Pitch-Level Data Mining to Predict Run Scoring in MLB:
A Case Study on Gerrit Cole

Vincent Forca
MS of Data Science 2025
University of Colorado Boulder
Boulder, CO, USA
forca.vincent@gmail.com

1. ABSTRACT:

In this study, we analyze pitch-level data from Statcast for MLB pitcher Gerrit Cole, using a combination of supervised and unsupervised machine learning techniques. We explore patterns in pitch location, batter count, and handedness using clustering and frequent pattern mining, then train a random forest classifier to predict whether a given inning will result in a run being scored. To capture the pitcher's psychological or mechanical state, we engineer novel features such as rolling ball ratios and strike zone adherence. Our model achieves an F1-score of 0.67 on run-scoring inning prediction, offering insight into pitch decision-making under pressure.

2. INTRODUCTION

Pitching performance is critical in determining the outcome of a baseball game. Predicting whether an inning will result in runs scored can inform coaching decisions, defensive alignment, and in-game strategy. This project aims to leverage pitch-level data to forecast inning-level outcomes using machine learning. We study the full season (2023) and extend analysis to 2023-2024 for generalization.

3. RELATED WORK

Prior work in baseball analytics includes strike zone mapping, pitch tunneling, and using machine learning to classify pitch types or estimate run expectancy [1][2]. However, few studies combine pitch clustering and supervised models to predict inning outcomes. Our approach builds on research using Statcast and machine learning [3] by integrating unsupervised and supervised techniques.

4. METHODOLOGY

We took individual pitch data and calculated it a per-inning level.

Modeling: Use k-means clustering, Apriori frequent pattern analysis, random forest for binary classification. **Tools:** Python, pandas, scikit-learn, matplotlib, seaborn, pybaseball

4.1 Data Source: We collected data using the pybaseball package, focusing primarily on Gerrit Cole. Nulls were cleansed for model features.

4.2 Features list for random forest: missed_zone, hit_zone, multiple_runners_on, two_strike_rate, release_speed, called_strike_rate, release_spin_ratio, ball_ratio_5seq, Righty_pct, three_ball_ratio, swinging_strike_rate, ff_pct, sl_pct, lefty_pct, ball_rate, full_count_rate.

4.3 EDA and Feature Engineering see 5.2 for list of engineered feature explanations.

ball_ratio_5seq: I wanted to introduce some sort of feature that can help explain the mental state of our pitcher to better predict run-innings. First, I propose ball_ratio_5seq (see figure 1), which is the ratio of balls within the last 5 pitches. This can clearly show when a pitcher is in trouble, mechanically or mentally (tilted).

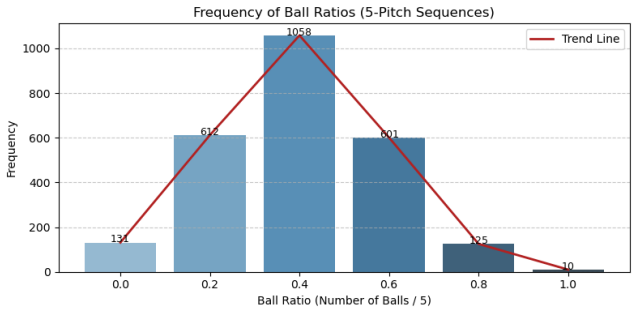

Figure 1. Distribution of ball-ratio-5seq by balls/strikes

First, regular ball and strikes were used, but it became apparent that batters can swing at a poorly placed pitch, which can cloud what we want to see – pitcher performance. To get a better idea of the actual pitched-ball ratio, I defined the strike zone coordinates and counted the ratios within the 5-pitch sequence: **Missed_Zone** –

outside the strike zone coordinates. **Hit_Zone** – inside the strike zone coordinates. The new ball-ratio results are below (figure 2).

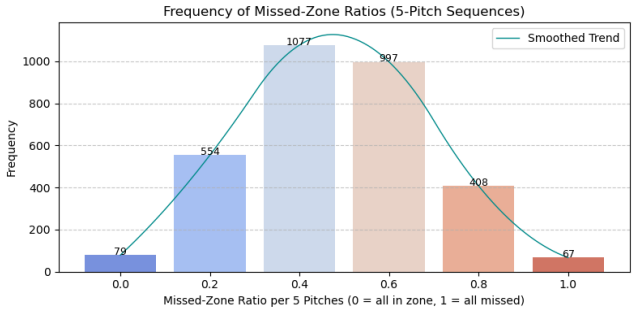

Figure 2. Distribution of ball-ratio-5seq by zone coordinates

Now there are far more counts in the 0.8 ball ratio column than previously, which shows us that our method for countering 'bad swings' worked. The probabilities of missed-zone ratios in a 5-pitch sequence for Gerrit Cole is shown below (see figure 3) based on strike zone coordinates.

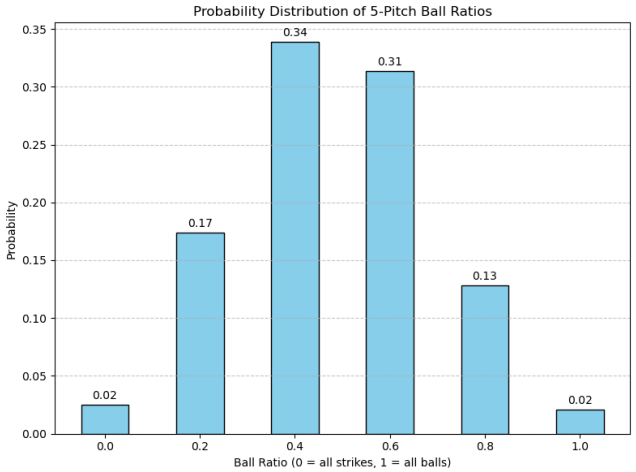

Figure 3. Distribution of ball-ratio-5seq probabilities

To further get a view on Cole's control or mental state, we want to see how his ball_ratio_5seq lined up as his pitch count got higher. We defined **fatigue level** by pitch count: (early <= 30), (30 < mid <=60), and (late >= 60).

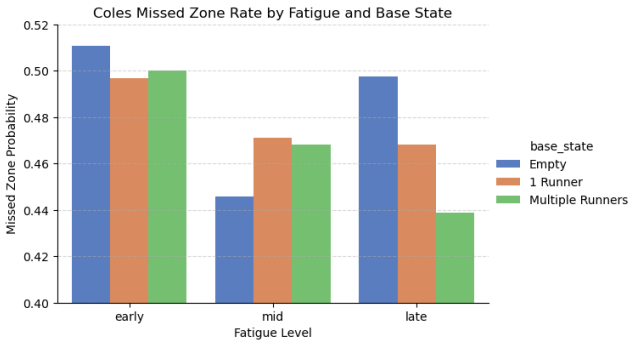

Figure 4. Cole missed zone probability by pitch count and base runners

In the previous illustration (see figure 4), we compared these features to the status of base runners. This suggests Cole has a strong mental fortitude – he has the lowest probability of missing the strike zone during the late-game with multiple runners on. We wanted to then check how the ball ratio looks over the course of all games played that 2023 season.

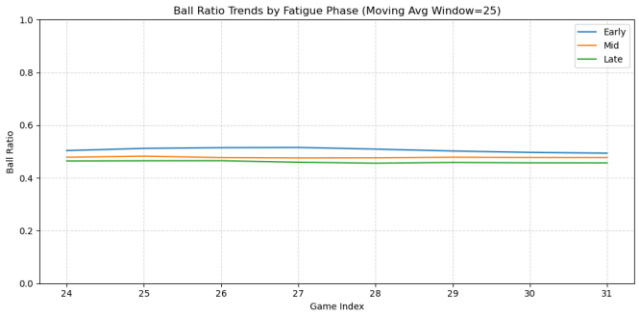

Figure 5. Average of Cole's missed-zone ratio across games by pitch count

The above graph (see figure 5) shows that as pitch count rises; Cole's 5 sequence ball ratio got lower, suggesting that over the 2023 season, he is clutch in the late game.

This is significant to Cole, because when doing the same study on a randomly chosen pitcher, **Blake Snell**, the results of his graph show that his accuracy is best in the mid game. (see figure 6)

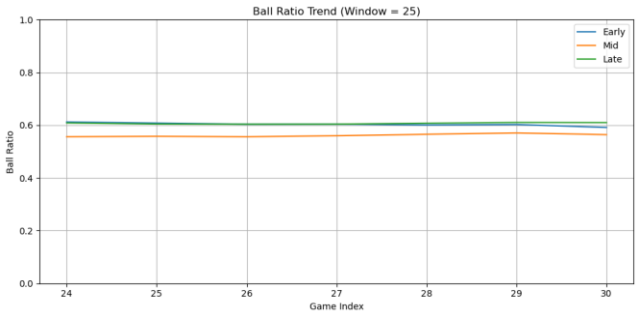

Figure 6. Average of Snell's missed-zone ratio across games by pitch count

Based on this evidence, we can elect to give players like Cole longer outings, while maybe pulling players like Snell earlier. Is this feature a mental predictor for a pitcher to give up runs – or just an explanation of when they perform optimally (early/mid/late)?

Other options for feature engineering I wanted to explore were count pressure: when there's 2 strikes, 3 balls, or full count. When there's runners in scoring position, how do they pitch? I also wanted to look at righty vs lefty batters and explore how those pitches differ.

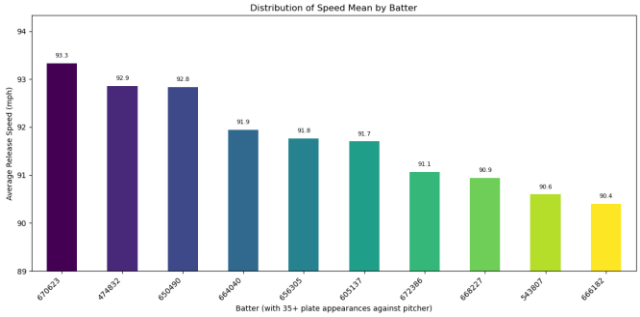

Figure 7. Mean speed VS most frequently seen batters

We can see that his mean speed varies by batter (see figure 7), further research can dive into his specific strategy against individual batters, but for this paper I want to focus specifically on his overall strategy against lefties and against righties and how it may relate to this average speed distribution.

We used clustering to further explore possible pitch behavior features to add to our forest model.

4.4 Clustering: We will use K-means clustering to see how pitches differ when Cole throws to Righty batters vs lefty. (6 clusters)

In the following clustering charts, it shows the strikezone (center rectangle) from the catcher's perspective. So Right handed batters would stand on the left hand side, and left handed batters would stand on the right hand side.

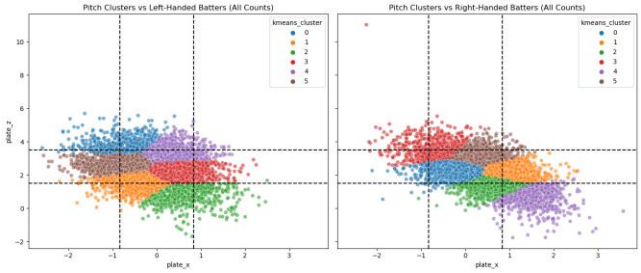

Figure 8. K-cluster of pitches to lefties vs righties (6 clusters)

As you can see (figure 8), the pitch location is more circular for lefties, and more diagonal for righties, and the 6 clusters differ accordingly.

We also wanted to look at cases where there is a full (3-2) count. **How many times is Cole expected to hit the zone with a full count?** Cole is expected to hit the zone about 64% of the time based on the cluster data below. (see figure 9)

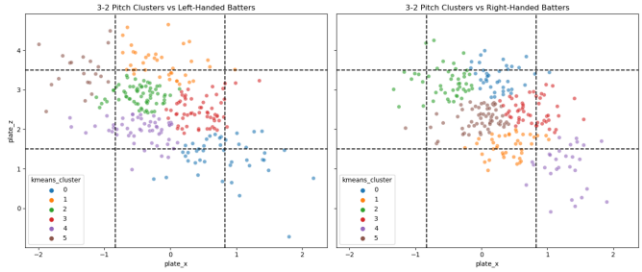

Figure 9. K-cluster of pitches to lefties vs righties on full counts

We also ran the pitch location clusters for cases where there 3 balls and less than 2 strikes (walk potential only), and cases where there are 2 strikes and less than 3 balls (K potential only). (see figure 9)

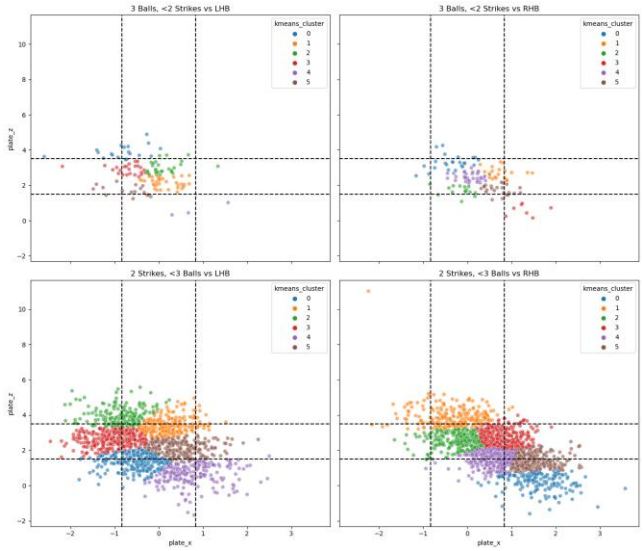

Figure 10. Cluster of pitch locations to lefties vs righties when behind on count (3 ball) and ahead on count (2 strike)

Notice the bottom right pitch cluster against righties (see figure 10) – this is a classic pitch or "down & away" against righties. What pitch is in that blue cluster against righties?

We want to see the different pitches thrown by Cole based on the batter's stance. (see figure 11)

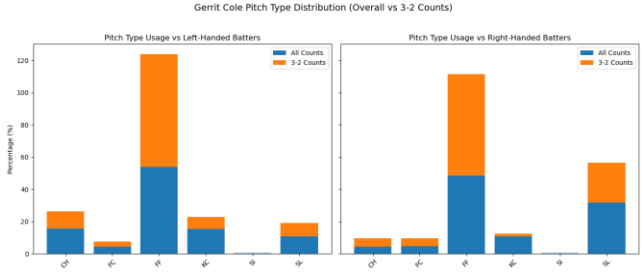

Figure 11. Stacked pitch types by full count vs all counts.

His fastball is the go-to pitch, He utilizes the slider more often against righties, change-ups more often against lefties, and also slightly more knuckle-curves against lefties.

The slider is more comfortable against righties because of where they stand – it tends to slide away from righties. This likely leads Cole to lean more on his knuckle-curve and change-ups for non-fastball pitches against lefties.

4.5 Apriori Frequent Pattern Analysis: We wanted to answer the question – "are there frequent patterns in pitch location vs batter handedness and count?"

To run this analysis, we treated it as a market basket problem: grouping pitches as transactions where the items are: batter handedness, count group (full-count, 3ball, 2strike), and location category (in zone, missed zone).

We obtained some good insight on some strategic patterns Cole uses based on antecedents & consequents. The relationships below were ones with the highest lift.

- (Two Strike Count) → (SL), confidence: 29%
  - When Cole is in 2-strike counts, about 29% of the time he throws a slider — a common put away pitch.
- (SL) → (Two Strike Count), confidence: 89%
  - When he throws a slider, it's almost always with 2 strikes — highly count-dependent usage.
- (Two Strike Count, FF) → (L), confidence: 52%
  - When Cole throws a fastball on 2 strikes, over half are to lefties — suggests he favors heat vs LHB in 2-strike spots (and slider vs righties).
- (SL) → (R), confidence: 70%
  - 70% of Cole's sliders are thrown to righties — classic usage pattern.
- (In Zone) → (FF), confidence: 62%
  - Most of his in-zone pitches are fastballs — typical for attacking early or finishing at the letters.
- (FF) → (In Zone), confidence: 53%
  - And when throwing fastballs, over half land in the zone — confirms he's not just wasting them.

Here is a heatmap (see figure 12 & 13) to further explain the uses of different pitches based on batter-stance and batter-count.

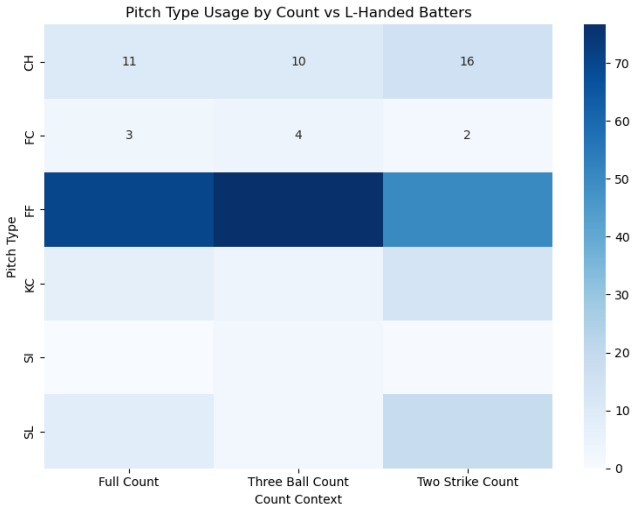

Figure 12. Heatmap of pitch types by batter handedness and count state to lefties

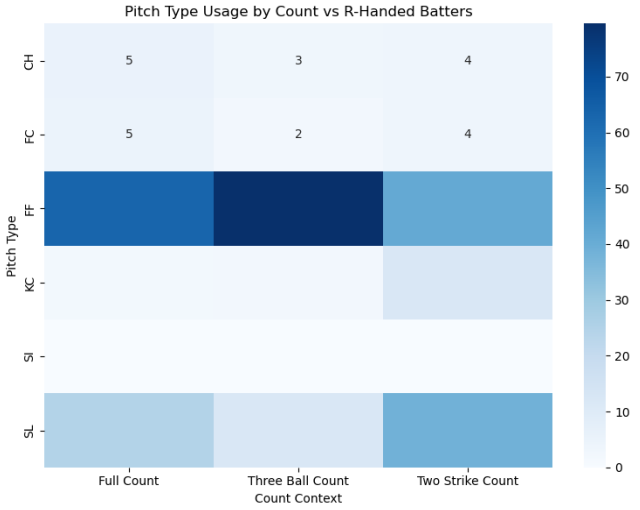

Figure 13. Heatmap of pitch types by batter handedness and count state to righties

He uses the Fastball often, but as you can see, for righties, the slider is a good 2-strike pitch (but riskier when it is full count).

Let's look at strikeout pitches from 2023-2024 seasons (see figure 14). What does he throw when there's 2 strikes, not 3 balls? What does he throw on a full count?

Keep in mind in the following illustrations, right handed batters stand on the LEFT side of the grid, while left handed batters stand on the RIGHT side of the grid.

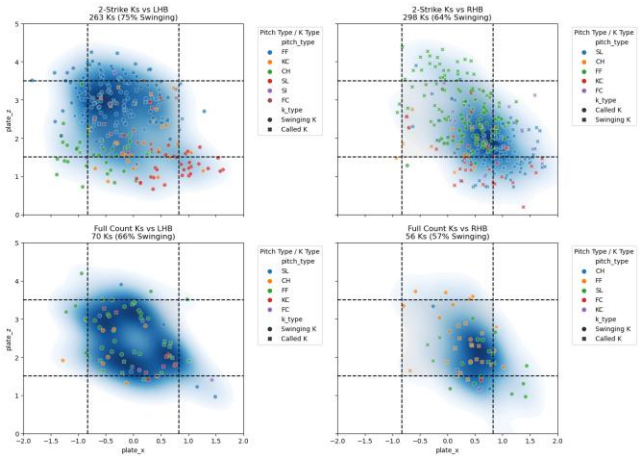

Figure 14. Heat map of strike out pitch types to righties and lefties on full count vs ahead in count (2 strike)

On 2 strikes (no walk risk), his go to strike-out pitch is down-&-away to right-handed batters, while it seems to be up-&-away fastballs for lefties. He does experiment with more SL, CH pitches to the lefties when not at risk to give up a walk.

However, on full counts, he sticks with his fastball more often, especially lefties.

Let's see when he is behind in the count (2023 season only) – 3 balls and less than 2 strikes (see figure 15):

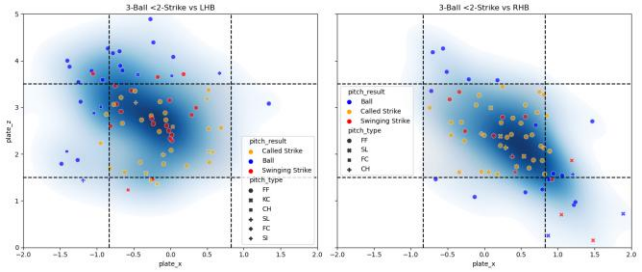

Figure 15. Heat map of pitch types to righties and lefties when behind in the count (3 balls)

Like full counts, when at risk of giving up a walk, he mainly relies on his fastball. From our frequent pattern analysis, we know this is his go-to pitch for hitting the zone.

Based on our clustering analysis and frequent pattern mining, we included the features: missed_zone/hit_zone (where the pitch landed), lefty_pct/righty_pct (percentage of pitches in inning to righty or lefty batter), ff_pct/sl_pct (percentage of fastballs and sliders), full_count_rate/three_ball_rate/two_strike_rate (count pressure).

4.6. Random Forest: we created the random forest model to predict innings where 1 or more runs were let up by Cole in the 2023-2023 seasons.

Y: 1+ run in an inning.

X: feature list (see 4.2)

We set an optimized threshold of 0.20 to maximize recall and F1. (see figure 16)

We prioritized recall(1) over precision because we want to miss less innings where a run was scored.

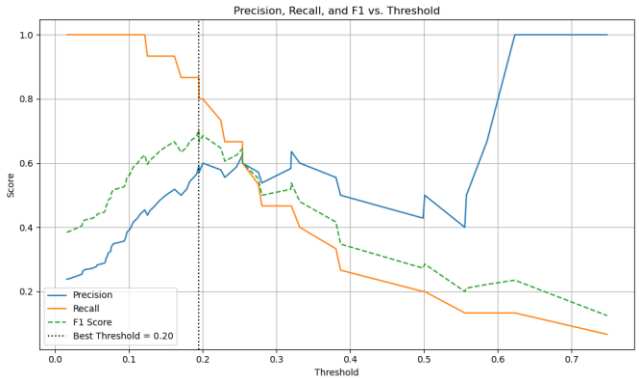

Figure 16. Threshold optimization graph by recall, precision and F1 score.

Precision(0,1): 0.93, 0.60

Recall(0,1): 0.83, 0.8

Accuracy: 0.83 – overall accuracy

F1 Score: 0.67 – combination of recall and precision.

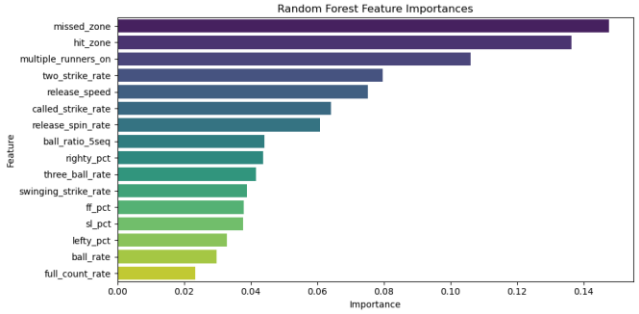

Figure 17. Random Forest Feature importance rankings

Top features	feature importance
missed_zone	0.147629
hit_zone	0.136319
multiple_runners_on	0.106019
two_strike_rate	0.079682
release_speed	0.075213
called_strike_rate	0.064069
release_spin_rate	0.060773

5. EVALUATION

Random Forest (2023, 2024 seasons): Train/test split (80/20)
5.1 Metrics: Accuracy, precision, recall, F1 score.
Target: High recall on true bad outings while minimizing false positives.

Our best-performing model achieves an F1-score of 0.67 on the minority class and an overall accuracy of 83%, while maintaining a recall(1) of 80%, showing strong potential for using granular pitch data to inform outcome predictions. We also increased the ball_ratio_5seq to a sequence of 10 pitches, which improved its impact.

Recall(0) = 0.83 means 83% of true no-run innings were correctly identified by the model as no-run innings. This implies a false negative rate of 17% (i.e., 17% of no-run innings were misclassified as run-innings).

Recall(1) = 0.80 means the model correctly identified 80% of true run-innings. In other words, it missed 20% of innings where runs were actually scored.

Precision(0) = 0.93 means that when the model predicts a no-run inning, it's correct 93% of the time. The false positive rate for this class is 7% — 7% of predicted no-run innings actually had runs.

Precision(1) = 0.60 means that 60% of innings predicted as run-innings were correct. That is, 40% were false alarms — innings where no runs were scored despite the model predicting otherwise.

5.2 Features: explanation of the top 7 features (all features above 0.05 importance level).

The top 2 features were missed_zone and hit_zone, coming in at 0.148 and 0.136. Obviously, both can be strong predictors in determining if runs score or not, but just the pitch's coordinates don't tell the whole story. There are times in baseball when a pitcher may strategically throw a ball as a strikeout pitch – perhaps some other engineered feature can explain these cases.

The 3rd most important feature was multiple_runners_on, which indicates if there are 2 or 3 runners on base, this guarantees a runner in scoring position. If there's multiple runners on base, hitting the strike zone (or not) becomes important. In future research, I would like to see a RISP feature, to account for 1 runner on 2nd base.

The 4th feature is 2-strike-rate at about 0.08, which is the total number of pitches on a 2 strike count divided by the total number of pitches in that inning. We engineered this feature as a predictor for runs not scoring.

The 5th feature is release_speed at about 0.075. This shows the average pitch velocity for the inning. By itself, it does not seem like it would be a strong predictor for a run scoring, but it signifies hidden features, such as fastball rate and off-speed pitch rate which could be a proxy for # of strikeout pitches thrown, or ratio of lefty/righty batters faced.

The 6th feature is called_strike_rate, at about 0.064, this feature is the rate of called strikes thrown in the inning – which differs from our 'hit_zone', since umpires can call strikes as they see fit. This is the count of 'Called Strike' divided by the total pitches in that inning (from the description attribute).

The 7th feature is release_spin_rate at about 0.06. This is the average spin rate thrown for the inning.

6. DISCUSSION

83% accuracy was not the highest we achieved. At one point, one of the models achieved 91% accuracy, but came at a cost of recall(1) being about 40%. For our use case, which would be scouting, coaching, decision making, and fantasy – we wanted to keep the recall high. Future work may include RISP (greater or equal to 1 runner on 2nd base) as opposed to 1 on base vs multiple on base for better mental predictors and may explore recurrent models for sequence modeling or integrating more in-game context (score, opponent strengths). We could also expand our engineered 'fatigue-level' into the model by including "pitches up to this inning", so pitchers that historically do better or worse as the game goes on will have that data point. Future research could also explore the use of LSTM or transformer-based sequence models to better capture pitch progression over an inning.

7. TIMELINE

- Day 1: Data acquisition and cleaning (done)
- Day 3-5: EDA and feature engineering (done)
- Day 5-6: Model training (done)
- Day 6-8: Hyperparameter tuning and evaluation (done)
- Day 9-10: Final write-up and slide deck (done)

8. REFERENCES

[1] Fast, M. (2010). What the heck is pitchf/x? *Baseball Prospectus.*

[2] Brooks, D. (2015). Exploring the Strike Zone. *Baseball Savant.*

[3] Marchi, M., & Albert, J. (2017). *Analyzing Baseball Data with R.* CRC Press.

9. ACKNOWLEDGEMENTS

Thanks to the pybaseball development team and Statcast for public MLB data access.