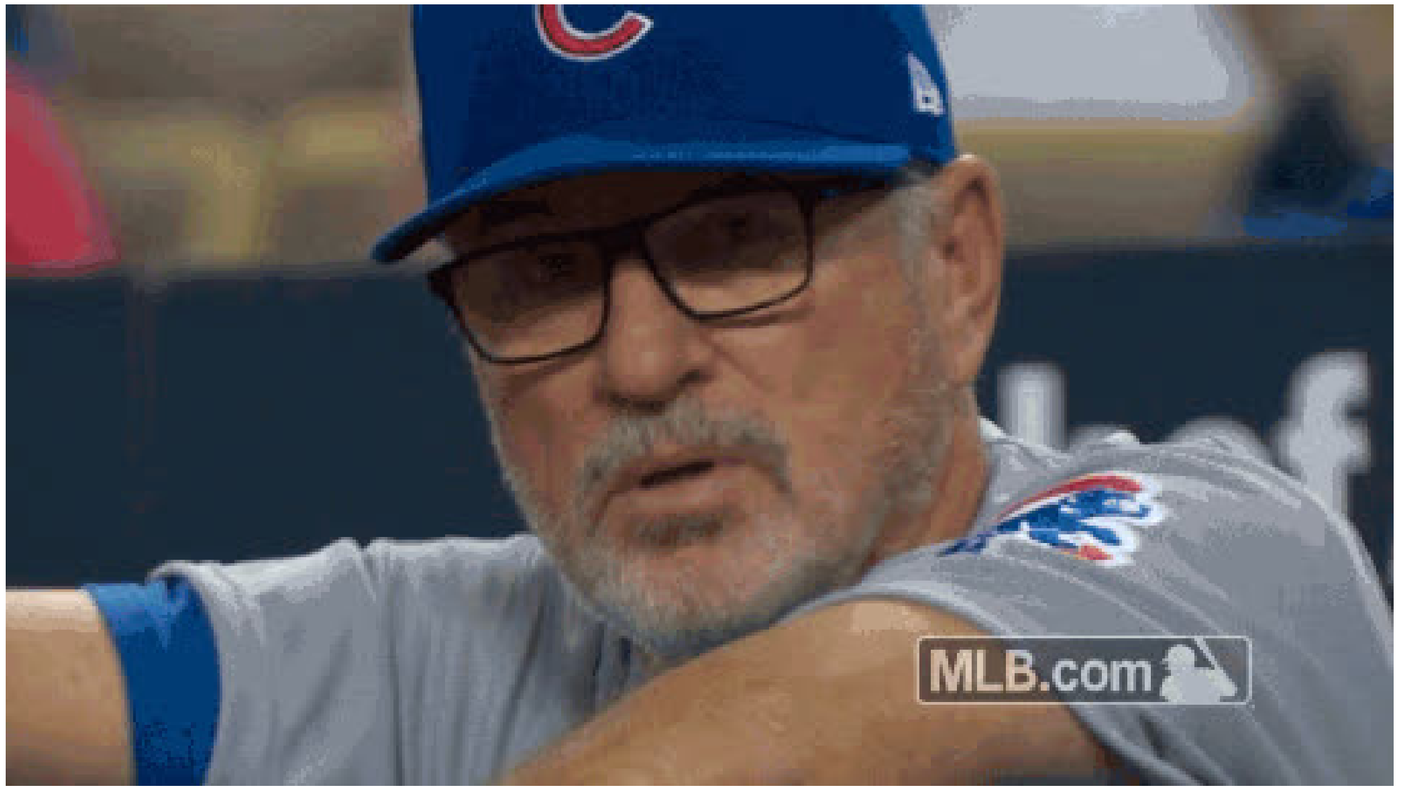


PITCH CLASSIFICATIONS AND BASEBALL

by Nader Esmael

YOU JUST HIT A BALL WITH A STICK RIGHT?



WHAT IS BASEBALL?

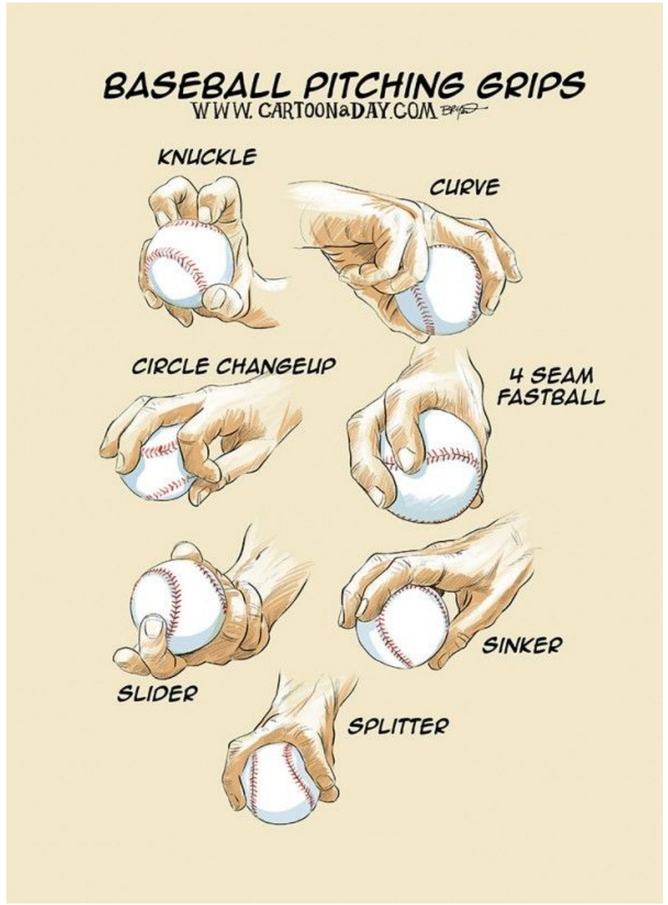
According to Wikipedia (sounds like the assumption), it is a "bat-and-ball game played between two opposing teams who take turns batting and fielding."

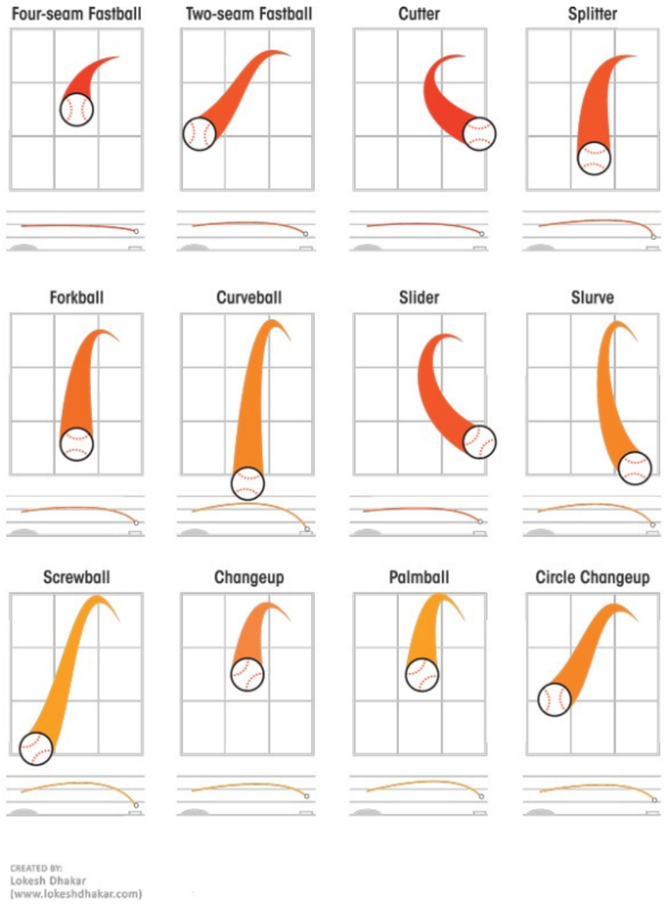
To a person in the MLB, it is a game of athletic capabilities, strategy, hand-eye coordination, and instincts/command.

You have to know when the right time to swing, which base to throw the ball to, determine if it's worth walking a batter, and at times, giving up this game for the next game.

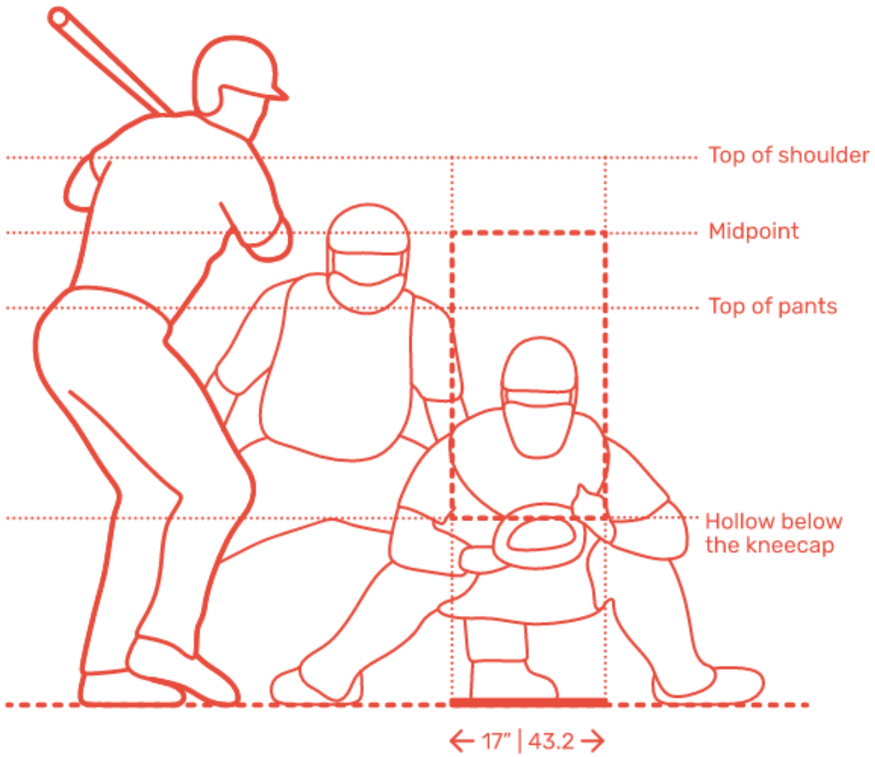
Baseball - Wikipedia

TOO MANY TYPES OF PITCHES





WHAT IS THE STRIKE ZONE?

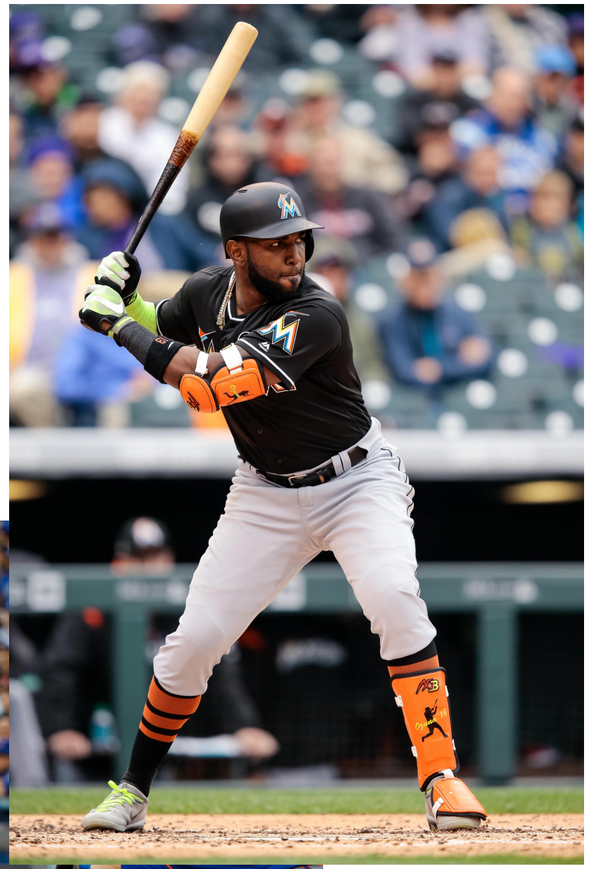


In baseball, the strike zone refers to "the volume of space which a ball must pass through to be called a 'strike' (if the batter doesn't swing)". They are calculated as the space between the width of home-plate, up to the midpoint between a batter's shoulders and uniform pants when in their stance, and extending down to just below their kneecaps.

In short, the official strike zone looks generally different with every single pitch thrown.

www.dimensions.com/element/strike-zone

DEFINING THE ZONE IS REALLY HARD



THREE STRIKES, YOU'RE OUT!!

I hope I made the right call...

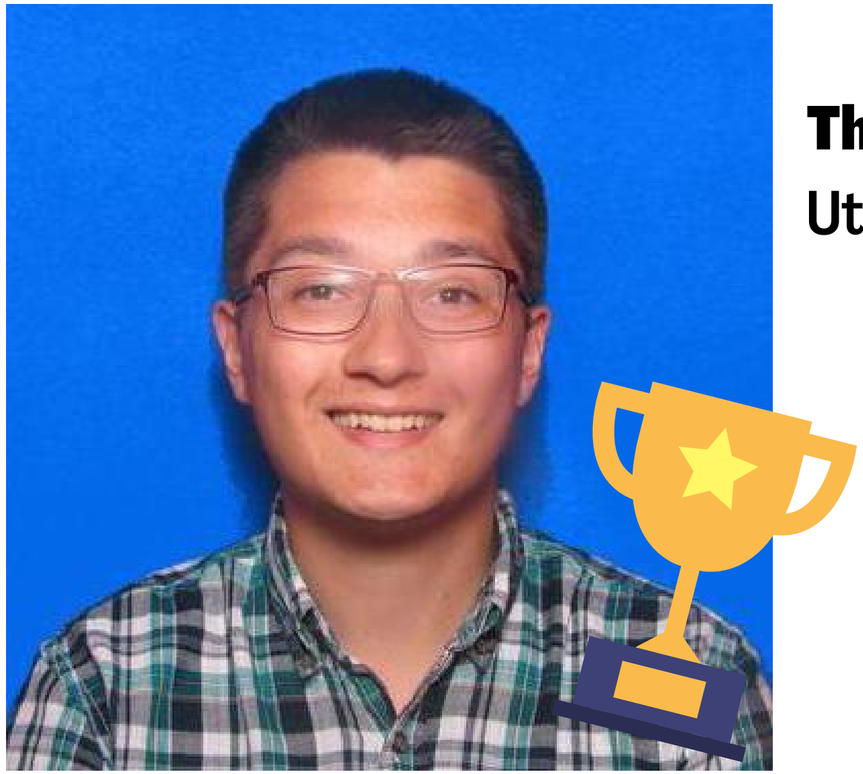
Regardless how difficult it is to make the call and define the strike zone, the umpire has the final say.

THAT'S A STRIKE!?

The Problem

Umpires have been giving America's past time too many issues. Too many strikes are called balls and too many balls thrown far away from the plate are called strikes. AI umpires have been considered to "perfect" officiating calls. With the help of data science, we can validate/rank the officating of particular umpire.

DATA COLLECTION - PYBASEBALL



Thank you Mr. James LeDoux, you are my MVP!

Utilizing pybaseball:

- Scrape data from baseballsavant.mlb.com
- Gathered game-log data from 2015-2020, including playoffs and World Series
- Whole dataset size ~ about 4 million rows of data with about 90 columns of features

DATA CLEANING PROCESS

COLUMN DROPPING

I had to reduce the amount of columns from 90 to ~55. Many columns were nulls. Removed columns that refer to hit locations, ballpark, fielders, runners, defensive alignment, and very specific stats.

FEATURE ENGINEERING

Used the "description" column to create "strike attempt" (target). Compared it to "type" column (short-hand for events).

MORE FEATURE ENGINEERING

Created 17 columns that referred to pitch name usage. Ex. .456 could refer that out of all pitches, they threw this pitch about 45% of the time.

NO NULLS & NO BATTERS

To prevent code crashing due to errors, instead of imputing nulls, the row was just dropped. Did not add any specific batter information except for their stance.

DATA CLEANING PROCESS CONTINUED

ANONYMIZED PLAYER FEATURES

Initially, I wanted to include the "batter id" and the "pitcher id" because a particular player would have a really strong influence to balls that are not swung on. However, they were dropped due lack of resources and computing power. We were able to dummify the particular pitch name thrown.

SOME INTERESTING FACTS

TOP 3 USED PITCHES

4-SEAM FASTBALL

SLIDER

CHANGE-UP

TOP 3 PITCHERS W/ MANY REPERTOIRS

YU DARVISH

JEFF SAMARDZIJA

16-WAY TIE

OTHER FACTS

The Atlanta Braves scored the most runs, in the last 6 seasons, against the Marlins. Final Score: 29-9

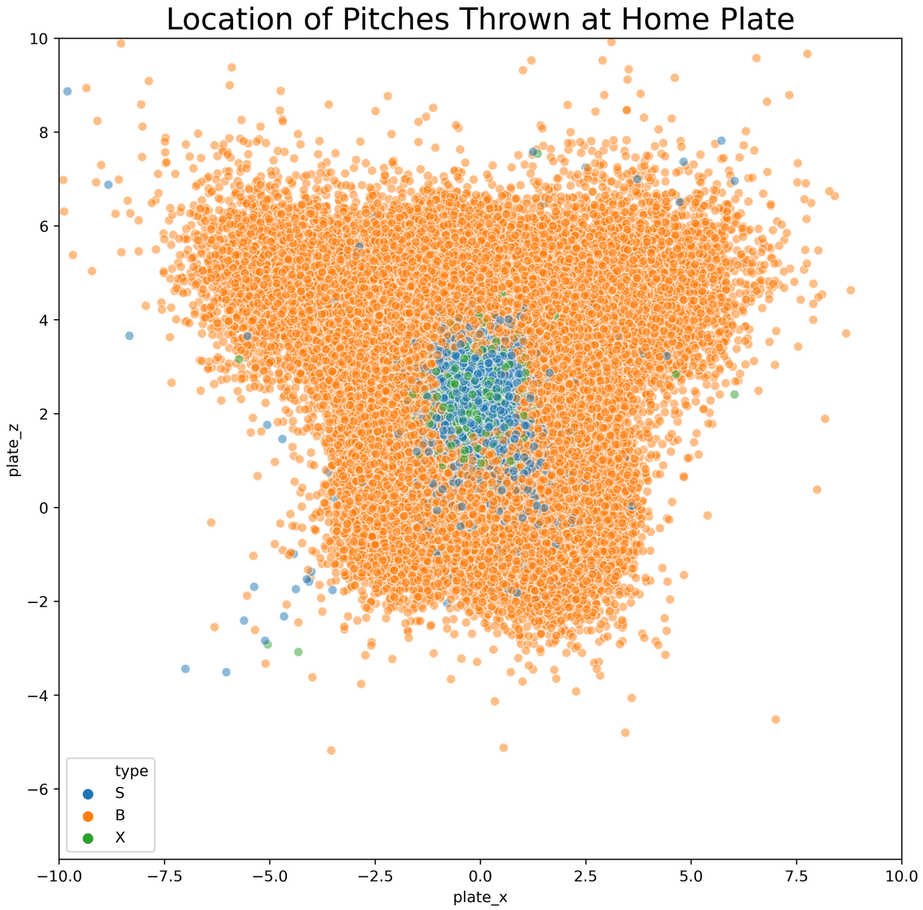
There have only been 175 switch hitters (people who can swing the bat both sides)in the last 6 seasons.

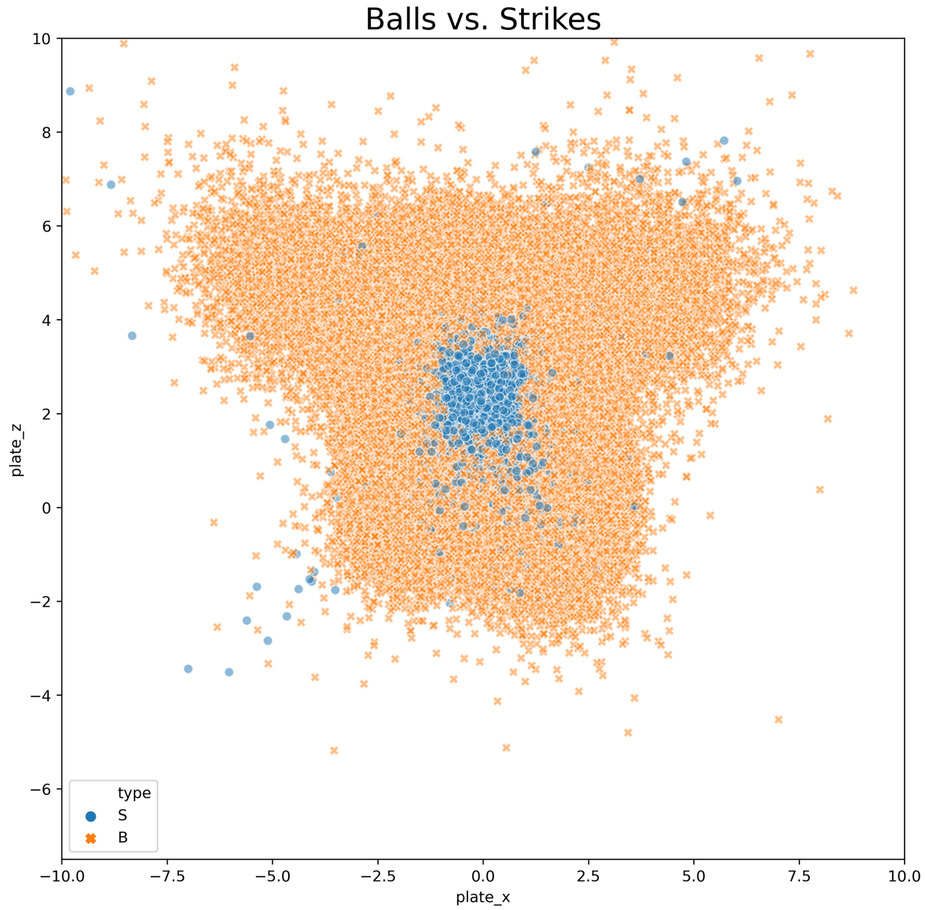
Left handed pitchers tend to throw slower than their right handed counterparts but utilize more offspeed pitches.

STATS ON PITCH NAMES

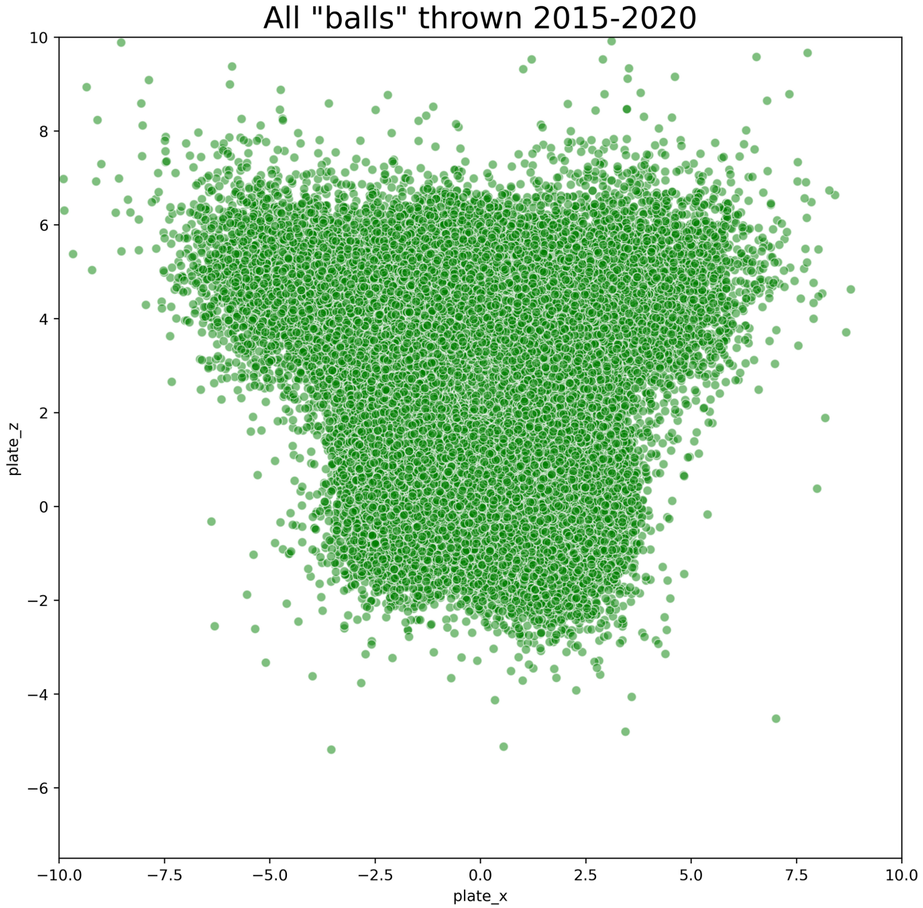
Pitch	Release Speed			Release Spin Rate		
	Avg	Min	Max	Avg	Min	Max
2-Seam Fastball	92.54	63.60	102.50	2158.52	453	3650
4-Seam Fastball	93.26	50.60	105.70	2267.70	454	3660
Changeup	84.25	46.30	97.80	1765.86	501	3690
Curveball	78.21	38.10	98.00	2477.38	466	3637
Cutter	88.53	48.00	101.40	2323.04	428	3599
Eephus	67.10	46.70	84.20	2344.50	574	3052
Fastball	89.70	88.00	92.10	2018.90	1404	2550
Forkball	86.50	78.50	91.90	1514.40	568	3381
Knuckle Curve	80.84	56.90	91.60	2443.55	500	3580
Knuckleball	76.08	50.70	83.20	1535.48	453	3302
Screwball	78.65	71.90	83.20	1963.41	1535	2335
Sinker	92.01	70.30	105.00	2123.82	443	3741
Slider	84.64	45.10	99.60	2345.41	413	3726
Split-Finger	85.03	72.20	96.30	1468.84	506	3673

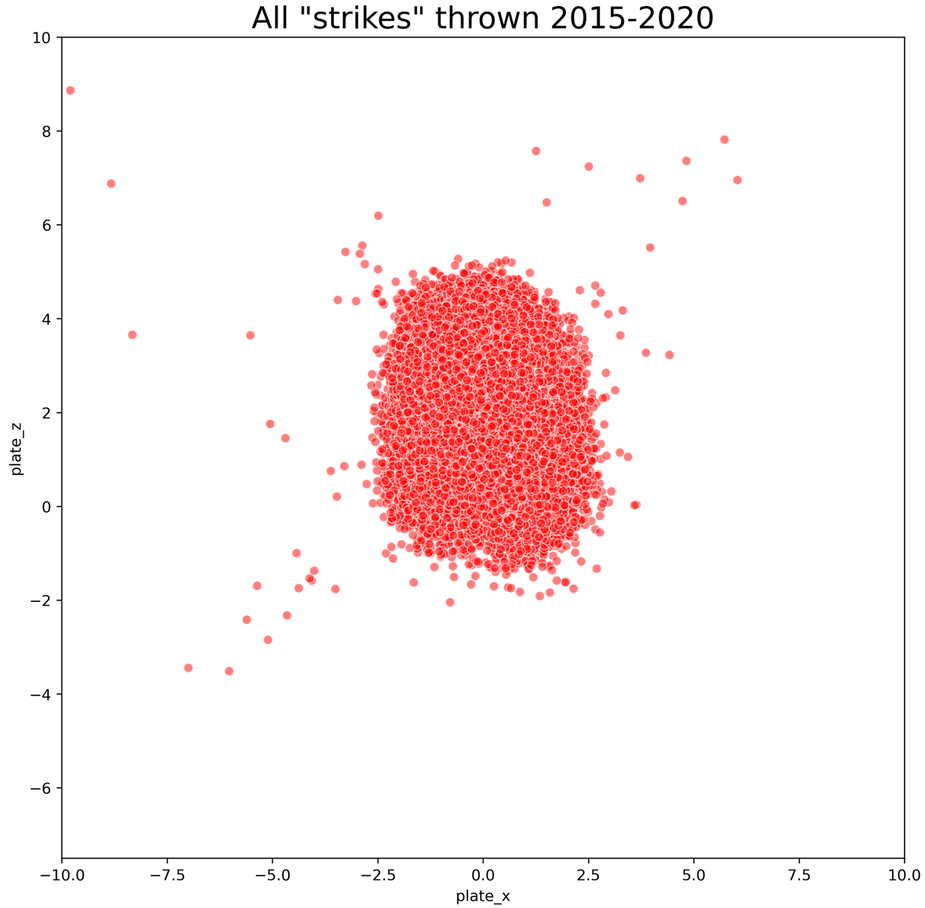
"BALLS" VS. "STRIKES"





"BALLS" VS. "STRIKES"





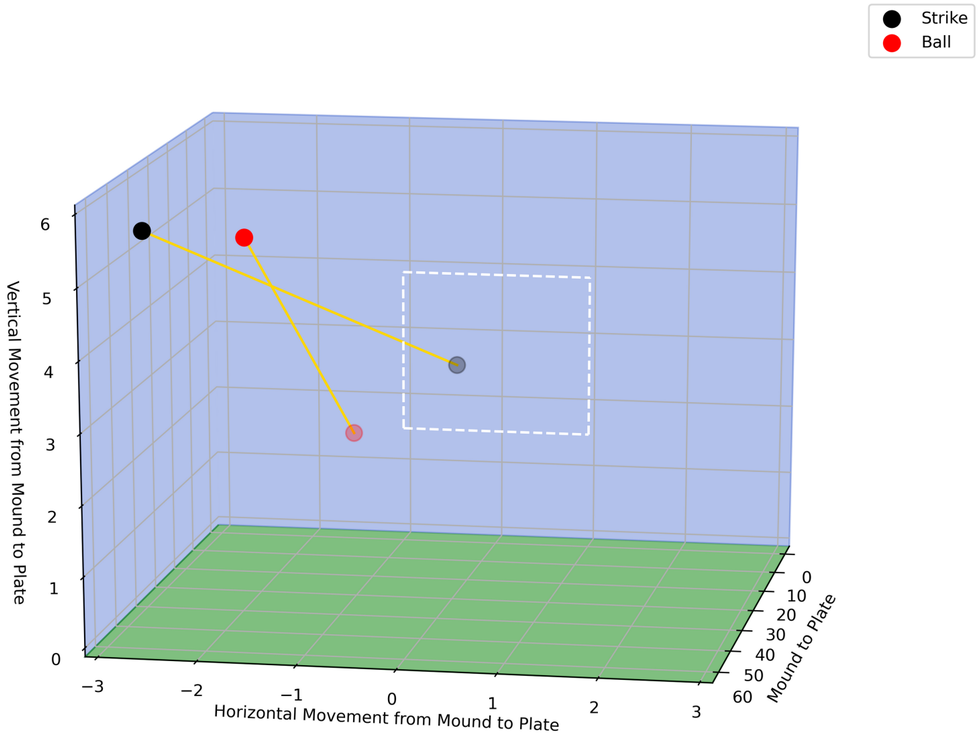
"BALLS" VS. "STRIKES"

Charts aren't everything

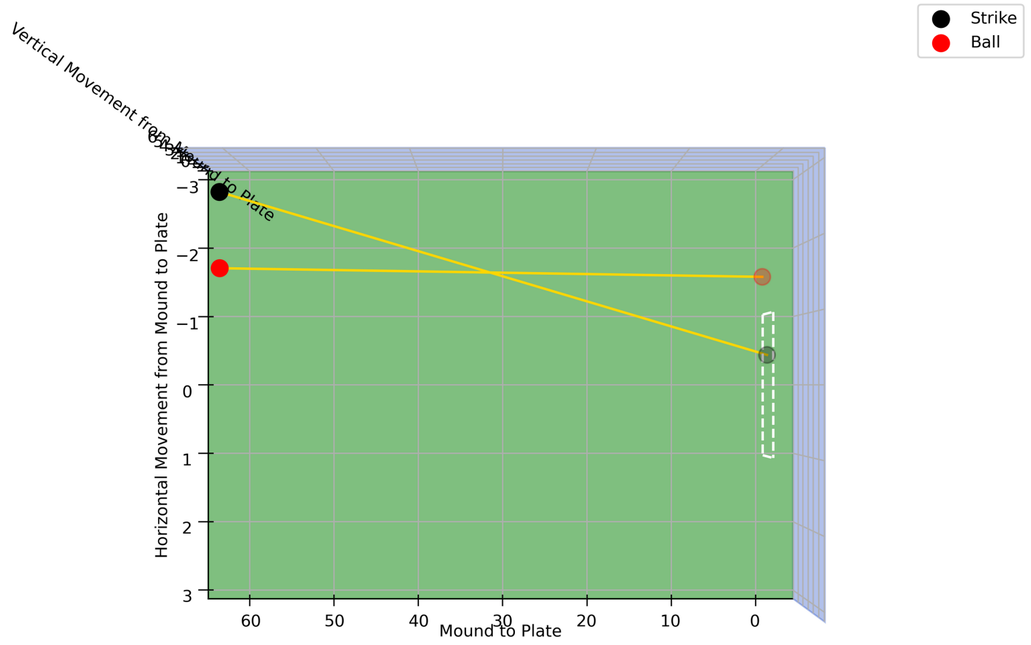
- The charts do not equally show if a pitch was correctly determined to be a strike or a ball
- In the balls chart, you expect to see a hole to show that these calls are actually strikes but they're not
- In the strike chart, there are some strikes that appear to be way outside of the zone and yet are strikes
    - This does not take into account those who swung at really bad pitches

"BALLS" VS. "STRIKES"

TV VIEW

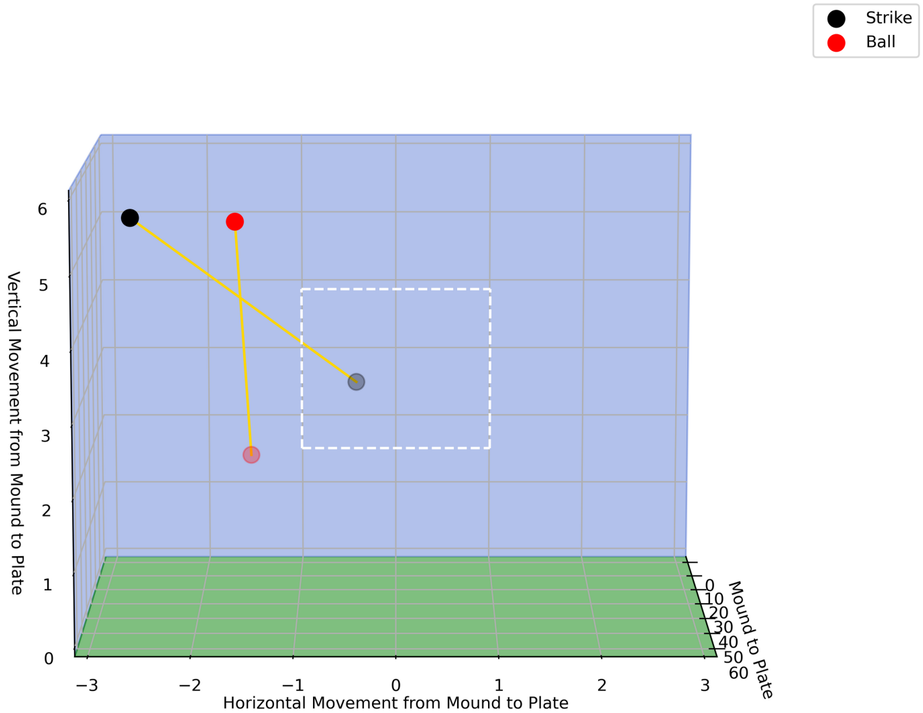


AERIAL VIEW

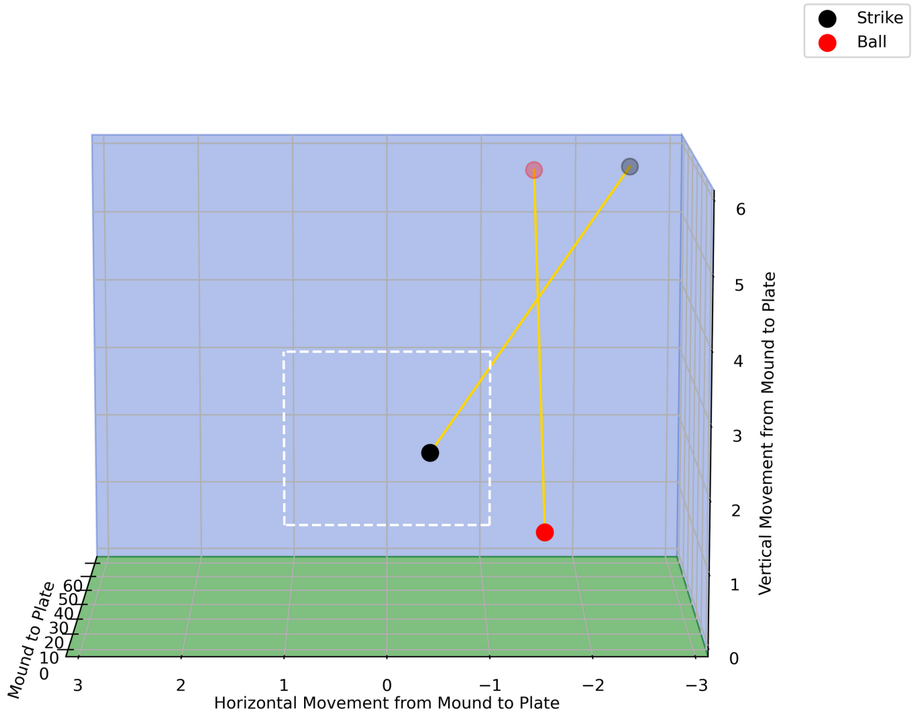


"BALLS" VS. "STRIKES"

Pitcher view

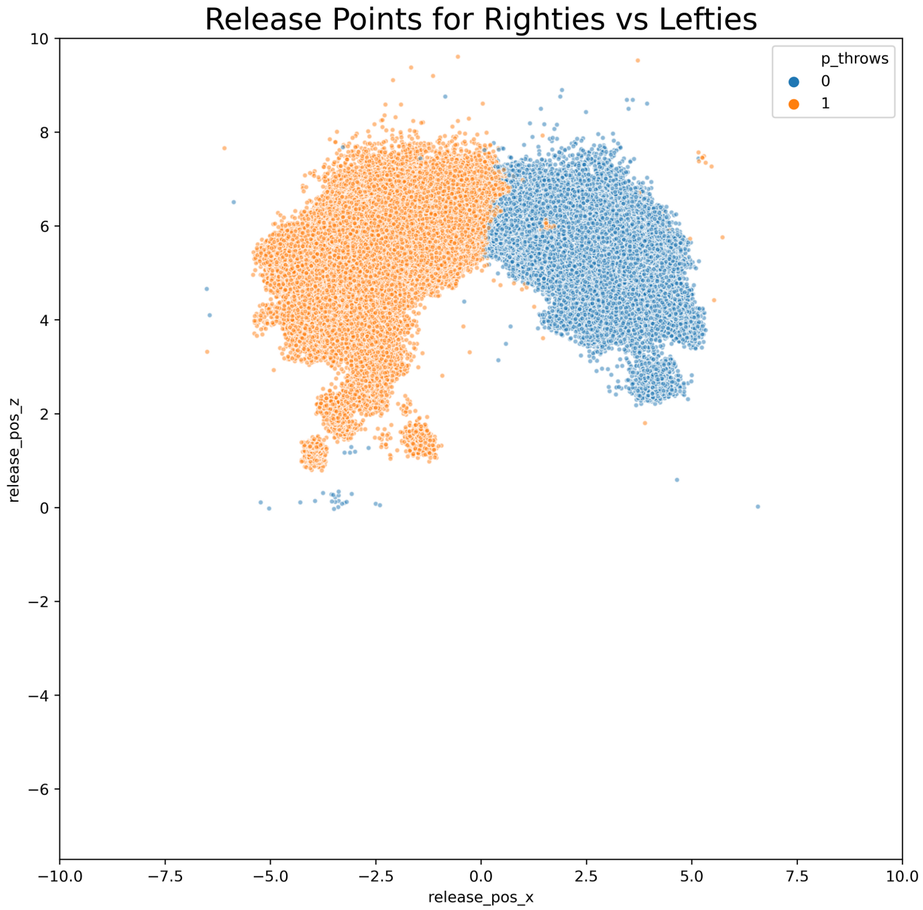


Batter view



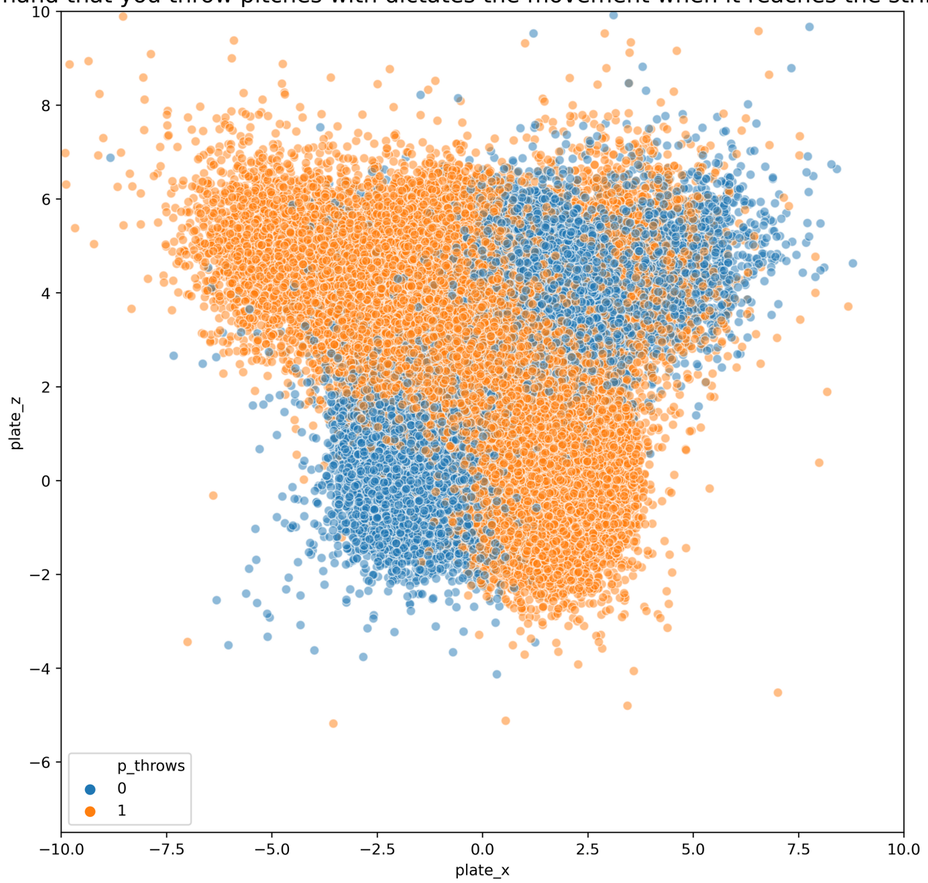
RELEASE POINT TO CATCHER'S MITT

Pitcher



Catcher

The hand that you throw pitches with dictates the movement when it reaches the strike zone



A "SPECTACULAR" MODEL

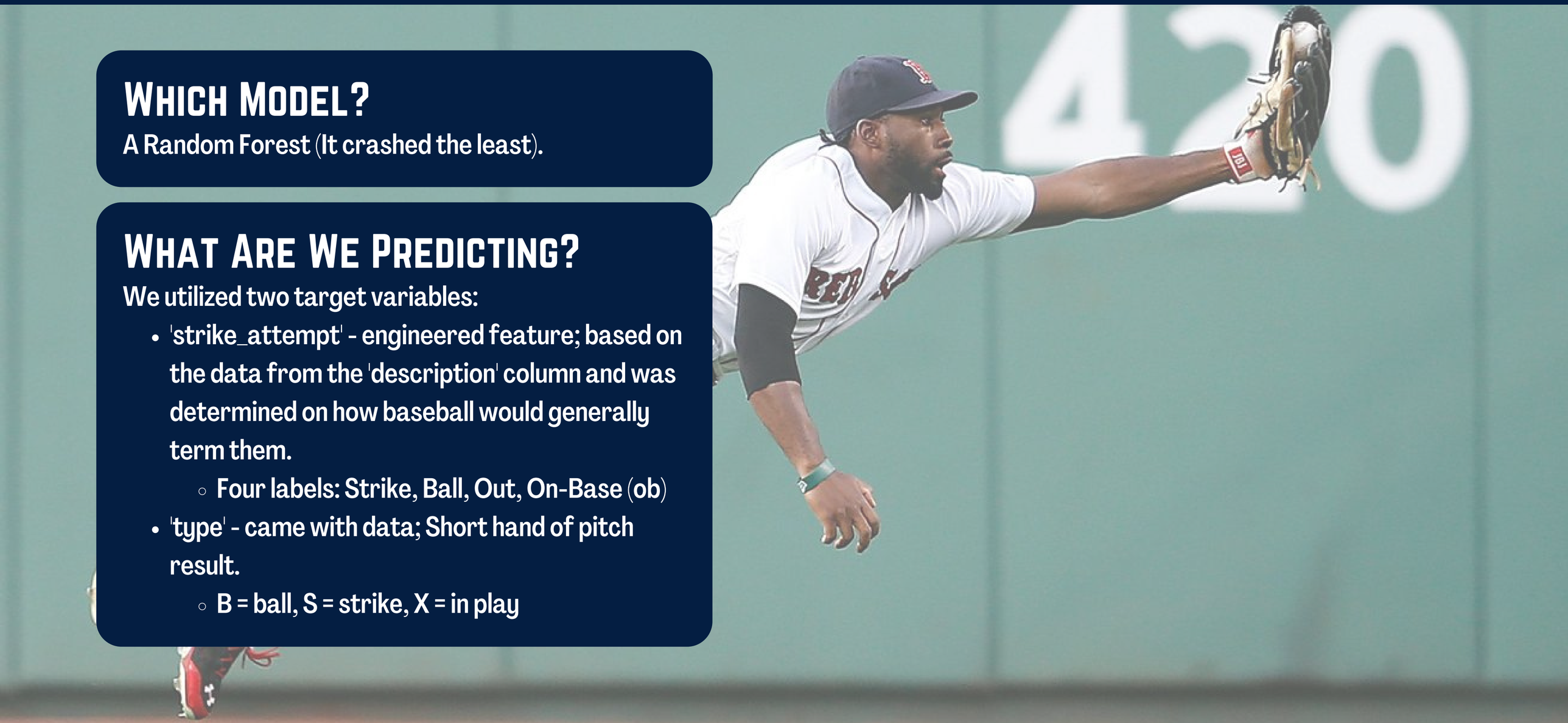
Which Model?

A Random Forest (It crashed the least).

What Are We Predicting?

We utilized two target variables:

- 'strike_attempt' - engineered feature; based on the data from the 'description' column and was determined on how baseball would generally term them.
  - Four labels: Strike, Ball, Out, On-Base (ob)
- 'type' - came with data; Short hand of pitch result.
  - B = ball, S = strike, X = in play



AND THE RESULTS ARE...!

Predicting 'Type'

F1 Score:

- Train: 0.5881085444902757
- Test: 0.5877891754236891

Accuracy:

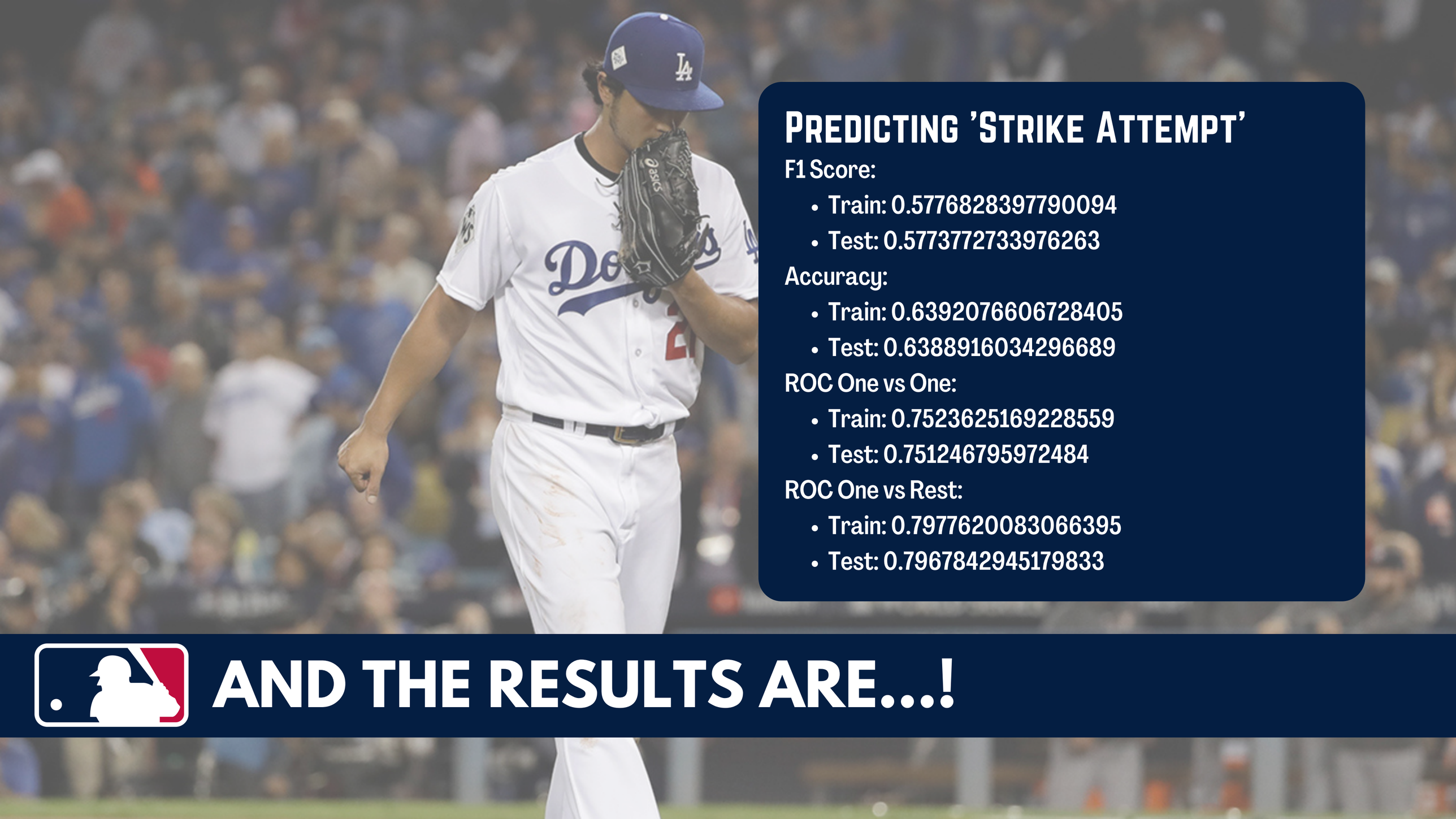
- Train: 0.6500613427421711
- Test: 0.6497017309800389

ROC One vs One:

- Train: 0.7921958017513672
- Test: 0.792270562144838

ROC One vs Rest:

- Train: 0.8069674394244575
- Test: 0.8069223246304112



Predicting 'Strike Attempt'

F1 Score:

- Train: 0.5776828397790094
- Test: 0.5773772733976263

Accuracy:

- Train: 0.6392076606728405
- Test: 0.6388916034296689

ROC One vs One:

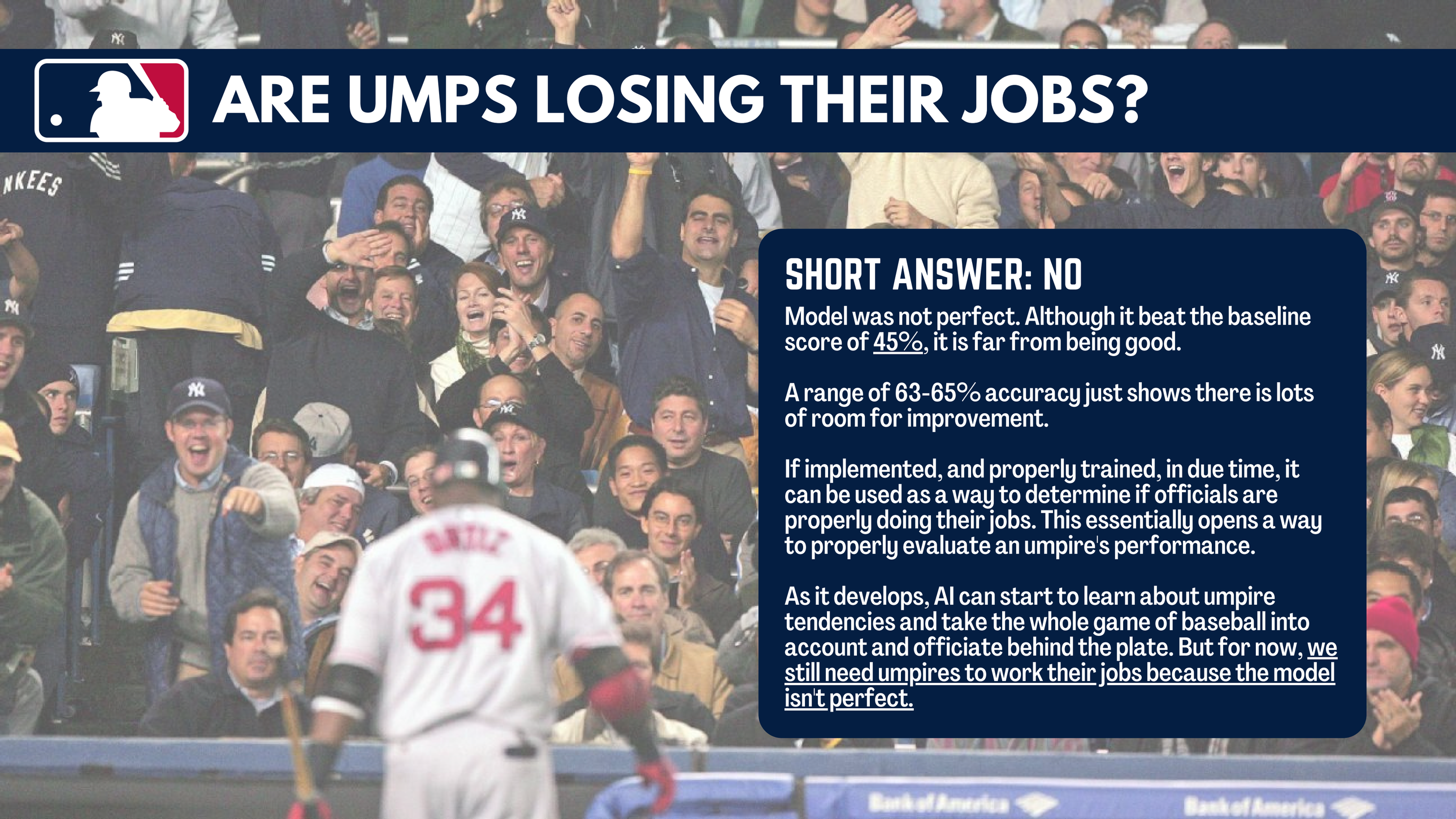
- Train: 0.7523625169228559
- Test: 0.751246795972484

ROC One vs Rest:

- Train: 0.7977620083066395
- Test: 0.7967842945179833

AND THE RESULTS ARE...!

ARE UMPS LOSING THEIR JOBS?



SHORT ANSWER: NO

Model was not perfect. Although it beat the baseline score of 45%, it is far from being good.

A range of 63-65% accuracy just shows there is lots of room for improvement.

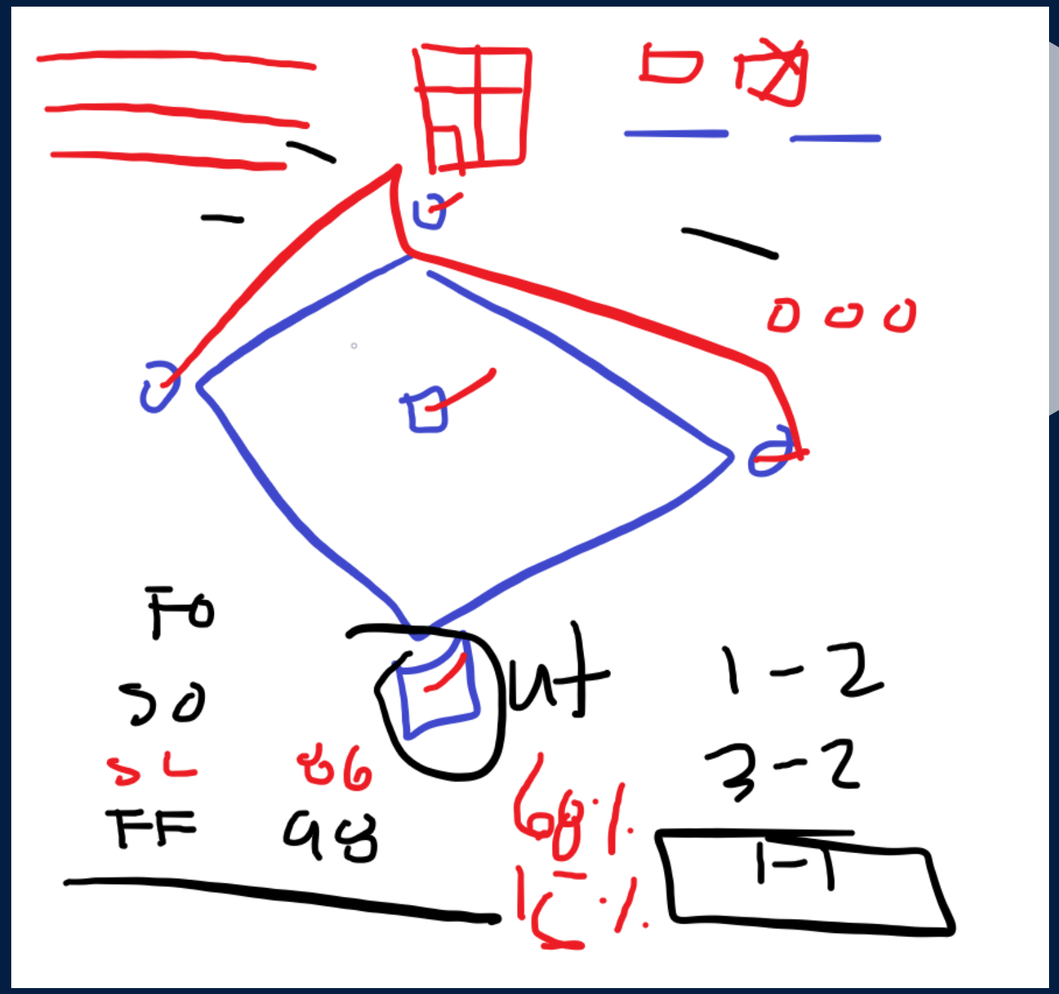
If implemented, and properly trained, in due time, it can be used as a way to determine if officials are properly doing their jobs. This essentially opens a way to properly evaluate an umpire's performance.

As it develops, AI can start to learn about umpire tendencies and take the whole game of baseball into account and officiate behind the plate. But for now, we still need umpires to work their jobs because the model isn't perfect.

NEXT STEPS

- Consider running PCA model to determine the most important features to include in the model
- Explore more tuning parameters
- Determine the best (tree) model possible (since Logistic Regression was awful)
- Utilize Data Bricks and/or AWS (kept crashing due to Memory Errors)
- Build a model that accounts for the batter's attributes, the runners on base, clutch factors, weather, which ballpark they are playing in, etc.
- Start one game at a time, batch at-bats as one row to start predicting outs and not just balls and strikes

END GOAL



Concept Art

Another concept is being able to take this data and simulate a full game that uses everything about the home and away team.

Happy Holidays!

FROM YOUR FRIEND, NADER

x