MLB Relief Pitcher Categorization and Analysis

Malcolm Gaynor

May 7, 2024

Professor Brad Hartlaub

Stat 306

Introduction

Relief pitchers are considered one of the most volatile positions in the game, likely due to their relatively small sample size, which often consists of high leverage situations. Also, even mediocre or bad relief pitchers are often just one small tweak away from finding huge success, which is why many relief pitchers who are unceremoniously traded, waived, or cut from one team end up finding great success on another team, even just a year later. For example, in the 2024 offseason, the Los Angeles Dodgers designated left handed reliever Bryan Hudson for assignment after he lost his roster spot due to multiple big name free agency and trade moves. A few days later, the Dodgers traded Hudson to the Milwaukee Brewers for an 18 year old prospect who was picked in the 20th round of the 2023 MLB draft. Despite putting up a 7.27 ERA in a small sample of innings in 2023, Hudson has been incredible for the Brewers in 2024, putting up a 0.93 ERA so far. Hudson has made a major change to his pitch selection, throwing his cutter and his sweeper 24% of the time each, after using them a combined 13% each last season. It remains to be seen if Hudson can continue this level of success, but nonetheless he demonstrates how a small change in pitching strategy can turn a below average pitcher into a dominant reliever.

Therefore, the primary goal of this analysis is to consider which relief pitchers may potentially benefit from changing their pitching style. To do so, I will cluster MLB relief pitchers based on their pitching style. Then, I will create a supervised learning model to predict ERA for pitchers in each cluster. Therefore, there will be a different model for each cluster. Finally, I will apply each of the models to all of the relief pitchers, to get a predicted ERA for each pitcher if they were in every cluster. Then, I can look into some case studies of players who the model predicts would have a better ERA if they transitioned their pitching style into a different cluster.

This analysis was inspired by a paper titled *Using Clustering to Find Pitch Subtypes and Effective Pairings* by Gregory Dvorocsik, Eno Sarris, and Joseph Camp, PhD, which also uses clustering to advise pitcher repertoire decisions. However, I will take their analysis a step further, by combining supervised learning methods with clustering.

Data

The data for this project was collected from Fangraphs and Baseball Savant, and includes relief pitcher seasons from 2018 to 2023 (excluding the shortened 2020 season) where pitchers threw at least 40 innings with at least 500 total pitches. Relief pitchers were defined as pitchers who appeared in over half of their total outings out of relief. This led to a dataset of 1,086 total relief pitcher seasons. For the clustering, the majority of the data was about the pitch repertoire, such as the percentage each pitch is used, and the average velocity and spin rate of these pitches. Outside of this data, the only other variable considered was the percent of pitches that were in the strike zone, according to Stacast. This data was chosen because it only considers aspects of a pitcher's strategy that are completely under his control. Data such as ground ball percentage and strikeout rate is certainly impacted by the pitcher and can reflect his pitching style, but involves influence from the hitter as well. Therefore, the clusters were made keeping in mind elements that are solely related to the pitcher himself, and not the pitched ball outcome. However, for the supervised learning techniques to predict ERA, more variables are considered, along with the variables already used in the clustering. These variables include the following percentages: strike out, walk, ground ball, fly ball, line drive, infield fly ball, pull, oppo, hard hit, soft hit, chase, swing, swing at strike, contact, in zone contact, out of zone contact, first pitch strike, swinging

strike, called strike, and total strike. The data also includes ground ball to fly ball ratio, and home runs per nine innings.

Clustering

Using K-Means Clustering, the 1,086 relief pitchers were clustered into 7 different groups, a number selected using the elbow method to reasonably limit the within-cluster sum of squares (WSS), which is shown in Appendix 1. The K-Means Clustering algorithm used is the standard Hartigan-Wong algorithm, which seeks to minimize the variance within each cluster by considering the sum of the Euclidean distance between each observation in the cluster and the cluster's center. For more information about K-means clustering, look at the attached R code, or read Chapter 20 of *Hands-On Machine Learning with R* by Bradley Boehmke and Brandon Greenwell.

It is difficult to qualitatively describe every difference between the clusters, but I will attempt to give a basic idea about some of the differences. However, keep in mind that 31 different variables were used to create these clusters, and these are just a few brief highlights that stick out upon inspection of the group means and medians.

Cluster	# players	Description
1	82	Most curveballs, second most 4 seam fastballs
2	251	Most sinkers and most sliders
3	253	Hard slider and slow curveball
4	140	Most 4 seam fastballs with highest velocity and spin rate
5	185	Lots of cutters, curveballs, lowest spin rate
6	88	In the strike zone more, lots of cutters
7	87	In the strike zone more, most sweepers, lots of sinkers

It is important to remember that these clusters do not take into account player performance, just strategy. Therefore, for example, cluster 2 includes dominant, well-known sinker/slider pitchers like Josh Hader (who is in cluster 2 all 5 years of the data), along with lesser known relievers with similar repertoires who lack Hader's level of success, such as Josh Kuhnel and Kirby Snead.

Gradient Boosting

Now, gradient boosting algorithms (specifically XGBoost) will be applied to the relief pitchers in each group separately, now considering even more data that goes beyond the outcome independent variables considered in the clustering. Gradient boosting is the method chosen here due to its use of many shallow trees that can combine to become powerful predictive tools, especially considering a dataset with many variables, as gradient boosting is able to handle complex interactions between the almost 70 variables considered.

The protocol used for this gradient boosting analysis is to first split each cluster's data into training and testing groups, utilizing a 70/30 split. Then, cross validation gradient boosting (xgb.cv() in XGBoost) will be used to find the optimal number of iterations for the *nrounds* parameter in order to minimize the RMSE. Then, the gradient boosting model will be created (using the xgboost() function). Both of these steps in the process involve considering just the training data. Then, the xgboost() model is applied to predict ERA in the testing data in order to calculate the RMSE. In terms of the RMSE, it is important to note this metric is in terms of the response variable (ERA), which ranges from around 0.67 to around 8.82 in our data set. Finally, variable importance plots are created to investigate which variables were most influential in the model. A summary of the gradient boosting models is below. For more details about the

procedure, or the XGBoost hyperparameters used in this analysis, examine the corresponding R code. More background information on gradient boosting can be found in chapter 12 of the textbook by Boehmke and Greenwell.

Again, it is important to note that the following is just a surface level description of the models, they include a large number of shallow decision trees which are "boosted" to increase predictive ability. The Variable Importance Plots with the 10 most important variables are included in Appendix 2.

Cluster	RMSE	Most important variable	Impo-rtance	2nd most important variable	Impo-rtance	3rd most important variable	Impo-rtance
1	0.833	HR/9	0.15	Called strike %	0.06	Fastball spin	0.06
2	1.003	HR/9	0.15	K %	0.1	BB %	0.07
3	0.884	HR/9	0.3	K %	0.15	Strike %	0.06
4	0.976	HR/9	0.23	K %	0.09	In zone contact %	0.08
5	1.006	HR/9	0.2	K %	0.13	Hard %	0.05
6	0.959	HR/9	0.38	Swing strike %	0.04	Fastball %	0.04
7	0.837	HR/FB	0.2	BB %	0.17	Line drive %	0.06

The RMSE values of around 1 are not incredible, but they are solid, and will lead to a useful model. Considering that the minimum ERA in our data is 0.67, and the maximum is 8.82, an RMSE of 1 means that the average difference between the predicted ERA and the actual ERA for relievers in the testing groups was about 1 run, just under 13% error considering the range of the response variable. The RMSE value of around 1 earned run is something to keep in mind, especially when interpreting specific results. For example, if our model predicts that changing

from one cluster to another would lead to a 0.5 run decrease in ERA, that is not strong evidence, as the RMSE is larger than that difference.

In terms of variable importance, one major factor that stands out is the fact that limiting home runs appears to be the most important factor in decreasing ERA, regardless of the cluster. This makes sense, as home runs directly correlate to runs scored. However, the importance of home runs is not overwhelmingly large, and the variety in which variable is the second or third most important implies that the models to predict ERA are at least somewhat different across clusters, which is a good sign that the models may be able to predict which pitchers have underlying statistics that imply they could have a lower ERA if they pitched with a strategy corresponding to a different cluster.

To pursue this question, the next step is to apply all 7 models to every relief pitcher season in the dataset. After doing so, we find that our models predict that 334 relief pitcher seasons could have had a decrease of over 1 run in their ERA if they had a different pitching style. In other words, over 300 relief pitchers put up stats in the past five years that, according to the gradient boosting models, would correspond to an ERA decrease of 1 run or more if they were in a different cluster than the one they are currently in. This is an interesting finding, but, especially considering how complex both the clustering and the supervised learning methods were, it makes more sense to dive deeply into a few case studies instead of looking broadly at all 1,086 of the data points.

Case study 1: Corbin Burnes

The biggest discrepancy between actual ERA and predicted ERA in the gradient boosting model of a different cluster was for Corbin Burnes, who put up an abysmal 8.82 ERA working

primarily out of the bullpen for the Milwaukee Brewers in 2019. Burnes was classified in cluster 5 for this season, likely due to his usage of fastballs, curveballs, and the occasional cutter. However, just one non-COVID season removed from the worst ERA in our dataset, Corbin Burnes was the National League Cy Young winner in 2021, and has not failed to make an all-star game since. He has made major, visible changes to his pitching style since 2019, as he no longer throws the 4 seam fastball that he relied so heavily on earlier in his career, instead primarily using the cutter. Unfortunately, because Burnes became a starting pitcher, he is only included in the dataset one time, and therefore it is unknown which cluster he would have been in in 2021 or later. However, the XGBoost model for cluster 6 (which includes pitchers who throw the most cutters at a high spin rate, which reflect his current pitching style) predicted that Burnes would have had a 3.91 ERA if he chose that pitching style in 2019, an almost 5 run decrease from his actual ERA of 8.82.

The fact that the player who this method determined had the chance for the most extreme decrease in ERA was an eventual Cy Young winner is strong evidence for this analytical approach. In conclusion, despite an ERA of 8.82 as a cluster 5 pitcher (with an XGBoost predicted 7.94 ERA in that cluster), this analysis predicted that Burnes could have an ERA of under 4.00 if he were in a different cluster. In other words, the gradient boosting models not only were able to correctly predict Burnes' struggles while attempting to pitch using the style of a cluster 5 pitcher, it also correctly predicted his massive improvement after changing his pitching style.

Case Study 2: Shelby Miller

While Burnes' 2019 season was the worst ERA in the dataset, Shelby Miller's 8.59 ERA in the same season was not much better, coming in as the third worst. However, unlike Burnes, Miller is included in our dataset again, as he made an incredible comeback in 2023, putting up a sparkling 1.71 ERA after failing to log over 13 innings since 2019. In 2019, Miller was a cluster 5 pitcher. However, the gradient boosting model for cluster 7 predicted that his underlying stats could have corresponded to an ERA of 4.30, almost 4 runs lower than his actual ERA. In 2023, Miller was a cluster 7 pitcher, utilizing a sweeper and splitter that were not included in his 2019 repertoire. Therefore, again, this method of gradient boosting models combined with clusters of pitchers was able to predict Shelby Miller's huge improvement.

It is important to note that, despite the fact that Miller and Burnes represent great areas of success in the ability of this type of analysis to highlight pitchers who potentially could break out in a huge way, the model failed to predict the level of success reached by these two. That is, the model predicted an ERA in the upper 3.00s for Burnes and Miller, who went on to put up sub 2.50 ERA seasons. However, it is unreasonable to expect these gradient boosting models to take into account every adjustment made by Miller and Burnes that lead to their success. Changing pitching styles was certainly an important element in their success, but was likely not the only adjustment they made, which explains why they overachieved even the most optimistic XGBoost predictions. Regardless of this limitation (which is a limitation of every predictive machine learning model), Burnes and Miller are great examples of how this analysis has predicted players who made huge strides just by changing their pitching style.

Case Study 3: José Alvarado

In 2021 and 2022, José Alvarado put up ERAs of 4.20 and 3.18 respectively, while pitching in cluster 5 both seasons. These are not terrible numbers, especially compared to the 2019 seasons of Miller and Burnes. However, in both of these seasons, the XGBoost model for cluster 6 predicted that he could have had an ERA of 3.55 in 2021 and 2.24 in 2022 if he switched up his repertoire. He did just that in 2023, switching over to cluster 6 and putting up the best season of his career, with a career low 1.74 ERA. His transition from cluster 5 to cluster 6 can be summarized by his decision to stop throwing his curveball altogether, along with his commitment to attacking hitters in the strike zone more often, as he threw a career high 50.4% of his pitches in the strike zone in 2023. Therefore, Alvarado made the adjustment that the XGBoost model would have suggested, which led to significant improvements in ERA.

Case Study 4: Buck Farmer and Bryan Baker

Buck Farmer is included in this data set every year except for 2021, where he failed to log 40 innings. He has been a solid, but not spectacular, reliever in these seasons, logging ERAs of 4.15, 3.72, 3.83, and 4.20 between the years of 2018 and 2023, omitting 2020 and 2021. In each of these seasons, Farmer was clustered into group 4, which reflects his reliance on the 4 seam fastball. However, for each of these 4 seasons, the XGBoost model predicted that he would have his lowest ERA if he were in cluster 7, predicting ERAs of 3.33, 3.75, 2.43, and 3.25, which represent improvements in every season except for 2019, and an improvement greater than the RMSE of 0.837 runs associated with the gradient boosting model for cluster 7 in the two most recents seasons. Cluster 7 represents an increased use of sinkers, which is especially

interesting considering Farmer debuted his sinker in 2024, which he has thrown more than his 4 seam fastball so far.

Bryan Baker was also in cluster 4 in 2022 and 2023, the only two years he has pitched more than 1 inning in MLB. He has been quite solid, putting up an ERA of 3.49 and 3.60, respectively. Similarly to Farmer, he has relied heavily on his 4 seam fastball, with a slider and a changeup to compliment it. However, the gradient boosting model for cluster 7 predicted that his ERA could have been much lower, predicting a 2.50 ERA in 2022 and a 2.42 ERA in 2023 if he were in cluster 7. Unlike Farmer, Baker has started the year in AAA, it will be interesting to see how his repertoire has changed once he makes it back to MLB.

An interesting facet of cluster 7 is that it represents the most sweeper use of any cluster. Therefore, this model suggests that Farmer and Baker may have repertoires that would be conducive to including a sweeper. Hence, one suggesting that this model has for these pitchers is to potentially experiment with a new pitch: the sweeper. Farmer has yet to try this pitch out, and Baker has yet to pitch in MLB this year. However, trying out a sweeper is recommended to these two relievers by this analysis, and represents another area in which this model is potentially useful in pitcher development.

Case Study 5: Aaron Bummer and Shintaro Fujinami

Aaron Bummer and Shintaro Fujinami represent two very dynamic, exciting relief pitchers that recently changed teams in 2024. Also, they are both pitchers who this analysis predicts could have had an ERA over 3.5 runs lower if they were in a different cluster in 2023. Bummer, who put up an ERA of 6.79 with the Chicago White Sox last season as a cluster 7 pitcher (reflecting his sinker/sweeper usage) is pitching for the Atlanta Braves in 2024.

According to the XGBoost models, Bummer could have had an ERA as low as 2.78 if he were in cluster 6, or 2.91 in cluster 1. Cluster 1 pitchers usually rely on a 4 seam fastball and a curveball, neither of which are pitches Bummer has been known to use frequently. However, cluster 6 pitchers have a high cutter usage rate, which is a pitch Bummer uses about 10% of the time. Therefore, this analysis suggests that Bummer could benefit from slightly modifying his pitching style to include more cutters.

In his first season in MLB after making the transition from the NPB in Japan, Shintaro Fujinami struggled with the Oakland Athletics and the Baltimore Orioles, combining for an ERA of 7.18 while being grouped in cluster 5. Despite his heavy usage of a 4 seam fastball that touches triple digit velocities, one contributing factor to his grouping in cluster 5 was likely his low spin rates. However, the gradient boosting model for cluster 4 suggested that he could have had an ERA of 3.48, and the model for cluster 7 predicted an ERA of 3.67. One potential method for converting Fujinami from a cluster 5 to a cluster 4 pitcher likely would be to increase the spin rate on his 4 seam fastball. However, this is easier said than done, as increasing spin rate is something that most MLB pitchers are attempting to maximize regardless of pitching style, so it is likely that Fujinami already has attempted to increase his spin rate. Therefore, it may be more reasonable to convert Fujinami to cluster 7, which would likely mean experimenting with using a sinker instead of his 4 seam fastball, increasing his sweeper usage, and/or throwing more strikes. Again, it is unclear if these recommendations are feasible for Fujinami, who is currently pitching in AAA for the New York Mets. However, regardless of specific recommendations, it is encouraging news for exciting pitchers like Bummer and Fujinami that this approach identifies them as pitchers who are primed to break out.

Limitations and Future Directions

One obvious future direction that leads naturally from the analysis and the case studies considered above is to expand this approach to starting pitchers. Eventual ace starting pitchers such as Burnes and his former teammate Freddy Peralta were identified by the gradient boosting models as players who had rough years as relievers, but could improve greatly if they changed their style. Both of these pitchers did eventually find success, so much so that they graduated from the bullpen and became starters, meaning they were no longer in the dataset after their first season. Therefore, it would be interesting to include starting pitchers in the data as well, especially to monitor players like Burnes and Peralta, to test whether or not they found success by changing clusters, and, if so, if they changed clusters to the cluster that the gradient boosting models would have recommended.

Another potential future direction would be to consider more factors in the clustering, outside of just pitch repertoire, which, other than in the strike zone percentage, was basically the only factor considered. Variables such as handedness, wingspan, and armslot are often considered when qualitatively grouping pitchers, but these factors were ignored when making these clusters.

In terms of limitations, the clustering method was also limited due to the nature of the data. For example, if a pitcher never threw a slider, then his average slider velocity and spin rate was zero. This does not seem like a great way to classify this data point, but there are not many obvious alternatives. Instead, especially for factors such as velocity and spin rate, it may have been better to group all types of fastballs and all types of offspeed pitches together. For example, one variable could have been “fastest velocity”, which would be the velocity in mph of that pitcher’s fastest pitch, whether it was a 4 seam fastball, a sinker, a cutter, etc. However, not all

pitchers even throw an offspeed pitch, and each reliever throws a different number of pitches, so the obstacle of how to classify velocity and spin rate when a certain pitcher does not utilize the pitch in question would remain an obstacle regardless of the approach. Another potential future direction would be to pursue a clustering technique that could handle missing values in data, and leave statistics such as velocity and spin rate as NAs for pitchers who do not throw the pitch in question, but there are limitations to this approach as well.

Finally, one limitation of the gradient boosting aspect of this analysis is that it predicts ERA, which is a statistic that can be impacted by things outside of a pitcher's control, such as bad defense, lucky hitting, or ballpark factors. In fact, according to the article *What is the Best Predictor of ERA?* by Cameron Kaplinger, statistics like SIERRA and FIP are better at predicting future ERA than past ERA. Therefore, ERA might not be the best statistic to represent how well a pitcher has done, especially considering the small sample size of a relief pitcher. For example, consider a reliever who has pitched 45 innings with an ERA of 4.00, meaning he has allowed a total of 20 earned runs. Imagine that, in his 46th inning, the bases are loaded with two outs, and he allows a 330 foot fly ball down the right field line. If the game is being played in Wrigley Field, the ball will be caught over 20 feet in front of the fence, and his ERA will drop to 3.91. However, if the game is being played in Fenway Park, the ball will sail over 20 feet into the stands, and his ERA (assuming he secures the third out of the inning) will jump to 4.70. This represents just one example of how a single play can lead to a fluctuation of almost an entire run in a relief pitcher's ERA, completely due to factors outside of his control. Therefore, using a statistic such as xERA, xFIP, or SIERRA, which takes into account batted ball data to determine how well balls are hit, and not just their outcome, could be a better way to represent relief pitcher success than just considering ERA.

Conclusions

This analysis appears to show that there are distinct pitching styles in MLB, which is not a surprise. Also, it shows that the success of pitchers in each group (success defined by a low ERA) can be predicted using different underlying factors depending on the pitcher's style and repertoire. Therefore, it is possible to identify pitchers who have struggled using a certain pitching style, but who have statistical profiles that suggest that they would have success if they utilized a different pitching style.

In conclusion, our protocol of clustering MLB relievers, creating separate gradient boosting models to predict ERA in each cluster, and then applying the models to every reliever to consider which relievers should think about changing their pitching style or repertoire was successful in predicting breakouts from pitchers such as Corbin Burnes, Shelby Miller, and José Alvarado. These are just a few examples out of over a thousand relief pitcher seasons analyzed, and there are likely examples of times when the model incorrectly advises pitchers. However, in terms of the model's usefulness, if an MLB front office wants to consider ways to identify cheap, low risk pitchers who may be able to have great success if they tweaked their style, this is a potentially effective method to do so. Not every pitcher who this model predicts will succeed in another cluster actually will, there are many outside factors at play. However, if a team acquires a handful of pitchers who this model predicts will break out, just one success could provide a huge boost to a team's bullpen. Finally, considering that multiple breakout pitchers (like Burnes, Miller, and Alvarado) seemed to follow what would have been the recommendations of this analysis, it is likely that MLB teams are already considering similar approaches to finding ways to improve pitcher effectiveness.

References

Baseball Savant. “Pitch Arsenals Leaderboard” *Baseball Savant,* MLB, https://baseballsavant.mlb.com/leaderboard/pitch-arsenals?year=2023&min=500&type=avg_speed&hand=. Accessed 7 May 2024.

Boehmke, Brad, and Brandon M. Greenwell. *Hands-on Machine Learning with R*. CRC Press, Taylor & Francis Group, 2020.

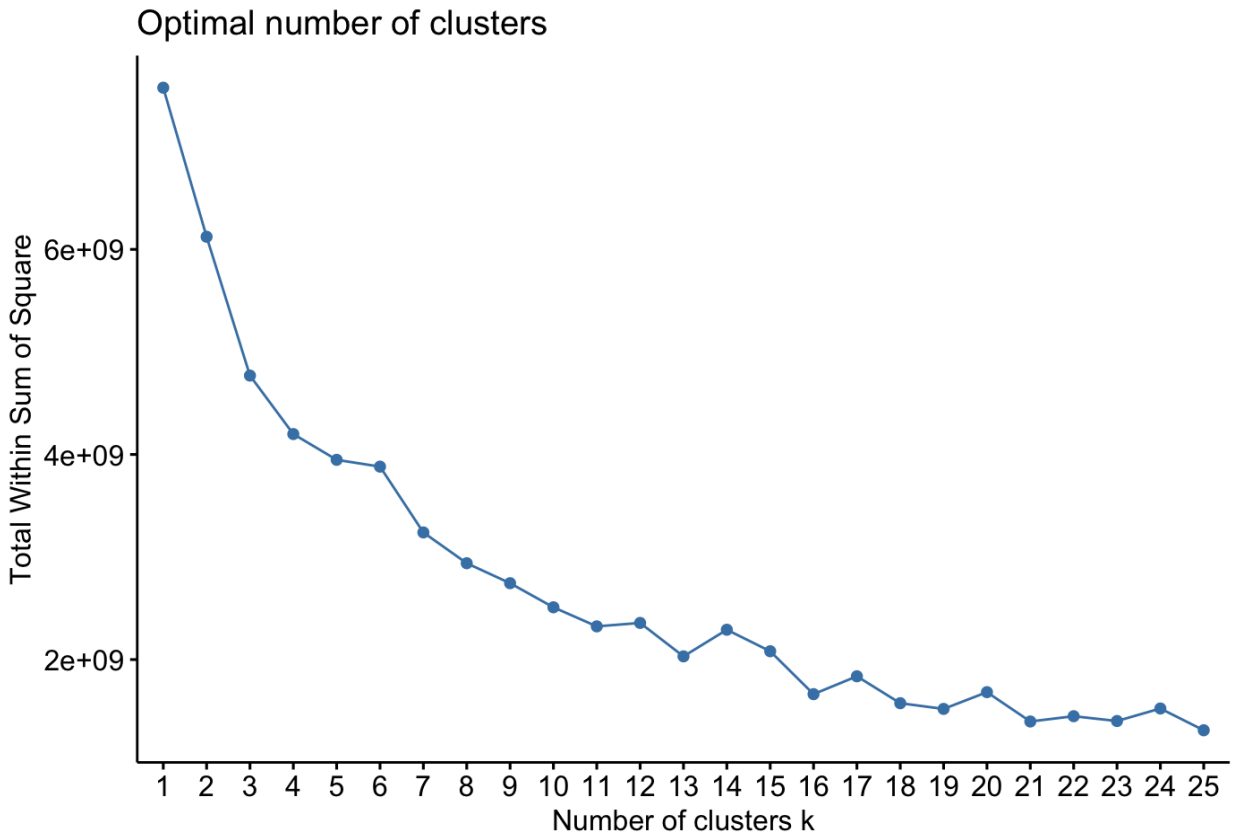
Dvorocsik, Gregoy, and Eno Sarris, and Joseph Camp. *Using Clustering to Find Pitch Subtypes and Effective Pairings*. Baseball Research Journal, Summer 2020.

Fangraphs. “Major League Leaderboards - 2023 - Pitching” *Fangraphs*, Fangraphs, https://www.fangraphs.com/leaders/major-league?pos=all&stats=pit&type=8&startdate=&enddate=&month=0&season1=2018&season=2023. Accessed 7 May 2024.

Kaplinger, Cameron. “What Is the Best Predictor of Era?” *Medium*, Medium, 13 June 2023, medium.com/@cameron.kaplinger/what-is-the-best-predictor-of-era-a2b39677bbd2.

Net World Sports. “Baseball Field Dimensions Guide.” *Net World Sports*, www.networldsports.com/buyers-guides/baseball-field-dimensions-guide#:~:text=The%20deepest%20right%20field%20in,and%20310ft%2F94.5m%20respectively. Accessed 7 May 2024.

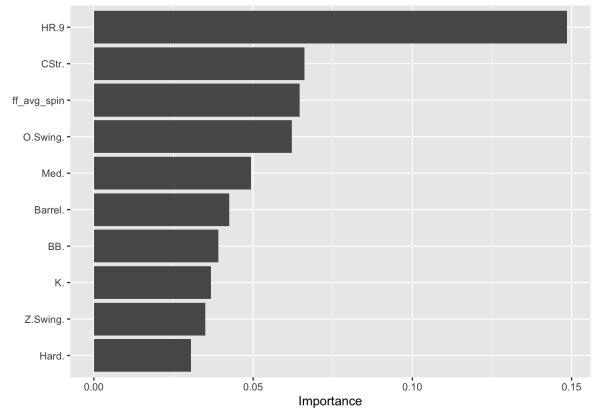
Appendix 1: K-Means Clustering elbow plot



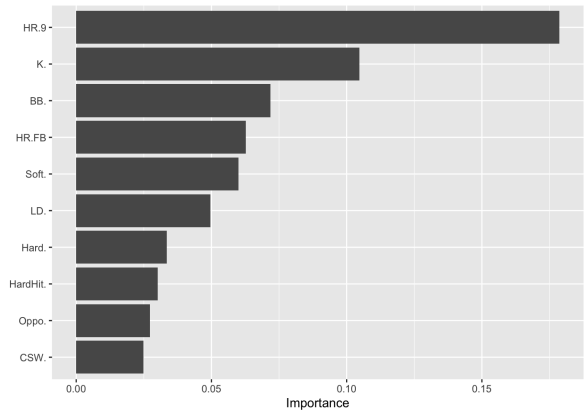
This elbow plot suggests a few options for the ideal number of clusters. However, it seems that the largest drop off in total within-cluster sum of squares occurs at 7 clusters. It can be argued that there is an “elbow” at 13, 16, or even 21 clusters as well, with smaller WSS values. However, the initial “elbow” at 7 clusters is apparent, and results in a more reasonable and useful number of clusters.

Appendix 2: VIP plots for XGBoost models

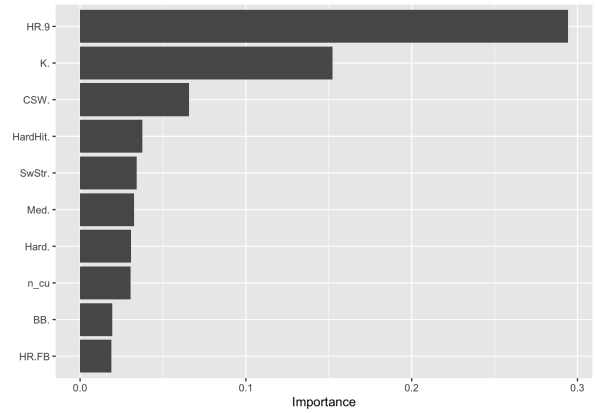
Cluster 1:



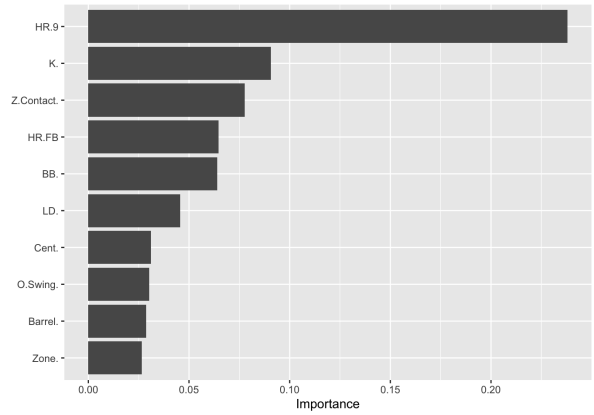
Cluster 2:



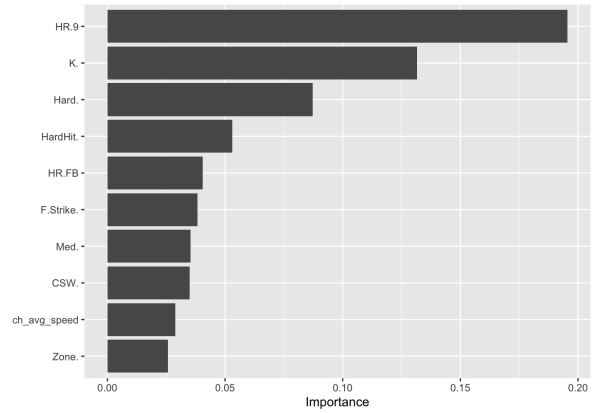
Cluster 3:



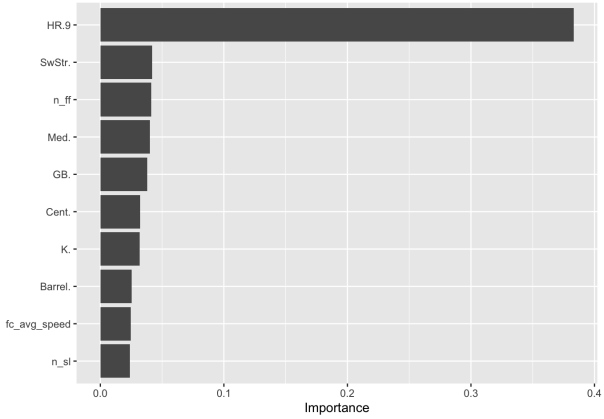
Cluster 4:



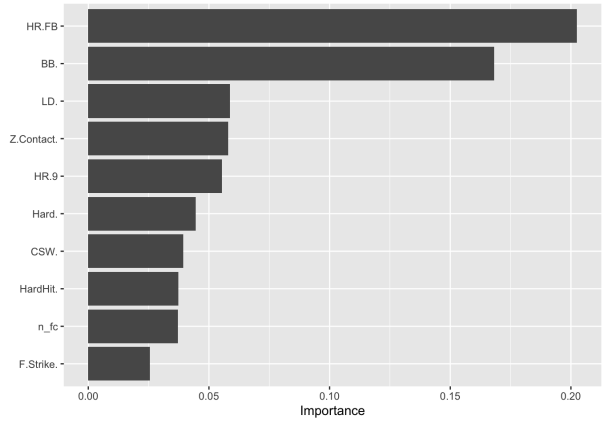
Cluster 5:



Cluster 6:



Cluster 7:


HR.FB
BB.
LD.
Z.Contact.
HR.9
Hard.
CSW.
HardHit.
n_fc
F.Strike.
0.00
0.05
0.10
0.15
0.20
Importance