2024 Cincinnati Reds Hackathon
MLB’s Freaky Friday: Pitcher Role Reversal

The Iowa Hawkeyes

I. Process Summary

Our team used a three step solution to identify pitchers that are currently misused by their organization. Our process began by isolating each unique role across Major League Baseball. After categorizing each pitcher according to their role, we created a model to predict the probability that a pitcher is a starter, given four independent variables:

Arsenal: *The unique number of pitches thrown at least 5% of the time by an individual pitcher.*

MPH_loss: *The correlation between a pitcher's fastball velocity and index of fastballs thrown in each appearance.*

Stuff_plus: *Provided by FanGraphs; A pitch grade based solely on physical characteristics including induced vertical break, horizontal break, velocity, and release point.*

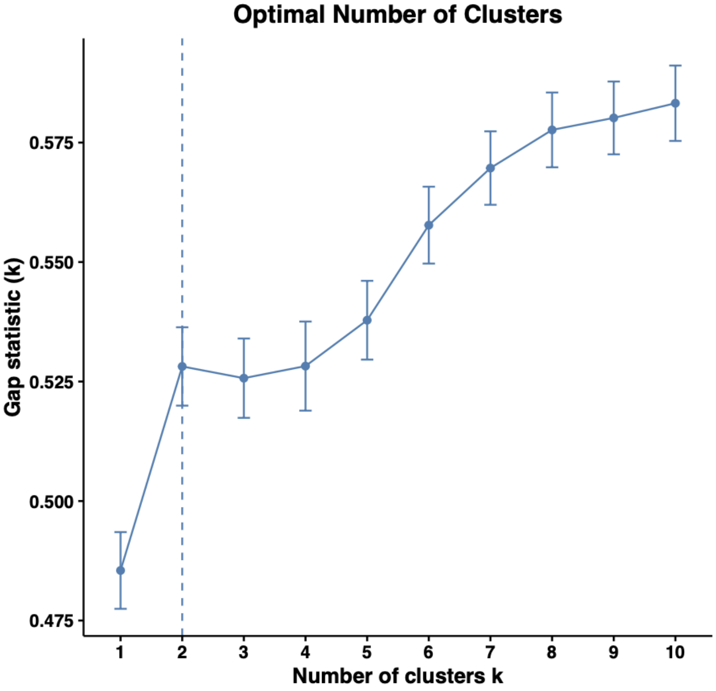
Location_plus: *Provided by FanGraphs; A location grade based on the location of the pitch and count, regardless of pitch metrics.*

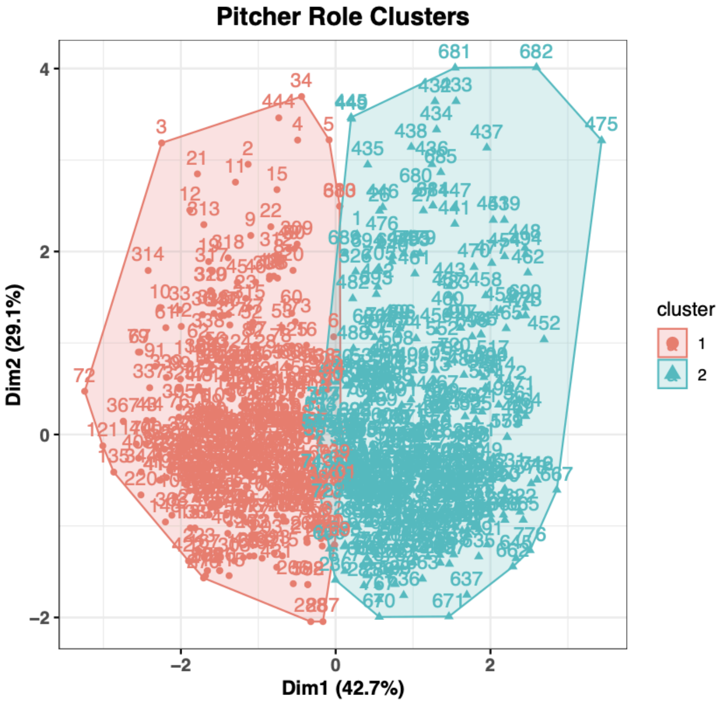
Finally, we compared the expected probability of being a starter with the actual primary role of each pitcher in 2023 and identified three candidates primed to elevate their value in a different role.

II. K-Means Cluster and Role Buckets

To qualify for our analysis, a pitcher must have thrown at least 20 innings and 200 pitches at the Major League level in 2023. This allowed us to remove position players and pitchers lacking a sufficient sample.

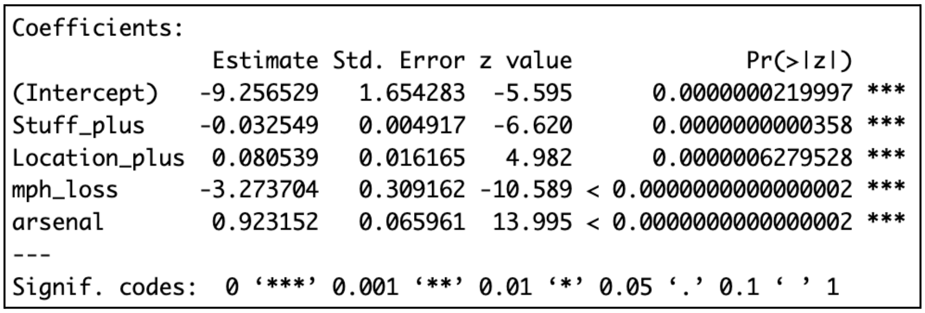
Instead of relying on domain knowledge or a subjective definition of each type of pitcher, we used a K-means cluster analysis to objectively categorize each pitcher. K-means clustering is an unsupervised learning algorithm that aims to minimize the sum of the squared distances between observations to group similar data and illustrate underlying patterns. K-means requires the user to define the fixed number of *k* centroids. We created a data frame containing Arsenal, MPH_loss, and a new variable, batters_faced, which represents the average number of batters each pitcher faces per appearance. Using the Gap Statistic method, we settled on k = 2. This implies there are only two different types of pitchers, colloquially referred to as *starters* or *relievers*. Popular belief suggests there are many types of relievers. However, these distinctions are not defended by data and were not used in our analysis.





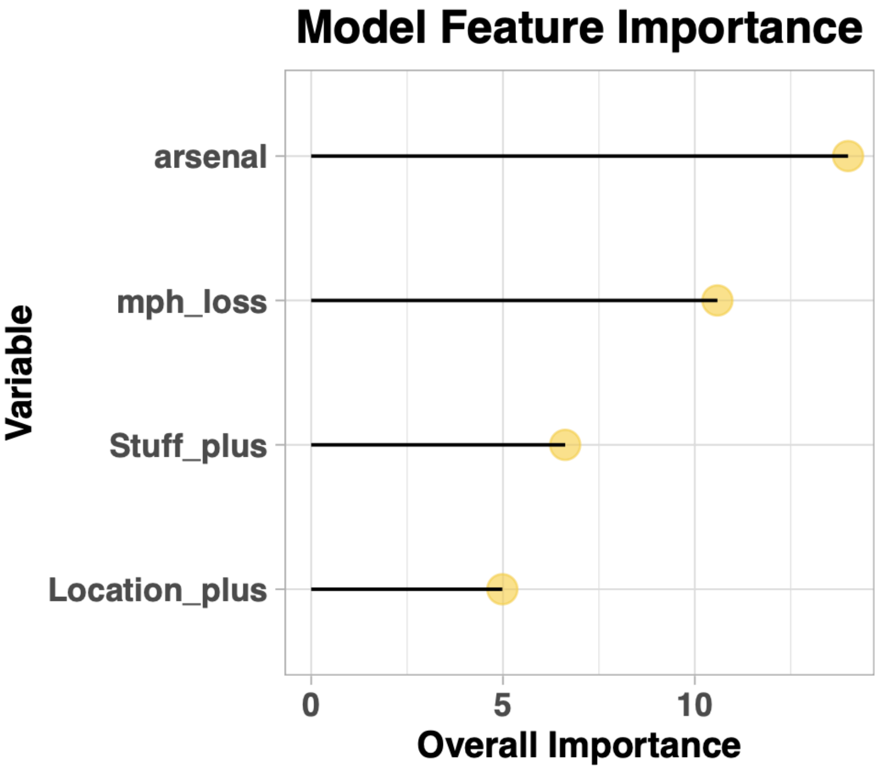
III. Model and Variable Selection

After defining the number of unique pitcher roles across Major League Baseball, we created a logistic regression model to predict the probability that a pitcher is a starting pitcher as a function of Stuff_plus, Location_plus, MPH_loss, and Arsenal. These variables reflect a pitcher's true talent and stamina independent of their role. The best model should identify pitchers as starters or relievers based solely on what they can control. Pitchers cannot control how they are used, so we avoided any variables that could reflect on-field results or performance metrics, which may be skewed by their role. Arsenal was our most significant variable. We did not want to predict which pitchers could develop future pitches or tweak their arsenal because this removes us further from reality. We want to identify candidates for role changes according to their present development, and the current arsenal variable is the best method.



```
Coefficients:
               Estimate Std. Error z value              Pr(>|z|)
(Intercept)   -9.256529   1.654283  -5.595      0.0000000219997 ***
Stuff_plus    -0.032549   0.004917  -6.620      0.0000000000358 ***
Location_plus  0.080539   0.016165   4.982      0.0000006279528 ***
mph_loss      -3.273704   0.309162 -10.589 < 0.0000000000000002 ***
arsenal        0.923152   0.065961  13.995 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We used a logistic regression model because it is easily interpreted and well-suited for categorical dependent variables. We settled on our four chosen predictors by analyzing the significance in the model and visualizing the importance with the *varIMP()* function in the caret package.

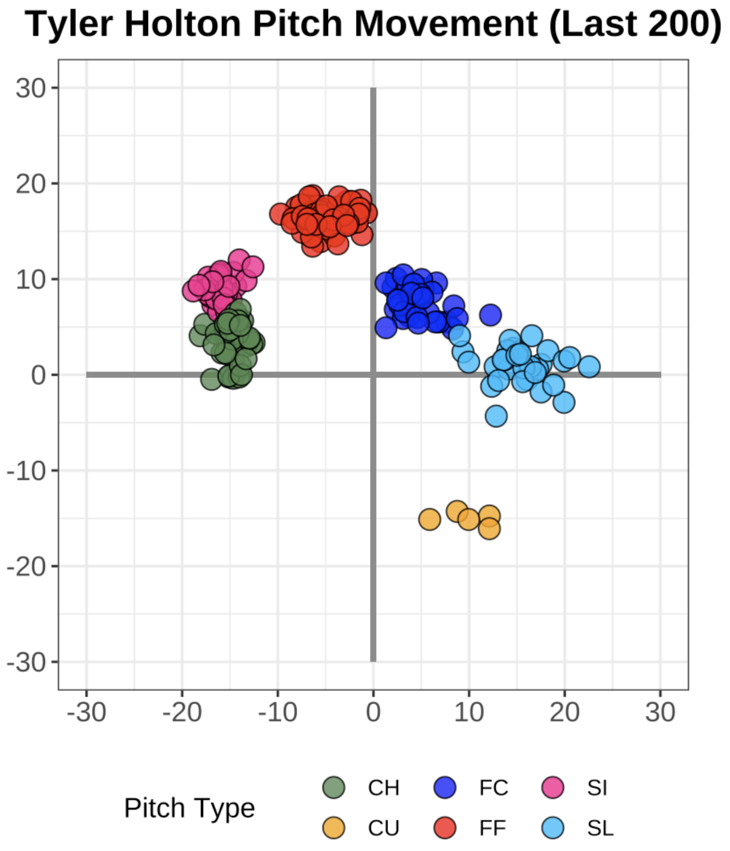


We trained the model on 70% of our data, which contained 1,176 pitchers from 2021-2023. We cross-validated the model on the remaining 30% (503 pitchers). Finally, we applied the model to only qualified pitchers from 2023. This allowed us to identify potential candidates to switch roles based solely on last season.

IV. Candidates to Switch Roles

Connor Phillps, Starting Pitcher, Cincinnati Reds
The Reds' #4 prospect debuted last season with a high-stuff arsenal, but our model suggests he profiles better as a reliever. His three-pitch mix features excellent movement, but his location was sub-par across five starts. Given Phillips' age, whiff rate, and fastball velocity, it's tempting to use him as a starter. However, he could maximize his strengths (velocity, movement, minimizing contact) in the bullpen with shorter, high-leverage appearances.

Tyler Holton, Relief Pitcher, Detroit Tigers
Holton is another rookie, but he found success thanks to his excellent control, ability to induce weak contact, and large arsenal. These strengths led our model to suggest him becoming a starting pitcher. All six of his pitches grade above average with clearly defined movement profiles. Add this with his plus command, and Holton could be a valuable left-handed addition to an unproven Detroit rotation.



Michael Kopech, Starting Pitcher, Chicago White Sox
Kopech is unique because he has pitched as both a starter and reliever in his career. Since 2022, the White Sox have used him as a starter, but our data suggests he would be more valuable as a reliever. Kopech's elite extension leads to excellent velocity and above average stuff scores. He uses his fastball 61% of the time, which is a problem given his poor location and average movement profile. As a reliever, the usage imbalance would be less important and he could continue to rely on velocity to generate whiffs in short outings.

Given the lack of starting pitching depth on the White Sox, Kopech likely adds more value to the roster as a below average, misused starter than he would as a reliever. This presents a situation in which our model could be used to identify potential trade pieces. Kopech is a valuable arm for any team, and if other teams identify his misuse they should try to acquire him, or any other misused pitcher, to maximize talent in the correct role. The White Sox, in need of starting pitching, should willingly trade Kopech, a true reliever, for pitchers that profile as starters. If the White Sox could more efficiently find starting pitching on the trade market, they should part with Kopech, who will never reach his true talent ceiling given their need to use him as a starter. If pitcher roles are an inefficiency across Major League Baseball, as our model suggests they are, the teams that are able to identify misused pitchers will acquire pitchers at low prices and maximize their performance in the optimal role.

V. Limitations and Conclusion

Our model is built under the assumption that most pitchers are used in their optimal role. By assuming pitcher roles are accurate at the aggregate, we can easily identify outliers as candidates for a new role. It's reasonable to assume that most pitchers are used correctly, however this assumption causes our model to predict the role of a pitcher from their similarity to other pitchers in the same role. We are not identifying what makes a pitcher *perform* best in each role. Instead, we are predicting what makes a pitcher *similar* to other pitchers in the same role.

We do not have any variable reflecting how performance changes with each time through the order. This variable would require a pitcher to pitch multiple times through the order and could overlook relief pitchers that have never been used multiple times through a lineup. Nevertheless, we believe the best starting pitchers will be successful multiple times through the order. We created the MPH_loss variable to reflect stamina and bypass an explicit time through the order requirement.

Our model identifies pitchers that are presently misused by their organization and could benefit from a role change. Potential uses for a team could include roster analysis to build the starting rotation around the top five candidates. Another potential insight could be evaluation of the trade market to identify misuse cases. By bucketing pitchers into two roles, starters and relievers, our process is extremely actionable. A front office member could tell a coach that Pitcher X profiles as a reliever instead of a starter in very simple terms. With each additional bucket, roles become more blurry and unreliable. It's difficult to advocate for a switch from closer to set-up man, or any other role under the reliever umbrella, without encroaching on a coach's responsibility or drawing insignificant distinctions between roles. When the data support only two divisions of pitchers, this becomes an even harder sell. Overall, our cluster process and model combine to effectively identify misused pitchers in an interpretable and actionable manner.