Connor Moss & Liam Rosengren
Baseball Executive Summary

Anytime a group of professional analysts produces a list of the top ten players in a given league, the response will be unpredictable. But, instead of eye-rolling about how inaccurate the MLB Top Ten Right Now of starting pitchers in 2024 is, we decided to use two primary statistical methods (clustering and gradient boosting) to form our predictions for the upcoming 2024 MLB season.

We took inspiration from three primary pieces of literature. One deals with the idea of run-scoring potential to predict wins through the application of Markov chains. The next focused on the kinematic changes that baseball pitchers adapt as they age to remain competitive at the highest level (i.e. shorter strides, more closed pelvis, more lead knee flexion). Lastly, the third predicted the outcomes of MLB games using a similar style of gradient boosting analysis as we employed.

We started with an age analysis to understand how our variables measured up depending on the age of the pitcher. Upon finding that within our dataset, older pitchers generally had lower ERA and higher WHIP we had a sense of the type of pitcher that was going to top our list. Given the subjective nature of these rankings, the lack of a response variable guided us in the direction of unsupervised learning. Thus, we began with wrangling the data to perform clustering.

To play on the idea of recency bias, we weighted each of our five years of data differently, with more recent weights receiving higher values. Thereby making our results more generalizable for a prediction in the 2024 season. We found three distinct clusters:

- Cluster One – High IP, High WAR, “Proven contributors.”
- Cluster Two – High K, High RA, “Gambles”
- Cluster Three – Lowest overall effectiveness of group, “Potentially Promising Prospects”

Then we created our own “score” to incorporate some measurements amongst the group to compare our list to the professionals. Our score calculation involved assigning weights to different

variables in our dataset. For example, ERA would lower the total score by 2, whereas SO/9 would increase the score by 1.5. This was done after all of the averaged stats were normalized to ensure fairness.

Our score calculations came back promising as our list ranked (in order): Justin Verlander, Gerrit Cole, and Spencer Strider as the top three pitchers for this upcoming season. In total, our predictions from the weighted score calculation shared 6 out of the 10 guys that appeared on the MLB Top Ten Right Now list.

We followed up this encouraging prediction with a dive into uncharted territory. Using two different methods of gradient boosting, we sought to construct a model that rivaled "The Shredder". To begin, we built a bare-bones gradient boosting model from scratch, using sequential improvements to optimize our Root Mean Squared Error (RMSE) while predicting Wins Above Replacement (WAR) for pitchers in 2024. Similar to MLB Network's proprietary machine, we made use of historical data, standard and advanced data to construct our models.

The "brute force" model used a learning rate of 0.05 with 500 trees. While 500 trees sounds large, we found that this optimized our RMSE without risking overfitting thanks to our small learning rate. With an RMSE of 0.0513, this model unambiguously nailed the optimization of minimizing RMSE. Additionally, with 5 out of 10 shared players to the list of "The Shredder" (and 2 more just outside), we would be fools to argue with our results.

In the name of robustness, we expanded our gradient boosting analysis to something more extreme: R's XGBoost function. After tuning our boosting, tree, and regularization parameters, we turned the crank on XGBoost many times. In part to observe the variety in the rankings, and in part to optimize our RMSE. A series of trials found that the RMSE of our XGBoost model ranged from (0.17, 0.78). In addition to being worse than our scratch model, the lists created often included deep-cut names, with the best one we found containing 4 of the 10 names listed by MLB Network.

In summary, we constructed three strong models for MLB's top ten starting pitchers for the upcoming 2024 season. While Sandy Alcantara (Tommy John Surgery) won't be there, the rest of the names could feasibly be the aces of their respective teams come October, leading them to the Fall Classic.